Ensemble Learning Techniques for Structured and Unstructured Data

Michael A. King

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Business Information Technology

Alan S. Abrahams, Co-chair
Cliff T. Ragsdale, Co-chair
Lance A. Matheson
G. Alan Wang
Chris W. Zobel

February 13, 2015
Blacksburg, Virginia, United States

Keywords: ensemble methods, data mining, machine learning, classification, structured data, unstructured data

Ensemble Learning Techniques for Structured and Unstructured Data

Michael A. King

Abstract

This research provides an integrated approach of applying innovative ensemble learning techniques that has the potential to increase the overall accuracy of classification models. Actual structured and unstructured data sets from industry are utilized during the research process, analysis and subsequent model evaluations.

The first research section addresses the consumer demand forecasting and daily capacity management requirements of a nationally recognized alpine ski resort in the state of Utah, in the United States of America. A basic econometric model is developed and three classic predictive models evaluated the effectiveness. These predictive models were subsequently used as input for four ensemble modeling techniques. Ensemble learning techniques are shown to be effective.

The second research section discusses the opportunities and challenges faced by a leading firm providing sponsored search marketing services. The goal for sponsored search marketing campaigns is to create advertising campaigns that better attract and motivate a target market to purchase. This research develops a method for classifying profitable campaigns and maximizing overall campaign portfolio profits. Four traditional classifiers are utilized, along with four ensemble learning techniques, to build classifier models to identify profitable pay-per-click campaigns. A MetaCost ensemble configuration, having the ability to integrate unequal classification cost, produced the highest campaign portfolio profit.

The third research section addresses the management challenges of online consumer reviews encountered by service industries and addresses how these textual reviews can be used for service improvements. A service improvement framework is introduced that integrates traditional text mining techniques and second order feature derivation with ensemble learning techniques. The concept of GLOW and SMOKE words is introduced and is shown to be an objective text analytic source of service defects or service accolades.

Acknowledgements

I would like to express my deepest gratitude and thanks to the numerous people who supported me through this incredible life journey. It was an honor to work with my co-chairs, who are a brain trust of knowledge. My co-chair, Dr. Cliff Ragsdale, provided amazing insights, guidance, endless wordsmithing, and timely words of wisdom throughout the dissertation process. Dr. Cliff Ragsdale stood as the "rock precipice" during the numerous challenges I encountered along this journey, such as the discovery of software bugs that required rerunning three months of analysis, and my ongoing quest with "getting the equations correct." And most importantly, I would like to thank Dr. Ragsdale for encouraging me to apply to the Business Information Technology Ph.D. program. My other co-chair, Dr. Alan Abrahams, delivered the energy to sustain this process when most needed. Dr. Abrahams wore many hats during this process; coach, cheerleader, motivator, and most importantly, mentor. Dr. Abrahams provided the numerous creative and innovative research ideas required during this dissertation writing process, and the solutions to clean up my mistakes. He generously provided several invaluable and interesting data sets that make this applied research usable by industry.

I would to thank Dr. Lance Matheson for giving the amazing opportunity of participating in the Pamplin Study Abroad program. Dr. Matheson's humor and support was a sustaining force through the entire PhD process. I would like to thank Dr. Chris Zobel and Dr. Alan Wang for the support and insights on several research projects that has shaped and honed my research skills. I would like to thank Dr. Bernard Taylor who has provided the leadership for the BIT department. I also would like to thank Tracy McCoy, Teena Long, and Sylvia Seavey whom cheerfully provided the back office support when most needed. Many thanks to my fellow BIT student colleagues and friends. The opportunity to share ideas and support our mutual goals through the process has made the journey more rewarding. I am grateful to my many friends and family members that have encouraged me by kindly asking about and patiently listening to my many ideas along the way.

This Ph.D. journey was motivated by an idea of my wife, Karen King. Without her amazing love, stamina, and endless supply of optimism, I could not have achieved this goal. Blessed be the house designed and supported by the Beaver Hutch Framework! Your love has sustained me through this long journey. It is your turn now…

Table of Contents

List of Exhibits

List of Tables

Notes

The work presented in Chapter 2 is published as:
King, M. A., A. S. Abrahams and C. T. Ragsdale (2014). "Ensemble Methods for Advanced Skier Days Prediction." Expert Systems with Applications 41(4, Part 1): 1176-1188.

The work presented in Chapter 3 is published as:
King, M. A., A. S. Abrahams and C. T. Ragsdale (2015). "Ensemble Learning Methods for Pay-per-click Campaign Management." Expert Systems with Applications 42(10): 4818-4829.

Chapter 1:

Introduction

"And much of what we've seen so far suggests that a large group of diverse individuals will come up with better and more robust forecasts and make more intelligent decisions than even the most skilled decision maker."
James Surowiecki

1. Ensemble methods overview

Ensemble learning algorithms are general methods that increase the accuracy of predictive or classification models such as decision trees, artificial neural networks, Naïve Bayes, as well as many other classifiers (Kim, 2009). Ensemble learning, based on aggregating the results from multiple models, is a more sophisticated approach for increasing model accuracy as compared to the traditional practice of parameter tuning on a single model (Seni and Elder, 2010). The general ensemble technique, illustrated in Exhibit 1.1, is a two-step sequential process consisting of a training phase where classification or predictive models are induced from a training data set and a testing phase that evaluates an aggregated model against a holdout or unseen sample. Although there has been general research related to combining estimators or forecasts (Major and Ragsdale, 2000; Clemen, 1989; Barrnett, 1981), ensemble methods, with respect to classification algorithms are relatively new techniques. Thus, it is important to clarify the distinction between ensemble methods and error validation methods: ensemble methods increase overall model accuracy while cross validation techniques increase the precision of model error estimation (Kantardzic, 2011).

The increased accuracy of an ensemble, because of model variance reduction and to a lesser extent bias reduction, is based on the simple but powerful process of group averaging or majority vote (Geman, 1992). For example, the analogy of the decision process of an expert committee can demonstrate the intuition behind ensemble methods. A sick patient consults a group of independent medical specialists and the group determines the diagnosis by majority vote. Most observers would agree that the patient received a better or more accurate diagnosis as compared to one received from a single specialist. The accuracy of the diagnosis is probably higher and the variance or error of possible misdiagnosis is lower, although more agreement does not always imply accuracy.

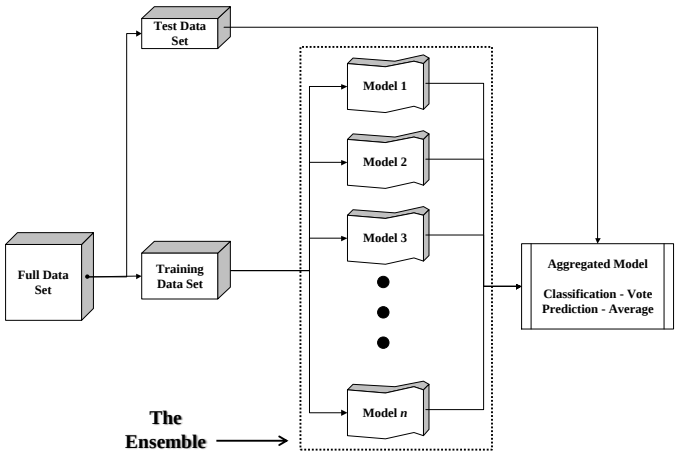


Exhibit 1.1. Generalized Ensemble Method

As additional insight for ensemble methods, James Surowiecki, in his book The Wisdom of Crowds, offers a contrarian argument supporting group decisions, saying in many cases under the correct circumstances, group decisions are often more accurate when compared to the single most accurate decision (Surowiecki, 2004). Immediately the reader probably remembers the United States' failed Cuban Bay of Pigs invasion and the NASA Columbia explosion as two of the most notorious examples of groupthink, as evidence against trusting the wisdom of the majority. However, while Surowiecki acknowledges these two examples and more, he provides examples of where group decisions were in fact much more accurate and at times almost exact.

One example offered by Surowiecki as support for group decisions is the process used by Dr. John Craven, a United States Naval Polaris project engineer, to deduce the location of the USS Scorpion, a United States submarine that sank while en route back to the Norfolk naval base in 1968. Craven assembled a group of eclectic naval professionals, gave them the available information and asked each individual to independently make his best location estimate. Using Bayesian logic, to the disbelief of the United States Naval command, Craven aggregated the locations from his group and made a prediction. Five months after the USS Scorpion was declared missing, United States Naval command finally permitted a search ship to track Dr. Craven's estimates. After several days, the search ship located the submarine at a depth of nearly 11,000 feet only 220 yards away from Craven's guess.

An additional example provided by Surowiecki and a precursor to the modern day jelly bean jar counting exercise, is the insight discovered by Sir Francis Galton, a prominent statistician during the late 1800s, while visiting a local county fair. He observed a contest where a hawker challenged fair goers to guess the final butchered and dressed weight, while viewing the live ox. Galton indicated that the participants came from a wide range of socio economic backgrounds and that the guesses were made independently, without any influence from the game hawker. Galton was able to gather approximately 800 wagers and then calculated the mean weight. He noted that the crowd average was 1,197 just 1 pound below the true dressed weight of 1,198. Galton's observations and subsequent statistical analysis were motivation for his seminal work, "Vox Populi" (Galton, 1907).

Open source software creation, prediction markets, and wiki technology such as Wikipedia are all recent examples that Surowiecki cites as collaborative processes falling under the wisdom of the crowd umbrella. However, regardless of their time of occurrence, Surowiecki argues that these examples all share the four dimensions required for a group to outperform any single member: diversity, independence of action, decentralization and effective decision aggregation. If any of these dimensions are absent, the negative consequences of groupthink, where individuals change their own opinion or belief in favor of the crowd consensus, are substantial. The diversity requirement brings different sources of information to the decision process, which expands the solution space of possible solutions. A group decision cannot be more accurate if all group members choose or suggest the same solution. The independence of action requirement mitigates the possibility of a herd mentality where group members sway or influence other members towards one specific solution. In addition to independence of action, physical decentralization, the third dimension, creates the condition where group members have the ability to act in their own best interest while concurrently interacting to produce collaborative solutions. Effective decision aggregation, the last required dimension for group wisdom, is a process where single group member errors balance each other out while allowing superior solutions to survive.

The interesting aspects here are the parallels that can be drawn between Surowiecki's informal criteria for group wisdom and the extensive body of literature pointing to very similar dimensions as prerequisites for improving the accuracy of ensemble learning algorithms (Das, R., et al., 2009; Claeskens and Hjort, 2008; Han and Kamber, 2006). Ensemble learning methods fundamentally

work the same way as effective group wisdom, by taking a set of independent and diversified classification or prediction models and systematically aggregating the results based on an objective criteria, such as majority vote or averaging.

Armstrong (2001) makes an excellent case for general forecast combination, backed by a comparative empirical study of 30 forecast combination studies, from a wide range of business contexts, with an average per study error reduction of 12.5%. Supported by these findings, Armstrong presents a formal framework for combining forecasts that includes similar themes when compared to Surowiecki's four dimensions for effective group decisions. To increase forecast accuracy, it is essential that forecasts have varied data sources, be derived from several forecasting methods or algorithms, and use a structured or mathematical method for combining the forecasts. Different data sources provide the diversity needed to expand the available solution space, while utilizing several forecasting methods increases researchers' ability to search the solution space. Armstrong also indicates that "objectivity is enhanced if forecasts are made by independent forecasters." Armstrong argues that combined forecasts are most valuable when there is a high cost associated with forecasting error and when there is substantial uncertainty surrounding the selection of the most applicable forecasting method.

2. Conceptual foundation

Dietterich makes the argument that there are three theoretical reasons why a set of classifiers frequently outperform an individual classifier measured by classification accuracy (Dietterich, 2000). The first reason is statistical in nature and occurs when a classifier must form a hypothesis from a solution space that is much larger than the solution space constructed by the available training data set. As illustrated by Exhibit 1.2.1, the outer boundary represents the full solution space *S* while the inner boundary represents a set of accurate classifiers on the training data. The point *C* is the true classifier model and is unknown. By averaging the set of classifiers c_n, the aggregate classifier *c`* forms an excellent approximation of *C* and thus minimizes the probability of selecting a substantially less accurate single classifier.

A second reason that ensembles can be more accurate than a single classifier is computationally

founded. Numerous machine learning algorithms, such as artificial neural networks or decision trees perform a gradient search or random search routine to minimize classification error. However, while learning, these algorithms can become stalled at a local optima, especially when the training data set size is large, where the eventual optimization problem could become NP-hard (Blum and Rivest, 1988). If an ensemble of classifiers is constructed with local search using a random start for each classifier, the aggregated search could cover the solution space more effectively and provide a more accurate estimate of the true classifier *C*. Exhibit 1.2.2 demonstrates how random starts of classifiers c_n can converge and an aggregated solution *c`* is much closer to the true classifier *C*.

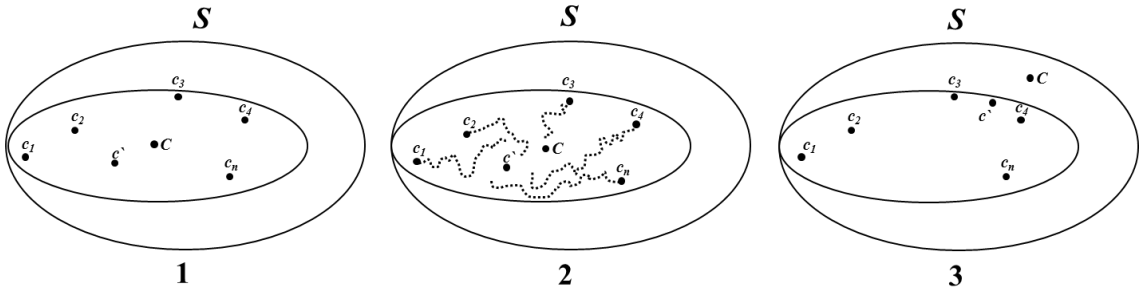


Exhibit 1.2. Conceptual Foundations for Ensembles

A final reason that ensembles can be more accurate than a single classifier is representational in the sense that the current solution space hypothesized by the classifiers does not adequately "represent" the true model. Single classifier models c_n may not have the degree of model complexity required to accurately model the true classifier *C*. In this case, it could be more efficient to train an ensemble of classifiers to a level of desired accuracy than to train a single classifier of high complexity, requiring numerous parameters changes (Seni and Elder, 2010). Exhibit 1.2.3 illustrates that the true classifier *C* is not modeled or represented well by the trained set of classifiers c_n . *The ensemble solution c`, containing the possibility of substantial error, provides a more accurate estimation of the true classifier C.*

3. Formalization

Formalizing these analogies and concepts, the error rate of an ensemble of classifiers is less than the error rate of an individual model, being necessary and sufficient when:

- each model has an accuracy rate that is at least marginally better than 50% and,

- the results of each model are independent and,
- the class of a new observation is determined by the majority class (Hansen and Salamon, 1990) such that:

$$f_{combined} = VOTE(f_1, f_2, f_3, f_4, \ldots f_N)$$

It can be shown that, when these model requirements are met, the error rate of a set of *N* models follows a binomial probability distribution (Kuncheva, 2003), thus the error rate of an ensemble equals the probability that more than *N*/2 models incorrectly classify. For example, with an ensemble consisting of 20 classifiers, *N*, where each model performs a two category classification task, and each with an error rate of ε = .3, the ensemble error rate is

$$Ensemble\ \varepsilon = \sum_{n=N/2}^{N} \binom{N}{n} .3^n (1-.3)^{N-n} = .047962$$

This ensemble error rate is substantially lower than the individual model error rate of .3. Note that the summation started at *n*=10, to represent when 10 or more models actually misclassified a test set observation, while 10 models or less correctly classify a test set observation. As a comparison, an ensemble of 50 classifiers, under the same assumptions, has an error rate of ε = .00237, considerably lower than the previous example.

The theoretical framework that supports the validity of the increased accuracy of ensemble learning techniques is called bias-variance decomposition of classification error (Dietterich and Kong, 1995; Geman, et al., 1992). The accuracy of a general statistical estimator (θ) is measured by the mean squared error:

$$MSE = Bias(\theta)^2 + Var(\theta) + e$$

Bias error is a deviation measurement of the average classification model created from an infinite number of training data sets from the *true* classifier. Variance error is the error associated with *a single* model with respect to each other or in other words, the precision of the classification model when trained on different training data sets (Geman, et al., 1992). Considering bias-variance, there is a tradeoff between lowering bias or lowering variance, with respect to the ability of a model to correctly map the actual data points, based on a specific machine learning model and a specific training data set. The true machine learning model for a given situation has a specific architecture and parameter set that are, of course, typically unknown which makes bias reduction on real world

data sets difficult. For example, the architecture of a polynomial regression function is determined by the functional degree while the model parameters consist of the variable coefficients. Models that have few parameters are typically inaccurate due to a high bias, because of limited model complexity and thus an inadequate ability to capture the true model. Models with numerous parameters are also routinely inaccurate because of high variance as a consequence of higher levels of flexibility and over fitting (Hastie, et al., 2009)

As additional theoretical support, based on Hastie, et al., (2009), variance reduction by averaging a set of classifiers can be formally deduced as follows:

Assume there are *D* datasets used to train *d* classification models for input vector *X*

$$y_d(X)$$

Making the assumption that the true classification function is

$$F(X)$$

it follows that

$$y_d(X) = F(X) + e_d(X)$$

The expected Sum of squared error for an input vector X, per model, is shown by

$$E_X\left[\left(y_d(X) - F(X)\right)^2\right] = E_X[e_d(X)^2]$$

The average error per individual classification model therefore is

$$E_{\mu,individual} = \frac{1}{D}\sum_{d=1}^{D} E_X\,[e_d(X)^2]$$

The average for an ensemble is given by

$$\mu_{combined} = \frac{1}{D}\sum_{d=1}^{D} y_d(X)$$

The expected error from the combined prediction is indicated by

$$E_{\mu,combined} = E_X\left[\left(\frac{1}{D}\sum_{d=1}^{D} y_d(X) - F(X)\right)^2\right]$$

which reduces to

$$E_{\mu,combined} = E_X\left[\left(\frac{1}{D}\sum_{d=1}^{D} e_d(X)\right)^2\right]$$

Assuming that the models are independent and their variances are uncorrelated, and the summation

from *1* to *D* and *1* divided by *D* cancel out, it follows

$$E_{\mu,combined} = \frac{1}{D} E_{\mu,individual}$$

The fundamental insight from the derivation above is that the average combined model variance error can be reduced by the term *D* by averaging *D* replicas of the classification model. However, as noted above, these results are valid when the models are completely independent and the associated variance errors are uncorrelated. These assumptions, in practice, are rather unrealistic, since ensemble methods typically create bootstrap replicas from the original data sets and use the same variable set or subsets for all classifier models in the ensemble set, which introduces a degree of *positive* correlation. Thus, the reduction in error will be less than the factor of *D*.

However, it is important to note that when models are dependent and have a degree of negative correlation, the variance error is indicated by:

$$E_{\mu,combined} = \frac{1}{D}\left[VAR\left(\sum_j d_j\right) + 2\sum_i \sum_{i \neq j} \left(COV(d_i, d_j)\right)\right]$$

where the resulting variance error can be lower than a factor of *D*.

Although counter intuitive, when compared to variance reduction by averaging, it can be empirically shown that by adding bias to a known unbiased estimator can actually decrease variance and mean squared error, thus improving overall model performance. To illustrate with a classic example, with respect to the normal distribution, the unbiased estimator for the sample variance is:

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n - 1}$$

while the biased estimator for the sample variance is:

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n}$$

One observation to note in this case is that the biased estimator for the sample variance has the lower mean squared error between the two estimators (Kohavi and Wolpert, 1996). Thus, a modeler must recognize that bias-variance is a function of model flexibility or complexity, when designing machine learning algorithms. More specifically, a modeler must acknowledge the tradeoff represents a continuum between under fitting, namely high bias, and the risk of over fitting and introducing too

much variance. Exhibit 1.3 illustrates this continuum in a stylized format adapted from Hastie, et al., (Hastie, et al., 2009). As model flexibility increases, the training sample error continues to decrease, while over fitting increases for the test sample. The goal, of course, is to determine the model architecture that minimizes the test set variance for the specific classification task at hand. However, as previously mentioned, bias reduction while theoretically possible, in practice is difficult and impractical (Seni and Elder, 2010).

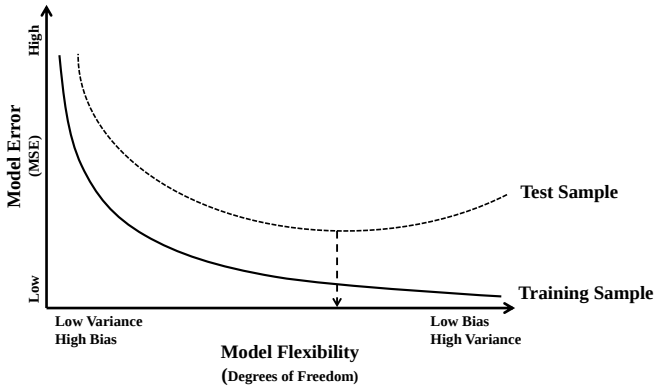


Exhibit 1.3. The Bias Variance Tradeoff

4. Research objectives

This research intends to present ensemble methods to the greater information systems research community, as well as to business management, as an effective tool for increasing classification or prediction accuracy. The overarching strategy for this research stream is to assess the efficacy of ensemble methods given a specific combination of a dependent variable data type and a feature set data type. Exhibit 1.4 illustrates the organization of the three major sections contained in this dissertation with respect to the dependent variable and the feature set type combination. This research defines structured data as well-organized information that adheres to a data model and primarily numeric, while in contrast, defines unstructured data as information represented as free form textual content that does not follow a predefined data model. Chapter 2 addresses the consumer demand forecasting and daily capacity management requirements of a nationally recognized alpine ski resort in the state of Utah, in the United States of America. Both the dependent variable and the feature set are numeric and structured. Chapter 3 discusses the opportunities and challenges faced by a leading firm providing sponsored search marketing services. This chapter develops a method for classifying profitable campaigns and maximizing overall

campaign portfolio profits. The dependent variable is a categorical variable having two levels, and the feature set is numeric data. Chapter 4 illustrates the management challenges of online consumer reviews encountered by service industries and addresses how these textual reviews can be used for service improvements. The dependent variable is a categorical variable having two levels, and the feature set is free form unstructured text. The combination of a continuous dependent variable with an unstructured feature set will provide an opportunity for future ensemble modeling research.

		Feature Set Type	
		Structured Data	**Unstructured Data**
Dependent Variable Type	Prediction - Continuous Data	Chapter 2	Future Work
	Classification - Categorical Data	Chapter 3	Chapter 4

Exhibit 1.4. Research Landscape

This research also answers and provides supporting information for the following research questions:

1. What are the advantages and disadvantages of ensemble methods when compared to standard single classification model techniques?
2. How can researchers accurately estimate ensemble accuracy and compare the accuracy of several ensemble models?
3. Are there base classifiers that are more applicable for ensemble learning methods?
4. What are some of the insights and cautions that researchers or business managers should be cognizant of when employing ensemble methods to data sets from actual business problems?

Five ensemble learning algorithms are discussed in detail, empirically tested, and applied in one or more of the following three chapters. An ensemble learning taxonomy, which describes the ensemble selection criteria, is introduced in Chapter 2 and the discussion continues in the remaining chapters. Appendix A provides the pseudo-code for these five ensemble algorithms. The pseudo-code shown in Appendix A is illustrative of the methods. An industry-accepted software platform, RapidMiner 5.3, was relied on for the specific implementation of each method. The experimental results in the subsequent chapters help answer the four research questions. The key contributions developed from the research discussed in the three main chapters, are also summarized in the final chapter.

Appendix A:

Ensemble pseudo-code

Input:
Data set D = [(X_1, y_1), (X_2, y_2), ... , (X_n, y_n)] *X_n attribute vectors, n observations, y_n predictions.*
First level classification algorithms, $d_{1...S}$
For s = 1 to S
d_s = CreateFirstLevelModels(D) *Create first level models from data set D.*
End
Output:
Combine S model classes by majorty *Combine model outputs by max class.*
Return ensemble class

Voting

Input:
Data set D = [(X_1, y_1), (X_2, y_2), ... , (X_n, y_n)] *X_n attribute vectors, n observations, y_n predictions.*
Set DT = number of decision trees to build
Set P = percentage of attribute set to sample

For i=1 to DT
Take random sample D_i bootstrapping from D size N
Create root decision tree node RN_i using D_i
Call CreateDT(RN_i)
End

CreateDT(RN)
If RN contains leaf nodes of one class then
Return
Else
Randomly sample P of attributes to split on from RN
Select attribute x_n with highest splitting criterion improvement
Create child nodes cn by splitting on RN, RN_1, RN_2,...for all attributes x_n
For i=1 to cn
Call CreateDT(i)
End for
End CreateDT

Output:
Combine DT model classes *Combine model outputs by majority vote.*
Return ensemble majority class

Random Forests

Input:
Data set D = [(X_1, y_1), (X_2, y_2), ... , (X_n, y_n)] — *X_n attribute vectors, n observations, y_n predictions.*
Base classification algorithm, d — *Define base learning algorithm.*
Ensemble size, S — *Number of training loops.*
For s = 1 to S
  D_s = BootstrapSample(D) — *Create bootstrap sample from D.*
  d_s = CreateBaseLearnerModels(D_s) — *Create base models from bootstrap samples.*
  Make model prediction d_s
  Save model prediction d_s
End
Output:
Combine S model classes — *Combine model outputs by majority vote.*
Return ensemble majority class

Boot Strap Aggregation

Input:
Data set D = [(X_1, y_1), (X_2, y_2), ... , (X_n, y_n)] — *X_n attribute vectors, n observations, y_n predictions.*
First level classification algorithms, $d_{1...S}$
Second level meta learner, d_{2nd}
For s = 1 to S
d_s = CreateFirstLevelModels(D) — *Create first level models from data set D.*
End
D_{New} = 0 — *Start new data set creation.*
For i = 1 to n
  For s = 1 to S
    C_{is}= d_s (X_i) — *Make prediction with classifier d_s*
  End
D_{New} = D_{New} ∪ [((C_{i1}, C_{i2}, ... , C_{iS}), y_i)] — *Combine to make new data set.*
End
$d_{Trained2nd}$ = d_{2nd}(D_{New}) — *Train meta model to new data set.*
Output:
Return ensemble prediction = $d_{Trained2nd}$ — *Ensemble prediction.*

Stacked Generalization

Input:
Data set D = [(X_1, y_1), (X_2, y_2), ... , (X_n, y_n)] — *X_n attribute vectors, n observations, y_n predictions.*
Ensemble size, S — *Number of training loops*
Subspace dimension, A — *Number of attributes for subspace*
  For i = 1 to S
    SS_i = CreateRandomSubSpace(D, A) — *Create new variable set for training input*
    d_s = CreateBaseLearnerModels(D_s) — *Create base models from bootstrap samples*
    Make model prediction d_s
    Save model prediction d_s
  End
Output:
Average S model predictions — *Combine model outputs by mean*
Return ensemble prediction

Random Subspace

References

Armstrong, J. S. (2001). Combining Forecasts. Principles of Forecasting: A Handbook for Researchers and Practitioners. J. S. Armstrong. Amsterdam, Kluwer Academic Publisher.

Barnett, J. A. (1981). Computational Methods for a Mathematical Theory of Evidence. International Joint Conference on Artificial Intelligence. A. Drinan. Vancouver, CA, Proceedings of the Seventh International Joint Conference on Artificial Intelligence : IJCAI-81, 24-28 August 1981, University of British Columbia, Vancouver, B.C., Canada. 2.

Blum, A. L. and R. L. Rivest (1992). "Training a 3-Node Neural Network is NP-Complete." Neural Networks 5(1): 117-127.

Claeskens, G. and N. L. Hjort (2008). Model Selection and Model Averaging. Cambridge Series in Statistical and Probabilistic Mathematics; Variation: Cambridge Series on Statistical and Probabilistic Mathematics., New York.

Clemen, R. T. (1989). "Combining Forecasts: A Review and Annotated Bibliography." International Journal of Forecasting 5(4): 559-583.

Das, R., I. Turkoglu and A. Sengur (2009). "Effective Diagnosis of Heart Disease Through Neural Networks Ensembles." Expert Systems with Applications 36(4): 7675-7680.

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. Multiple Classifier Systems. J. Kittler and F. Roli. Berlin, Springer-Verlag Berlin. 1857**:** 1-15.

Dietterich, T. G. and E. B. Kong (1995). Error-correcting Output Coding Corrects Bias and Variance. International Conference on Machine Learning, Tahoe City, CA, Morgan Kaufmann.

Galton, F. (1907). "Vox populi." Nature, 75, 450–45.

Geman, S., E. Bienenstock and R. Doursat (1992). "Neural Networks and the Bias/Variance Dilemma." Neural Computation 4(1): 1-58.

Han, J. and M. Kamber (2006). Data mining : Concepts and Techniques. Amsterdam; Boston; San Francisco, CA, Elsevier ; Morgan Kaufmann.

Hansen, L. K. and P. Salamon (1990). "Neural Network ensembles." Pattern Analysis and Machine Intelligence, IEEE Transactions on 12(10): 993-1001.

Hastie, T., R. Tibshirani and J. H. Friedman (2009). The Elements of Statistical Learning : Data Mining, Inference, and Prediction. New York City, Springer.

Kantardzic, M. (2011). Data Mining: Concepts, Models, Methods, and Algorithms. Hoboken, N.J., John Wiley : IEEE Press.

Kim, Y. (2009). "Boosting and Measuring the Performance of Ensembles for a Successful Database Marketing." Expert Systems with Applications 36(2, Part 1): 2161-2176.

Kohavi, R. and D. H. Wolpert (1996). Bias Plus Variance Decomposition for Zero One Loss Functions. Machine Learning: Proceedings of the 13th International Conference, Morgan Kaufmann.

Kuncheva, L. I. and C. J. Whitaker (2003). "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy." Machine Learning 51(2): 181-207.

Major, R. L. and C. T. Ragsdale (2000). "An Aggregation Approach to the Classification Problem Using Multiple Prediction Experts." Information Processing and Management 36(4): 683-696.

Seni, G. and J. F. Elder (2010). Ensemble Methods in Data Mining : Improving Accuracy Through Combining Predictions. San Rafael, Morgan and Claypool Publishers.

Surowiecki, J. (2004). The Wisdom of Crowds : Why the Many are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations. New York, Doubleday.

Chapter 2

Ensemble Methods for Advanced Skier Days Prediction

"Prediction is very difficult, especially if it's about the future."
Niels Bohr

The tourism industry has long utilized statistical and time series analysis, as well as machine learning techniques to forecast leisure activity demand. However, there has been limited research and application of ensemble methods with respect to leisure demand prediction. The research presented in this paper appears to be the first to compare the predictive power of ensemble models developed from multiple linear regression (MLR), classification and regression trees (CART) and artificial neural networks (ANN), utilizing local, regional, and national data to model skier days. This research also concentrates on skier days prediction at a micro as opposed to a macro level where most of the tourism applications of machine learning techniques have occurred. While the ANN model accuracy improvements over the MLR and CART models were expected, the significant accuracy improvements attained by the ensemble models are notable. This research extends and generalizes previous ensemble methods research by developing new models for skier days prediction using data from a ski resort in the state of Utah, United States.

Keyword: Ensemble learning; data mining; forecasting; skier days.

1. Introduction

Over the past two decades, consumer travel behavior and patterns have changed. The length of both the traditional family vacation and the associated planning horizon has significantly decreased (Zalatan, 1996; Luzadder, 2005; Montgomery, 2012). This trend is specifically evident with respect to snow skiing leisure activities at North American ski resorts. According to John Montgomery, managing director with Horwath HTL, a leading consulting firm in the hospitality industry, "if you booked a family ski trip 10 years ago, it was for a Saturday to Saturday block. Come hell or high water you were going." However, extended family ski vacations are now the rarity while shorter trips planned several days before departure have become quite common (Montgomery, 2012). This change is at least partially due to the Internet providing potential travelers with immediate travel decision information about snow conditions and last minute travel promotions.

Management at ski resorts must continue to adapt to these changing travel patterns by employing accurate demand forecasting techniques which, in turn, influence resort capacity planning

operations. The tourism industry has long utilized statistical and time series analysis, as well as machine learning techniques to forecast leisure activity demand. However, there has been limited research and application of ensemble methods with respect to leisure demand prediction. This research uses local, regional, and national data to construct a skier days prediction model for a Utah-based ski resort. A skier day is the skiing industry standard metric for a single skier or snowboarder visit at one resort for any amount of time during one day (www.nsaa.org). We illustrate the predictive accuracy of forecasting models developed from multiple linear regression, classification and regression trees, and artificial neural networks techniques and demonstrate how prediction accuracies from these models may be increased by utilizing ensemble learning methods.

The 2009/2010 North American ski industry (North American Industry Classification System 71392) season counted nearly 60 million skier days, representing an approximate $16.305B industry (Mintel Marketing Database, 2010). As illustrated in Exhibit 2.1, this mature industry, is characterized by limited skier day growth, with only a 1.374% compounded annual growth rate over the last thirty years. As the 2007/2010 economic recession eroded consumer discretionary income

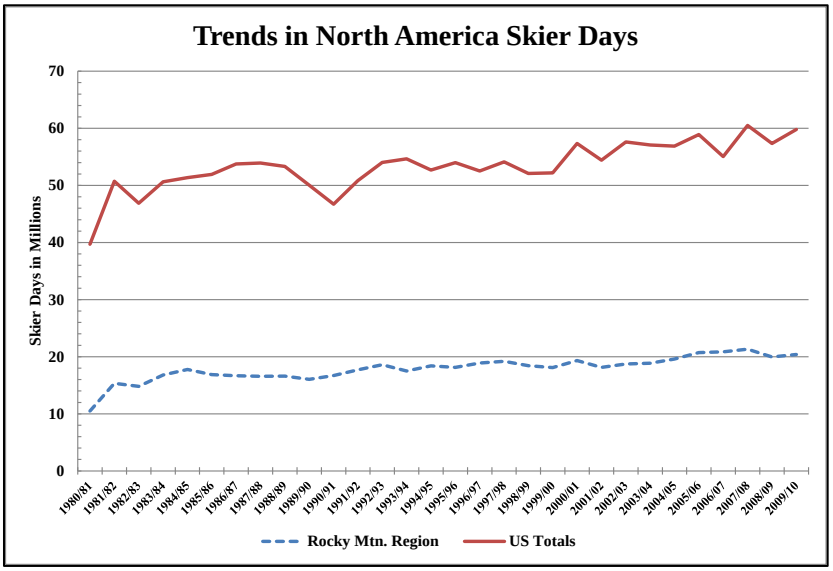


Exhibit 2.1. Trends in North American Skier Days

(www.nsaa.org), competition within the skiing industry became even more aggressive. To sustain a long-term competitive advantage, individual ski resorts must provide superior experiences, high quality ancillary services (e.g., food services, lodging and sleigh rides) and year round outdoor activities all of which are predicated on accurate skier days and ancillary services estimates

(Clifford, 2002.)

The remainder of this paper is organized as follows. Section 2 provides a literature review and overview of ensemble methods. Section 3 discusses the unique contributions of this work. The method and research design implementations are described in section 4. Section 5 provides a detailed discussion of the research results while section 6 presents managerial implications and future directions.

2. Literature review

The following section provides an overview of tourism forecasting research and the subsequent section presents background information on ensemble learning methods.

2.1. Related research

A significant theme of leisure or hospitality research published over the last two decades is the application of a wide array of forecasting techniques, such as time series analysis, econometric modeling, machine learning methods, and qualitative approaches for modeling tourism demand (Song and Li, 2008). Several comprehensive survey articles concentrating on tourism demand modeling have been published, each providing coverage of the forecasting method(s) utilized by the cited authors (Song and Li, 2008; Li, et al., 2005; Lim, 1999). While there is limited research on tourism demand forecast combination or ensemble learning methods contained in these survey articles, Song, et al. (2009) as well as Oh and Morzuch (2005) make excellent cases for combining tourism demand forecasts. Song, et al. demonstrated that a single forecast formed by averaging a set of forecasts for inbound Hong Kong tourism will, by definition, be more accurate than the least accurate forecast, thus mitigating some forecasting risk. Oh and Morzuch (2005) provided a similar argument by illustrating how a forecast created by combining several time series forecasts for Singapore tourism outperformed the least accurate forecast and, in some situations, was more accurate than the most accurate individual forecast.

Table 2.1 is a concise list of highly cited tourism forecasting articles that apply MLR, CART or ANN modeling techniques and is indicative of the limited nature of current academic literature with

a micro economic research focus. Also note that Table 2.1 contains only five prior articles related to skier days forecast, with two articles utilizing MLR and none applying ANN, CART, or ensemble techniques. This is also indicative of the limited availability of ski resort management research. The present study addresses this gap in research where the bulk of existing research emphasis is on classification models and not prediction modeling. To the best of our knowledge, this is the only research applying ensemble methods in skier days forecasting.

Author	***Forecasting Method***	***Forecast Target***
Uysal and Roubi, 1999	*Multiple regression, ANN*	*Tourist arrivals, Canadian inbound to U.S., aggregate*
Law, 2000	*ANN*	*Tourist arrivals, inbound to Taiwan, aggregate*
Burger, et al., 2001	*ANN, moving average, multiple regression, ARIMA*	*Tourist arrivals, inbound to Durban South Africa, aggregate*
Tan, et al., 2002	*Multiple regression, economic models*	*Tourist arrivals, inbound to Indonesia, Malaysia, aggregate*
Cho, 2003	*Exponential smoothing, ARIMA, ANN*	*Tourist arrivals, inbound to Hong Kong, aggregate*
Hu, et al., 2004	*Moving average, multiple regression, exponential smoothing*	*Restaurant customer arrivals, Las Vegas, U.S., local*
Kon and Turner, 2005	*ANN, exponential Smoothing, basic structural method*	*Tourist arrivals, inbound to Singapore, aggregate*
Naude and Saayman, 2005	*Multiple regression*	*Tourist arrivals, inbound to South Africa, aggregate*
Pai and Hong, 2005	*ANN, ARIMA, SVM*	*Tourist arrivals, inbound to Barbados, aggregate*
Patsouratis, et al., 2005	*Multiple regression, economic models*	*Tourist arrivals, inbound to Greece, aggregate*
Palmer and Montano, 2006	*ANN*	*Travel tourism, inbound to Balearic Islands, aggregate*
Chen, 2011	*Linear, nonlinear statistical models*	*Tourist arrivals, outbound from Taiwan, aggregate*
Shih, et al., 2009	*Multiple regression*	*Skier days, inbound to Michigan, U.S., local*
Hamilton, et al., 2007	*Multiple regression, ARMAX*	*Skiers days for New England ski resorts*
Riddington, 2002	*Learning curve, time varying parameter*	*Skier days, outbound to Europe from U.K., aggregate*
Perdue, 2002	*ANOVA, economic models*	*Skier days, inbound to Colorado, U.S., local*
Pullman and Thompson, 2002	*Multiple regression*	*Skier days, inbound to Utah, U.S., local*
This research	*Multiple regression, ANN, CART, ensembles*	*Skier days, inbound to Utah, U.S., local*

Table 2.1. Related Research

2.2. Ensemble methods overview

There is extensive supporting literature for the use of data mining techniques with respect to the

leisure industry. However, there is substantially less research advocating ensemble learning techniques such as bagging, boosting, random subspace, and stacked generalization; all of which offer some of the most promising opportunities for development and refinement of leisure demand estimation (Chen, 2011).

Boosting was developed by Schapire (1990) and is one of the most popular and powerful forms of ensemble learning. Based on data resampling, classification or prediction models are successively created starting from a weak model and then misclassified observations or inaccurate predictions are given more weight for the next model generation iteration (Schapire, 1990).

In 1995 Dietterich and Kong published a seminal article that provided much needed supporting theory for the superior performance of ensemble learning over a single classifier by adapting statistical bias-variance decomposition to ensemble learning (Dietterich and Kong, 1995). In 1996, Freund and Schapire developed AdaBoost, a significant refinement of the original boosting algorithm, with extensions for multinomial classification and ratio data prediction problems (Freund and Schapire, 1996). Bootstrap aggregation or bagging is one of the most widely used ensemble learning techniques because of ease of implementation, low model complexity and comparative high levels of learning accuracy. *N* bootstrap replicas of the training data set are created and trained. For classification, a majority vote is taken to determine the winning class for each observation. Averaging is used for numeric prediction (Breiman, 1996). Similar to bagging, Major and Ragsdale introduced weighted majority aggregation which assigns different weights to the individual classifiers with respect to their votes (Major and Ragsdale, 2000; Major and Ragsdale, 2001).

Stacked generalization or stacking is one of the earliest hierarchical methods of ensemble learning (Wolpert, 1992). Typically a two-tier learning algorithm, stacking directs the output from different types of prediction or classification models, (e.g., ANN combined with Naïve Bayes), and then applies these outputs as inputs to a meta learning algorithm for aggregation. Stacking has been empirically shown to consistently outperform bagging and boosting, but has seen the least academic research (Witten, et al., 2011).

Breiman, Dietterich and Schapire all include a basic source of randomness in their ensemble

algorithms to create diversity among the ensemble set. Ho (1998) introduced an additional technique to add model diversity with the random subspace ensemble method, where a different subset, or feature selection, of the full feature space is used for training each individual machine learner. A synthesis of this literature indicates that ensemble machine learning methods can be generally grouped by the method that individual classifiers are created or the method used to aggregate a set of classifiers.

3. Research contribution

The motivation behind this research is the development of a more effective and objective method for estimating skier days for ski resorts by applying ensemble learning methods to MLR, CART and ANN forecasting models. As detailed in Table 2.1, tourism demand is modeled at a macro economic (i.e. "aggregate", multiple ski resort) level in the majority of the articles, whereas this research makes a contribution by modeling skier days at a micro economic (i.e. "local", single ski resort) level, thus providing a more customized forecast to resort management similar to Hamilton, et al., (Hamilton, et al., 2007). Determining the most appropriate forecasting strategy for a specific service organization is a top level decision that helps match available capacity with customer demand. Owing to the nature of services, capacity planning for many leisure organizations is often more difficult than for manufacturers, which is certainly the case for the ski resort industry (Mill, 2008). Manufacturers can manage capacity by analyzing long-term forecasts and respond by building inventory buffers as needed. In contrast, service organizations must quickly react to weekly and daily demand variations and on occasion, to time of day volatility, without the benefit of an inventory buffer.

In fact, forecasting the daily volatility in demand is a crucial business problem facing most North American ski resorts and thus, a tactical or operational issue. Forecasting strategic long-term skier days demand, as illustrated in Exhibit 2.1, on the other hand, is rather straight forward since it has been flat for approximately twenty years (www.nsaa.org). Both daily and weekly skier days forecasts are essential inputs for operational decisions, (e.g., the number of lifts to operate, the required number of lift attendants, which slopes to groom and the level of ski rental staffing), that attempt to balance a resort's lift capacity with downhill skier density.

Table 2.2 illustrates the three principal operational metrics for four prominent ski resorts located in the United States, in the state of Utah. These four resorts are direct competitors and all share adjacent ski area boundaries. Lift capacity per hour and skiable terrain are typically stated at their maximum and are frequently quoted as a competitive advantage in promotional material. In contrast, skier density, which is calculated by dividing lift capacity per hour by skiable terrain, represents the potential of overcrowding and is rarely publicized (Mills, 2008). As shown in Table2.2, Solitude Mountain Resort, discussed in more detail later in the paper, has a much higher skier density metric than its three direct competitors. Solitude Mountain Resorts takes pride in its lift capacity, however the resort must consider the negative impact of possible overcrowding the skiing terrain and over utilization of the base area amenities. It follows that more efficient utilization of resources and improved capacity planning can drive higher skier satisfaction and thus higher revenues (Stevenson, 2012).

Resort Name	Lift Capacity Skiers Per Hour	Skiable Terrain in Acres	Skier Density
Snowbird Ski Resort	17,400	2,500	6.96
Solitude Mountain Resort	14,450	1,200	12.04
Alta Ski Lifts	11,248	2,200	5.11
Brighton Ski Resort	10,100	1,050	9.62

Table 2.2. Ski Area Direct Competitor Comparison

This research primarily takes a data mining perspective and focuses on improving the prediction accuracy on *new* observations while acknowledging the classic statistical goal of creating a good explanatory model. Previous empirical research (Law, 2000; Palmer, et al., 2006; Cho, 2003) has consistently shown the improved demand prediction accuracy of ANN when compared to traditional multivariate modeling techniques such as MLR and CART. Several ensemble learning methods are subsequently applied to the MLR, CART and ANN models and are shown to improve predictive accuracy. This research specifically demonstrates the advantages of ensemble methods when compared to the results from a single prediction model.

4. Methodology

The following section discusses the independent and dependent variable selection process, data set characteristics, base classifiers, and ensemble methods used in our analysis.

4.1. Initial variable set

The dependent variable in this research project is skier days, an industry standard attendance metric that all ski resort managers assess on a daily basis throughout their ski season. Exhibit2. 2 illustrates mean skier days (with 90^{th} and 10^{th} percentiles) by day of week for seasons 2003 to 2009 for Solitude Mountain Resort in Utah (www.skisolitude.com). Solitude Mountain Resort is an award-winning medium size resort with respect to ski terrain and total skier days and is world renowned for its consistent and abundant snowfall. While still relatively flat compared to other leisure activities, the state of Utah experienced a 2.88% compounded annual growth rate (CAGR) in skier days over the same time period covered by Exhibit 2.1, which is more than double the national CAGR. According to the Utah Governor's Office of Planning and Budget, consistent snowfall, relatively moderate pricing, and ease of access to resorts are the primary drivers for the state's consistent growth in skier days (http://governor.utah.gov/DEA/ERG/2010ERG.pdf , 2010).

Several forms of the skier days dependent variable were utilized in the exploratory phase of our model building. Dependent variables representing two, three, four, and five day leading skier days were generated by shifting forward the actual skier days value by each of these specific lead amounts. For example, in a two day leading skier days model, a record of independent variables for a Monday would contain the actual skier days (dependent variable) from two days forward, i.e., Wednesday. These four outlook horizons (i.e. two, three, four and five days in the future) are explored because ski resort managers can benefit from an operational planning horizon longer than the one day afforded by a next day prediction (Mills, 2008; King, 2010).

Exhibit 2.2 also shows that daily skier days follow a weekly cyclical pattern, along with high variability for each day of the week. With this complex skier days demand pattern, one can easily understand how difficult accurate skier days estimation and planning activities are for ski resort management (Clifford, 2002). A review of the literature and personal discussions with ski resort management at several North American resorts supports the premise that skier days, as in most consumer demand scenarios, is a function of economic variables, along with weather-related drivers, and a set of control variables that model specific contextual phenomena (Shih, et al., 2009).

An extensive list of possible independent variables was explored, resulting in an initial set of independent variables outlined in Table 2.3. In the exploratory phase, there were 23 independent variables and 2 interaction variable combinations. The use of this initial set of independent variables is supported by previous research (Pullman and Thompson, 2002; Hamilton, et al., 2007; Shih, et al., 2009; Chen, 2011) and includes several independent variable recommendations by Solitude Mountain Resort management. One possible limitation of this research is the different measurement time periods of the economic variables. While different measurement scales are not ideal, the variable selection methodology employed by subsequent ensemble analysis will determine if the potential explanatory benefits provided by these variables are significant.

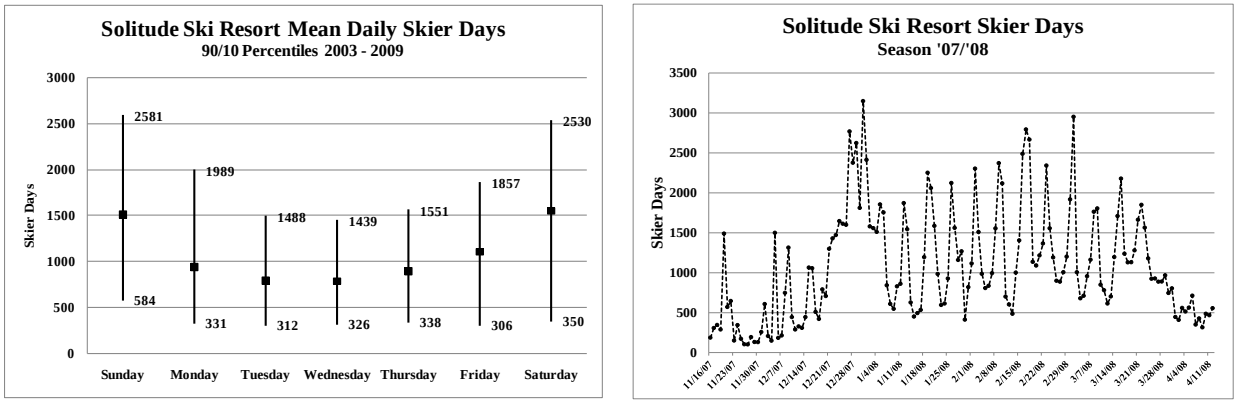


Exhibit 2.2. Skier Days Patterns

The management team from Solitude Mountain Resort provided the skier days data over the research time frame. The resort also provided climate related data for year-to-date snowfall, current snow depth, daily new snowfall measurements, and mean daily temperature. Year-to-date snowfall is defined as the cumulative snowfall from the season opening date up to and including a specific day of the season. Current snow depth is defined as the unpacked snow depth at a mid-mountain resort location within a restricted area for a given date. Exhibit 2.3 provides a comparison of skier days versus current snow depth and lends support to the explanatory value of current snow as an independent variable in this research. New snow fall is measured over a twenty-four hour lag, from the previous settled snow depth. These measurements are in inches and are based on National Oceanic and Atmospheric Administration suggested guidelines, although actual resort practices can be subjective at times. The mean daily temperature in Fahrenheit for each observation is calculated by averaging a one day lag of the maximum and minimum daily temperature reading collected by

the Powderhorn measurement station located within the resort and disseminated by the Utah State Climate Center.

Independent Variable	Variable Context
Year To Date (Y.T.D.) Snowfall	
Current Snow Depth	
New Snow Fall	Weather
Average Daily Temperature	
Avg. Daily Temp. x New Snow Fall	
Avg. Daily Temp. x Current Snow Depth	
Average National Airfare	
U.S. Unemployment Rate	
U.S. Consumer Price Index	Economic
U.S. Consumer Sentiment Index	
Gas Prices Rocky Mountain Region	
Day of Week Indicator	
Season Indicator, Current Day	
Holiday Indicator, Current Day	
Leading Holiday Indicator, Outlook Day	Time Dimension
Leading Season Indicator, Outlook Day	
Day Number of Season	
Squared Day of Season (Quadratic term)	

Table 2.3. Initial Independent Variable Set

Five economic factors were included in the initial independent variable set as follows: average national airfare, prices for retail grade gasoline in the Rocky Mountain region of the U.S, the U.S. employment rate, the U.S. Consumer Price Index (CPI), and the Consumer Sentiment Index (CSI). The average national airfare, defined as the mean price of an economy class ticket, is calculated on a quarterly basis by the U.S. Department of Transportation and was included in the pool of potential independent variables using a three month lag (http://www.rita.dot.gov/, 2012). The weekly mean retail gasoline price for the Rocky Mountain region, provided by the U.S. Department of Energy, reflects period ground transportation cost (http://www.eia.gov/, 2012). The U.S. Unemployment Rate functions as a possible proxy for current income and future discretionary income. The CPI reflects monthly changes in price levels for a consumer market basket of goods. The CPI contains a leisure activity expenditure component, which varies depending on available discretionary income. Thus, a sudden increase in the U.S. Unemployment Rate or the CPI, *ceteris paribus*, shifts discretionary income away from leisure expenditures (McGuigan et al., 2008). The CSI, calculated monthly by the University of Michigan, is a measurement of perceived consumer control over his or her current economic state and in this research project functions as a proxy for current and future income.

Indicator variables, which measure possible categorical time effects, are included in the initial

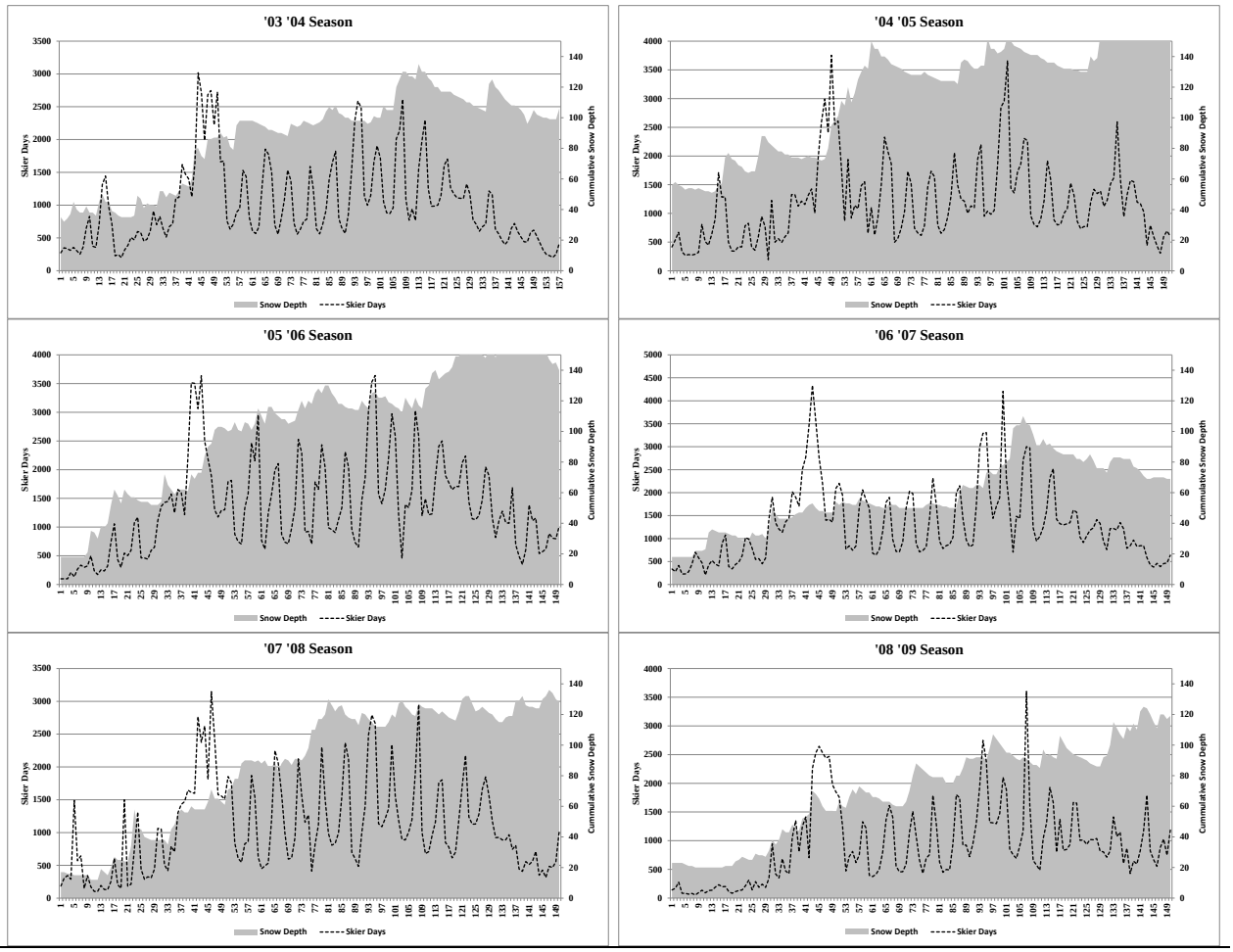


Exhibit 2.3. Cummulative Snow Fall and Skier Days by Day of Season

variable set because most ski resorts experience large fluctuations in skier days during their ski season. Six binary variables are used to model the seven days of the week. Two binary variables are used to model the three ski seasons: early, regular, and spring. One binary variable is required to model the combined holiday periods of Christmas week including New Year's Day and President's Day Weekend which includes Friday and Monday. An additional binary variable is needed to model whether the outlook (prediction) day actually falls in the holiday period. For example, if today is December 20th ("Holiday Type" = 0, for current day), and we are predicting 5 days ahead ("5 day lead"), then the outlook (prediction) day is December 25th (which sets "5 Day Lead Holiday Indicator" = 1, since the target day is Christmas). Similarly, two binary variables are needed to model which ski season (early, regular, and spring) the outlook (prediction) day falls into. The logic behind the use of these outlook variables is that the attributes of the target or outlook day are known

in advance by skiers and are probably taken into consideration in their skiing decision. Lastly, the numerical day of the season and its square are used to model the nonlinear trend aspects of the data set.

4.2. Data set description

The data set represents Solitude Mountain Resort's ski seasons from 2003-2004 until 2008-2009 and contains 908 skier days observations. Solitude Mountain Resort operates an automated RFID terrain access system which controls skier entry to the slope system and also compiles skier days and other tracking metrics. Although not a strict time series data set because of time gaps between ski seasons, it is a representative data sample that covers several different economic time periods (Pullman and Thompson, 2002). The data set is large enough that it easily meets the ratio of observations to independent variables requirement of 129 (104+k) observations where k is the number of independent variables (Tabachnick, et al., 2000). A representative sample of the data set is shown in Exhibit 2.4 which provides an illustration of how a specific record contains continuous, integer and categorical variables.

Data Point	D.O.W.	Date	Day of Season	Quadratic Term	Lagged Airfare	Weekly Gas Price	Unemp Rate	CPI	CS1	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Holiday Type	5 Day Lead Holiday Indicator	Early Season	Regular Season	5 Day Lead Early Season Indicator	5 Day Lead Regular Season Indicator	New Snow	YTD Snow Falls	Cum Snow Depth	Avg Daily Temp	Skier Days	5 Day Lead Skier Days
292	Thur	3/31/05	140	19600	297.28	2.17	5.20	193.10	92.60	0	0	0	1	0	0	0	0	0	0	0	0	15	652	204	43	1551	793
293	Fri	4/1/05	141	19881	301.39	2.17	5.20	193.70	87.70	0	0	0	0	1	0	0	0	0	0	0	0	0	652	198	42	1196	594
294	Sat	4/2/05	142	20164	301.39	2.17	5.20	193.70	87.70	0	0	0	0	0	1	0	0	0	0	0	0	0	652	192	43	1167	434
295	Sun	4/3/05	143	20449	301.39	2.17	5.20	193.70	87.70	0	0	0	0	0	0	0	0	0	0	0	0	0	652	185	41	1025	306
296	Mon	4/4/05	144	20736	301.39	2.20	5.20	193.70	87.70	1	0	0	0	0	0	0	0	0	0	0	0	0	652	182	41	442	596
297	Tues	4/5/05	145	21025	301.39	2.20	5.20	193.70	87.70	0	1	0	0	0	0	0	0	0	0	0	0	14	666	192	43	793	695
298	Wed	4/6/05	146	21316	301.39	2.20	5.20	193.70	87.70	0	0	1	0	0	0	0	0	0	0	0	0	0	666	187	39	594	600

Exhibit 2.4. Partial Data Set Example

4.3. Multiple regression

Given an initial set of possible independent variables, there are several search methods for developing a multiple regression model. This section discusses the steps followed to select a set of final independent variables from the exploratory independent variable pool. The selection of independent variables is an important consideration because there is typically a tradeoff between developing a complex multiple regression model that explains most of the variation in the dependent variable and a more parsimonious model that is easier to operationalize in an applied business setting (Ott and Longnecker, 2001).

Three widely used variable selection heuristics in business analytics and data mining are stepwise regression, forward selection and backward elimination. These three search methods use slightly

different search logic and typically develop somewhat similar models. Stepwise regression starts with one predictor and adds or removes predictors in a step-by-step method based on a model fit criteria. Forward selection starts with one predictor that has the largest *p* value and adds predictors one at a time, given a minimal entry criterion. Once a predictor is included in the model, it is never removed.

Backward elimination, the search method applied in this research, is essentially a compromise between the two previously discussed methods. This algorithm starts with the full model and determines whether there are any nonsignificant predictor variables present and, if so, the predictor with the smallest nonsignificant *p* value is dropped from the model. The algorithm cycles again, until all independent variables are significant at a predetermined alpha level.
The model building process was started by creating an additive model containing all 23 predictor variables plus 2 interaction terms. The interaction terms are Average Daily Temperature x Current Snow Depth and Average Daily Temperature x New Snow Fall (as we assumed that average daily temperature interacts with or confounds the current snow level, and possibly causes snow fall to accumulate at different rates). A backward elimination regression was performed using the 25 independent variables illustrated in Table 2.3 for each of the four variants of the skier days dependent variable. The attributes were normalized, using a standard Z transformation, to allow beta coefficient comparison and ranking.

4.4. Artificial neural networks

An ANN is a nonlinear statistical modeling tool used in classification or prediction problems as well as other data mining applications (Yi and Prybutok, 2001). ANNs originated in the field on artificial intelligence and are all-purpose mapping functions that excel on highly nonlinear problems and environments (Fausett, L. V., 1994). The connection with artificial intelligence applications is that given a set of input variables, the ANN is able to "learn" through many iterations of trial and error an approximate mapping from a set of input variables to one or more output variables. An example of an ANN, which consists of nodes that are interconnected by weighted links, is shown in Exhibit 2.5.

An ANN mimics the functions of biological neurons, synapses and neurotransmitters. In a nutshell, data is "fired" from the neurons (nodes) along the synapse paths (arcs) with a degree of intensity

produced by the neurotransmitters (weights). Each neuron computes the weighted sum of all the data transmitted to it and computes a response value that is transmitted to one or more other neurons. An *S* shaped (sigmoidal) transfer function is commonly used and allows each neuron to act like a mini nonlinear function that models a small piece of the total problem (Turban, et al., 2007).

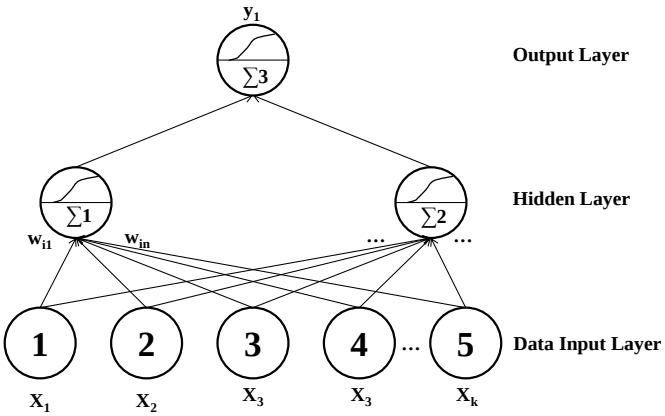


Exhibit 2.5. Artificial Neural Network Architecture

Neural networks can exhibit high accuracy when designed correctly, fast computational speed on new observations once trained, and an intuitive ease of use when implemented with dedicated software packages or Microsoft Excel Add-ins (Paliwal and Kumar, 2009). Due to their interconnected design, neural networks can implicitly model all possible variable interactions (Han and Kamber, 2006). While ANNs have proven to be a powerful predictive tool in numerous business situations, they do have several limitations that discourage their acceptance and implementation. One disadvantage is low explainability, meaning the modeler cannot easily describe how the predictions are derived (Uysal and Roubi, 1999). Additionally, ANNs require a large non-sparse data set for training and calibration purposes (Fausett, 1994).

Feed forward multi-layer ANNs, as illustrated in Exhibit 2.5, along with back propagation supervised learning are widely used for classification and prediction problems (Fausett, 1994; Law, 2000). The back propagation learning algorithm is a gradient descent method that trains the ANN in an iterative manner by minimizing prediction error, using the partial derivatives of the error function for each hidden node (Das, et al., 2009).

The real power of an ANN comes from the addition of a hidden layer and its ability to model

complex interactions such as the Exclusive Or (XOR) relationship. Without a hidden layer an ANN would basically function like a linear multiple regression model. Typically only one hidden layer is required by a back propagation ANN to model any nonlinear function (Fausett, 1994). However, there are issues that a modeler must consider associated with the actual number of hidden nodes to include in an ANN. If there are too few hidden nodes in the hidden layer, the overall network cannot adequately learn or fit the training patterns. If there are too many nodes in the hidden layer, the network can over-fit the training pattern, inhibiting the ability of an ANN to generalize to new data.

A widely used heuristic for determining the number of hidden nodes is to add 1 to the number of model inputs (Das, et al., 2009). Another popular hidden node guideline, available as the default value in several data mining software packages, is to add the number of input variables to the number of output variables and divide by two (Witten, et al., 2011). The interaction variables listed in Table 2.3 are not needed because ANNs implicitly model all interaction variable combinations. Thus, the guideline of adding the number of input nodes plus one as suggested by Das, et al equals 24 nodes for this research (Das, et al., 2009), while the guideline provided by Witten, et al., of adding the number of inputs nodes to the number of output nodes and dividing by two equals 12 nodes. Exploratory ANN models were created based on each leading day dependent variable and using both a 12 and 24 node hidden layer. We applied ten-fold cross validation while training an ANN, thus 90% of the data set is available during each validation fold. All model results were very similar, and the decision was made to side with the concept of parsimony and follow the Witten, at el. guideline for 12 hidden nodes. As a result, the ANN model in this research is fully connected, without any recursive node connections, and takes a 23:12:1 architecture, meaning that it has 23 input nodes for the independent variables, 12 nodes in 1 hidden layer and 1 output node for the dependent variable. The default modeling parameters, such as the learning rate (=0.3), momentum rate (=0.2), and number of epochs (=500), where used.

4.5. Classification and Regression Trees

A classification and regression tree is a specific form of decision tree that can be used for both classification of a categorical variable and prediction of a continuous variable (Nisbet, et al.; 2009). Breiman, et al., developed CART in 1984 as method of classifying high risk heart patients for the University of California, San Diego Medical Center. This seminal work introduced structured tree-

based modeling to the statistical mainstream audience. CART is one of many popular decision tree modeling algorithms which include CHAID, ID3, C4.5 and C5.

A decision tree is a data structure consisting of splitting nodes, decision branches, and leaf nodes that uses a hierarchical sorting process to group records from a data set into increasingly homogeneous groups. The decision tree algorithm recursively partitions the data set based on a goodness of fit metric derived from each attribute contained in the data set where the attribute with the best goodness of fit metric is used as the root node of the tree. Several decision tree algorithms support multinomial splits, while CART is restricted to binary splits. There are two widely used goodness of fit metrics for ranking attributes that best partition a data set. Both metrics, Information Gain (Hunt, et al., 1966) and the Gini Index (Breiman, et al., 1984), measure the change in homogeneity of a partitioned set, however they have quite differing origins. Information Gain was created as an entropy measurement from information theory while the Gini Index was developed as an economic measure of population diversity. The sum of squared error is typically used for partitioning continuous attributes.

There are four general steps that must be completed to construct a tree using CART. The first step is to compute a goodness of fit metric on each value from each attribute and rank each attribute by the increase in homogeneity or purity of each partition. After the ranking is complete, the attribute with the largest increase is designated as the root node attribute. The second step is to continue partitioning the data set into a maximal tree with the goal of producing pure final nodes, also called leaf nodes. Other stopping conditions for tree growing are stopping at a maximum level and stopping when a stated fraction of observations are contained in each leaf. The third step involves pruning some of the previous tree growth to prevent overfitting. The last step is an optimization process that requires the use of cross validation techniques to compare training and testing data set accuracy. When pruning back a level, the accuracy of both the training and testing data sets are compared and pruning continues until the best accuracy of the testing data set is found. One interesting advantage of decision trees is that the partitioning process automatically performs attribute ranking and selection for the researcher. The root node attribute provides the most discriminating power when partitioning the data set and subsequent tree level nodes provide decreasing discriminating power. Decision trees bring valuable diversity to ensemble models

because slight changes to the data set can create significantly different predictive models.

4.6. Ensemble methods

The predictive accuracy of individual forecasting methods can be improved by applying numerous ensemble learning strategies developed over the last several decades, such as voting methods, bagging, random subspace, and stacking. This research compares the predictive accuracy of these four well known ensemble methods with those of individual MLR, CART and ANN models. These four ensembles where selected based on their extensive research streams as well as being representative of common themes synthesized from several taxonomies reflected in the literature (Kuncheva, 2004; Kantardzic, 2011; Witten, et al., 2011; Rokach, 2009).

This research further generalizes and provides a concise taxonomy of ensemble methods that consists of four ensemble grouping dimensions as depicted in Table 2.4. The training set sampling dimension refers to how a data set is sampled to create training and test data sets, while the attribute subset sampling dimension describes how to sample and construct variable or feature subsets. The algorithm combination dimension focuses on whether one type or multiple types of prediction or classifier models are used for the ensemble, and the decision combination dimension details how outputs or decisions from the models are combined. This research uses the representative ensembles as described in Table 2.4.

Ensemble Dimension	**Methods**	**Ensembles**
Training set sampling	Boot strap, disjoint stratified	Bagging
Attribute subset sampling	Input space subsets	Random subspace
Algorithm combination	Single or multiple case(s)	Stacking
Decision combination	Fusion, selection	Vote

Table 2.4. Ensemble Taxonomy

Boot strap aggregation, more commonly known as bagging, is a scheme that takes the original data set *D*, randomly partitions it into training D_{Train} and test D_{Test} sets and creates new training data sets the same size as the original data set D_{Train} by resampling with replacement (Breiman, 1996). Bagging has been shown to effectively improve prediction accuracy in numerous empirical studies (Kim and Kang, 2010; Sun, et al., 2011; Das, et al., 2009; Kim, 2009) when weak base prediction models are utilized. Weak or unstable prediction models, such as ANN, decision trees, and subset selection regression, show large changes in predictions or classification

results when small changes to training data sets occur (Breiman, 1996). There are several parameters that must be set for bagging. In this research, the boot strap sample ratio for testing is the standard 33% and the number of individual prediction models in the ensemble set is set to 25. The same model architectures developed for the MLR, CART and the ANN analysis are utilized as model inputs for the bagging algorithm. The pseudo-code is shown in Exhibit 2.6.

Input:	
Data set $D = [(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)]$	X_n attribute vectors, n observations, y_n predictions.
Base classification algorithm, d	Define base learning algorithm.
Ensemble size, S	Number of training loops.
For $s = 1$ to S	
D_s = BootstrapSample(D)	Create bootstrap sample from D.
d_s = CreateBaseLearnerModels(D_s)	Create base models from bootstrap samples.
Make model prediction d_s	
Save model prediction d_s	
End	
Output:	
Average S model predictions	Combine model outputs by mean.
Return ensemble prediction	

Exhibit 2.6. Boot Strap Aggregation Pseudo-code

Random subspace is a training set creation strategy based on stochastic discrimination theory that takes random samples of the feature or variable set to construct new training sets of a predetermined size for each prediction model included in the ensemble (Ho, 1998). For example, if a training data set contains five variables, V1, V2, V3, V4, and V5, several equal sized subsets can be created such as {V1, V2, V3}, {V2, V3, V4}, { V1, V4, V5} and so forth. As previously discussed, the accuracy improvement provided by ensemble methods as compared to a single prediction model, is the direct result of randomization and diversity provided by weak prediction models. Stable prediction models or classifiers generally do not benefit from ensemble methods because their outputs are insensitive to training set sampling methods. The random subspace technique takes this issue into consideration and works by randomizing the feature set as opposed to randomizing individual data set observations. This strategy can increase the ensemble accuracy by creating diversity when training each model on a different feature subset. Ho indicates that good accuracy is achieved when the subset size is equal to roughly half the original feature set size (Ho, 1998). In this research, the number of prediction models contained in the ensemble is 25, the same as bagging. The final number of independent variables selected by the backward elimination regression analysis will determine the

subset size parameter. The pseudo-code is shown in Exhibit 2.7.

Input:	
Data set $D = [(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)]$	X_n attribute vectors, n observations, y_n predictions.
Ensemble size, S	Number of training loops
Subspace dimension, A	Number of attributes for subspace
For $i = 1$ to S	
SS_i = CreateRandomSubSpace(D, A)	Create new variable set for training input
d_s = CreateBaseLearnerModels(D_s)	Create base models from bootstrap samples
Make model prediction d_s	
Save model prediction d_s	
End	
Output:	
Average S model predictions	Combine model outputs by mean
Return ensemble prediction	

Exhibit 2.7. Random Subspace Pseudo-code

Stacked generalization or stacking is an ensemble method that combines classifiers of different types, in contrast to combining the results of many classifiers of the same type such as bagging. It is a hierarchical approach where the outputs from base prediction models are used, along with the original correct predictions or class tags as inputs to a higher level model known as a meta learner. The meta learner learns which set of base level prediction models provides the most accuracy and determines the weight of each model to aggregate into a final prediction. In this research the base level models used in the stacking ensembles are the MLR, CART and ANN models. Meta learner model selection is difficult to determine and even described as a "black art" by David Wolpert in his seminal work (Wolpert, 1992). Wolpert indicated that a simple linear model such as linear regression works best at the meta learner classification or prediction task because the majority of the prediction or classification efforts is completed by the base learner models (Wolpert, 1992). Ting and Witten suggested that an ANN could also be an effective base learner model (Ting and Witten, 1999). In this research, the MLR, CART and ANN models are each embedded as a meta learner, resulting with three instances of a stacking ensemble model. The pseudo-code is shown in Exhibit 2.8.

Voting is one of the most intuitive and simplest methods of combining prediction model outputs. When confined to a literal definition, voting is not an ensemble method since the scheme does not contribute to base model generation; however, voting is typically used as a baseline measure for ensemble comparisons because it is independent of both the data and base models. For a categorical dependent variable, a plurality vote (frequently mistaken as a majority vote) or the class with the

highest count is the ensemble output. Weighted voting, a more general voting rule, allows a modeler

Input:	
Data set $D = [(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)]$	X_n attribute vectors, n observations, y_n predictions.
First level classification algorithms, $d_{1 \ldots S}$	
Second level meta learner, d_{2nd}	
For $s = 1$ to S	
d_s = CreateFirstLevelModels(D)	Create first level models from data set D.
End	
$D_{New} = 0$	Start new data set creation.
For $i = 1$ to n	
For $s = 1$ to S	
$C_{is} = d_s(X_i)$	Make prediction with classifier d_s
End	
$D_{New} = D_{New} \cup [((C_{i1}, C_{i2}, \ldots, C_{iS}), y_i)]$	Combine to make new data set.
End	
$d_{Trained2nd} = d_{2nd}(D_{New})$	Train meta model to new data set.
Output:	
Return ensemble prediction = $d_{Trained2nd}$	Ensemble prediction.

Exhibit 2.8. Stacked Generalization Pseudo-code

to place more weight on models that have higher accuracy, based on prior domain knowledge, and thus lessen the chance of selecting an inaccurate model (Major and Ragsdale, 2001). For a continuous dependent variable, algebraic operations such as the mean, median, weighted mean, minimum, maximum, and product can by utilized as the combination rule (Witten, et al., 2011). As previously discussed, when independent prediction models, each with accuracy better than a random guess, are combined using voting, the combined accuracy increases as the number of prediction models increase (Han and Kamber, 2006). The pseudo-code is shown in Exhibit 2.9.

The three single model architectures developed in previous sections (i.e., MLR, ANN and CART) were used with the four ensemble methods (i.e., bagging, random subspace, stacking, and voting) as discussed in Section 4.6., resulting in ten specific ensemble implementations as depicted in Exhibit 2.10. Specifically, there are three bagging instances, three random subset instances, three stacking instances and one voting instance. All ten ensembles applied a hold out sample validation method, reserving 33% of the data for the testing set. Each ensemble model instance was cycled ten times and the average root mean squared error (RMSE) and R^2 on the hold out samples were calculated for model comparison (Witten, et al., 2011).

The bagging and random subspace ensemble methods both require an instance to be created from each of the single models, MLR, ANN and CART. However, stacking and voting ensembles are

created somewhat differently. Three stacking instances were configured with MLR, ANN and

Input:	
Data set $D = [(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)]$	X_n attribute vectors, n observations, y_n predictions.
First level classification algorithms, $d_{1 \ldots S}$	
Second level meta learner, d_{2nd}	
For $s = 1$ to S	
d_s = CreateFirstLevelModels(D)	Create first level models from data set D.
End	
Output:	
Average S model predictions	Combine model outputs by mean
Return ensemble prediction	

Exhibit 2.9. Voting Pseudo-code

CART models simultaneously as base level predictors with each of one of these learners acting as the meta learner, in turn. The one voting instance was configured with the MLR, ANN and CART models being combined as base level learners.

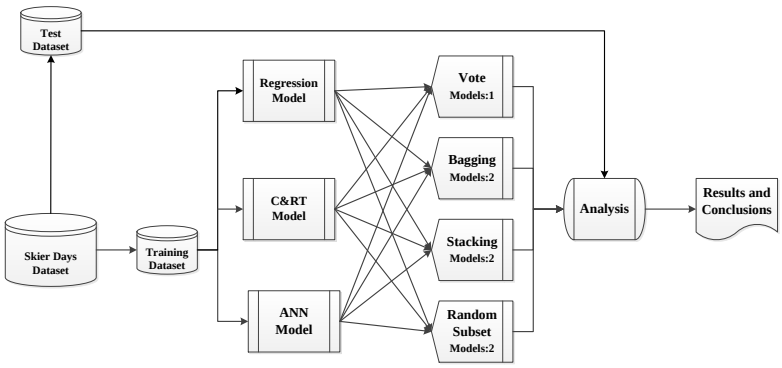


Exhibit 2.10. Conceptual Ensemble Experiment Model

The bagging and random subspace ensembles require, as a parameter, the number of individual base level models to include in the ensemble during runtime. As stated earlier, this research used 25 models per ensemble as suggested by Breiman (Breiman, 1996). Exhibit 2.11 illustrates a 95% prediction interval for the RMSE from a Bagging ensemble of ANNs. Note the downward trend of the RMSE and the beginning of convergence at 25 ensembles. Although these results are not guaranteed, they do support Breiman's findings. The random subspace ensemble technique also requires, as a parameter, the size of the attribute subspace to use for each model contained in the ensemble. As suggest by Ho (Ho, 1998) 50% of the attribute set was included in each ensemble instance utilizing a random subset method. Both the stacking and voting ensembles do not require any specific parameters, other than the parameters required for the previously developed single

model architecture.

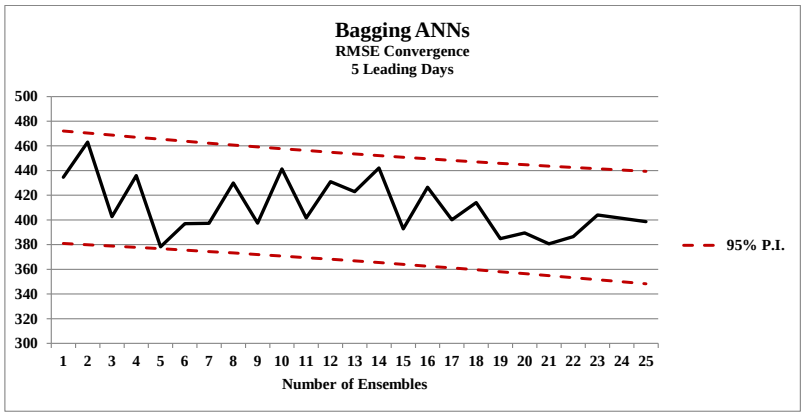


Exhibit 2.11. Ensemble Convergence

While on the surface, these techniques may appear complex, they can be implemented easily using modern software (RapidMiner, SAS Enterprise Miner, etc.), achieve more predictive accuracy than highly specialized and tailored stand-alone prediction algorithms, and thus take less time for setup, configuration and training (Lozano and Acuña, 2011). With respect to their No Free Lunch Theorem, Wolpert and Macready state, "there is no one algorithm that induces the most accurate learner in any domain, all the time" (Wolpert and Macready, 1997). In their article, the authors develop a sophisticated quantitative framework that illustrates the tradeoffs that are typically present and required with machine learning problems. If management or academics are only interested in obtaining the highest possible model accuracy, it may be overly time consuming or impractical to determine a single model that outperforms an ensemble of models. Combining predictive models, with the goal of balancing the weaknesses and strengths from each model, is a very active and promising area for data mining research.

5. Results and discussion

The RapidMiner v5.3 data mining platform was utilized for each single model as well as the ensemble analysis in this research project. RapidMiner is an established and well-respected open source data mining package that is feature rich, user friendly, data stream driven and licensed under the AGPL v3 open source licensing program. The platform provides an extensive collection of data set preprocessing features and machine learning algorithms. RapidMiner is a JAVA implementation

and also provides a broad collection of ensemble learning algorithms. Rapid-I maintains a support site and a wiki community (http://rapid-i.com/content/view/181/190/, 2013).

5.1. *Regression model results*

The multiple regression results, shown in Table 2.5, are based on the mean of 10 cycles through a ten-fold validation as suggested by Witten, et al., (Witten, et al., 2011). It must be noted that the MLR results, based on R^2 and RMSE calculations, appear very similar. These results were somewhat unexpected because, intuitively, it would seem that by increasing the leading days, the prediction accuracy would decrease. These results also initiated a discussion with Solitude Mountain resort management to determine which prediction time horizon is most applicable for their operations. The resort management made the case that a three day leading skier days prediction is the most useful for them. Their justification was predicated on weather report accuracy, ability to mobilize staff, and increased certainty related to operational issues such as vendor support, food deliveries, repair requests and highway snow removal.

MLR Model	**2 Day**	**3 Day**	**4 Day**	**5 Day**
Avg. R^2	0.612	0.620	0.619	0.617
Avg. Root Mean Squared Error	455.909	450.319	450.252	450.611

Table 2.5. Regression Model Results

5.2. ANN model results

A major contribution of this research is extending the limited previous research studies on the benefits of utilizing ANN for consumer demand estimation in a leisure activity setting. As shown in Table 2.6, this research indicates a modest drop in average RMSE for each of the leading skier day models when compared to the same MLR model. Table 2.6 also shows a consistent increase in the predictive power, as measured by R^2, by modeling the skier days data with an ANN.

Model	**2 Day**	**3 Day**	**4 Day**	**5 Day**
Avg. MLR RMSE	455.909	450.319	450.252	450.611
Avg. ANN RMSE	409.818	413.135	418.660	408.004
% Improvement	**10.110%**	**8.257%**	**7.017%**	**9.456%**
Avg. MLR R^2	0.612	0.620	0.619	0.617
Avg. ANN R^2	0.686	0.681	0.671	0.686
% Improvement	**12.092%**	**9.839%**	**8.401%**	**11.183%**

Table 2.6. Comparative Model Results

5.3. Classification and Regression Tree model results

The CART analysis is based on the mean of ten cycles through ten-fold validation. The associated decision tree has nineteen splitting nodes that represent the most discriminating independent variables and the set and order of these independent variables are quite different when compared to the MLR and ANN models. The decision tree has nine levels and terminates with twenty leaf nodes each containing the actual skier days predictions. These results, illustrated in Table 2.7, show improvements in the average RMSE for each CART model over the base MLR models. Similar to both the MLR and ANN results, the R^2 and RMSE results for each of the four CART models are very close, with the 3 Day Leading Skiers Days model having the lowest RMSE. However, the CART models did not show any improvement over the ANN results. These results were expected because CART is a nonparametric machine learning technique and is not restricted to the inherent linearity assumed by MLR, and should perform better in a nonlinear environment. However, the limited granularity imposed by binary splitting prevents the CART analysis from reaching the level of RMSE improvement that the ANN models achieves.

Model	**2 Day**	**3 Day**	**4 Day**	**5 Day**
Avg. MLR RMSE	455.909	450.319	450.252	450.611
Avg. CART RMSE	436.684	431.036	440.741	436.314
% Improvement	**4.217%**	**4.282%**	**2.112%**	**3.173%**
Avg. MLR R^2	0.612	0.620	0.619	0.617
Avg. CART R^2	0.644	0.653	0.635	0.641
% Improvement	**5.229%**	**5.323%**	**2.585%**	**3.890%**

Table 2.7. Comparative Model Results

The improvement in predictive accuracy of both the CART and ANN models supports the argument that the tourism industry could benefit by comparing the accuracies of various predictive models and not relying solely on traditional statistical methods.

5.4. Ensemble model formulation

A primary goal of this research is to investigate whether ensemble methods are effective at "boosting" prediction accuracy for improved skier days estimation. Table 2.8 provides the complete experimental results of the three single prediction methods and includes the results of the ten ensemble methods. Overall, the ensemble methods do show practical improvements in RMSE over the single model techniques. Thus, this research supports the argument that managers should not depend on the predictive results of one base prediction method. Ensemble modeling reduces the risk

associated with selecting an inferior model.

		RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
		2 Day Leading Skier Days		3 Day Leading Skier Days		4 Day Leading Skier Days		5 Day Leading Skier Days	
Single Models	Multiple Linear Regression w/ 10 Fold Cross Validation, Mean of 10 Runs	455.909	0.612	450.319	0.620	450.252	0.619	450.611	0.617
	Regression Tree w/ 10 Fold Cross Validation, Mean of 10 Runs	436.684	0.644	431.036	0.653	440.741	0.635	436.314	0.641
	Neural Network w/ 10 Fold Cross Validation, Mean of 10 Runs	409.818	0.686	413.135	0.681	418.660	0.671	408.004	0.686
Ensemble Models	Random Subspace MLR with Hold Out Sample, Mean of 10 Runs	455.504	0.611	451.557	0.623	456.064	0.622	443.201	0.627
	Random Subspace RT with Hold Out Sample, Mean of 10 Runs	447.316	0.645	418.132	0.666	439.217	0.663	414.058	0.682
	Random Supspace ANN with Hold Out Sample, Mean of 10 Runs	421.307	0.671	420.468	0.669	434.010	0.665	425.569	0.657
	Bagging MLR with Hold Out Sample, Mean of 10 Runs	458.912	0.611	462.623	0.594	454.342	0.621	452.171	0.633
	Bagging RT with Hold Out Sample, Mean of 10 Runs	443.887	0.655	431.667	0.655	435.211	0.658	443.988	0.642
	Bagging ANN with Hold Out Sample, Mean of 10 Runs	414.875	0.636	397.157	0.708	399.635	0.706	384.748	0.721
	Stacking All Models with MLR as meta with Hold Out Sample, Mean of 10 Runs	392.044	0.715	**379.285**	0.724	401.426	0.689	403.604	0.690
	Stacking All Models with RT as meta with Hold Out Sample, Mean of 10 Runs	440.060	0.652	423.443	0.663	432.100	0.651	437.249	0.654
	Stacking All Models with ANN as meta with Hold Out Sample, Mean of 10 Runs	**374.240**	0.712	392.414	0.720	**368.940**	0.732	**364.195**	0.703
	Vote All Models with Hold Out Sample, Mean 10 of Runs	396.635	0.688	384.860	0.697	394.337	0.686	399.704	0.711

Table 2.8. Summary of Average Experimental Results

Another initial observation is that the 5 leading day skier day form of the dependent variable produced the lowest overall RMSE of 364.195 skier days. This result was not expected, and on the surface is not intuitive, because typically the longer the length of time associated with a leading dependent variable, the lower the predictive accuracy. One possible explanation, in the context of recreational skiing, could be that the majority of skiers are out-of-town visitors who use a longer planning horizon (5 days or more vs. 2 days or less) and do not react rapidly to weather changes, due to work commitments, family schedules, or flight reservations that were made far in advance.

A key finding of this research is that for each variant of the skier days dependent variable, an instance of a stacking model generated the lowest RMSE. Stacking models with either a MLR or ANN as the meta learner performed significantly better that the stacking models with a CART as the meta learner. Both the MLR and ANN base models benefit from ensemble techniques, although ensembles with an ANN as the base classifier performed better with the skier days data set. This research also confirms that ensembles of nonparametric prediction models commonly achieve better accuracy than similar ensembles consisting of parametric classifiers because ensembles based on nonparametric prediction models typically exhibit much more model variance, an essential element for increased ensemble accuracy (Kim, 2009).

Exhibit 2.12 illustrates the relative RMSE improvements of the single CART and ANN models, and the ten ensemble model instances when compared to the base MLR model for the 5 day leading skier

days dependent variable. Note that the stacking ensemble with the ANN as the meta learner, which has the lowest RMSE, also gained the highest percentage accuracy increase. Numerous authors have observed (and this research confirms) that the least complex ensemble methods such as voting and bagging often afford impressive improvements in accuracy and argue that these ensembles should always be included in an analysis as a baseline metric (Kim and Kang, 2010). A voting ensemble was included in this research as a benchmark, because of the simplicity and parsimony this specific ensemble provides.

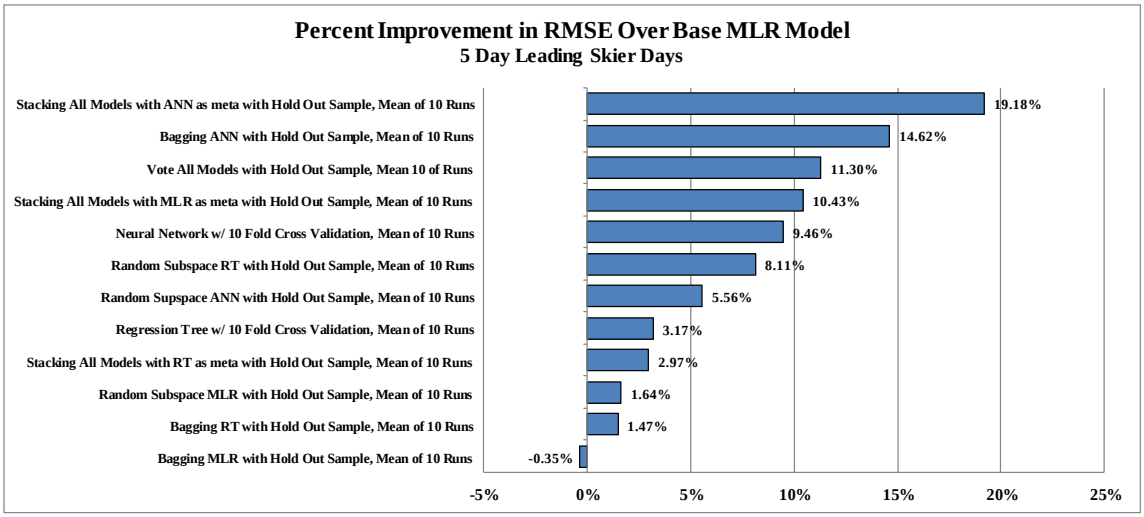


Exhibit 2.12. Ensemble RMSE Improvements Over the Base MLR Model

One final observation drawn from Exhibit 2.12 is that the ensembles created by random subspace performed considerably worse as compared to the other ensembles. Only five out of the twelve random subspace ensembles showed improvements over their respective base model. This result was unexpected because the random subspace ensemble method was actually developed to add diversity (and thus increase accuracy) when utilizing parametric classifiers (Ho, 1998). These results are possibly due to the concise and nonredundant nature of the data set, because the random subset method assumes some level of attribute redundancy to effectively create smaller size attribute spaces (Marqués, A. I., et al., 2012). This research used a feature selection parameter of 50% and as a result, the maximum size of a feature subset is 13, while the minimum size is 7, depending on whether the base learner utilized feature selection.

Overall, while some reduction in prediction error was expected, the considerable reductions in root

mean squared error of the ensemble instances were unexpected, given the previous reduction achieved by the ANN model over the MLR and CART models.

6. Managerial implications and future directions

As the importance of economic contributions from the leisure industry to the United States economy as well as other highly developed economies continues to increase, it is essential that the accuracy of leisure demand estimation continue to improve. This research project makes academic and managerial contributions by developing and comparing the predictive power of MLR, CART, ANN, and ensembles of these models for skier days estimation. The results of this research project indicate that the ANN model consistently shows superior predictive power for skier days estimation using local, regional, and national data. Further improvements in prediction accuracy were achieved by utilizing four ensemble learning techniques: bagging, stacking, random subspace and voting. While there are some limitations associated with this research related to data measurement frequencies and localization, it appears to be the first to show the increased predictive power afforded by the utilization of ensemble methods at the individual resort level within the skiing industry.

This research extends and generalizes previous research on skier days prediction by Shih et al. (Shih, et al., 2009) and King (King, M.A., 2010). The implications for ski resort management are the possibility of improved skier days estimation and thus enhanced financial budgeting and operational/capacity planning and most importantly, increased skier satisfaction. One future research direction could be to empirically test the hypothesized link between increased operational efficiencies and improved capacity planning with increased skier satisfaction.

After a predictive model has been created, resort management could easily deploy it for ongoing forecasting. All of the independent variables (except the two weather dimensions, Current Snow Depth and YTD Snow Fall) are essentially known in advance, making the implementation of this predictive system very straight forward. For example, ski resort management can easily determine the values of the time dimension independent variables for a *t* plus five day forecast. Because the Lagged Airfare is an average, its net change over such a short period of time would be essentially static. Operationalizing the remaining two weather related independent variables, Current Snow

Depth and YTD Snow Fall, may on first impression, seem completely untenable; but is certainly possible due to the increasing accuracy of weather forecast for North America. As an example, the ski resorts along the Wasatch Range in the state of Utah, United States, typically have access to snow storm forecasts seven to ten days in advance. These highly accurate forecasts provide the time of day, accumulation rate at specific elevation levels, accumulation totals and even water content percent of the snowfall (Hu and Skaggs, 2009). (http://www.wrh.noaa.gov/slc/snow/mtnwx/mtnForecast.php, 2013; http://weather.utah.edu/index.php?runcode=2012110312&t=nam212&r=WE&d=CN, 2013).

Climate change has been a contentious, political and debatable argument for quite some time with staunch advocates on both side of the issue, while North American ski resort managers find themselves somewhere in the middle managing their respective resorts that have been subject to a decade of fluctuating weather patterns. As an objective example of weather volatility, North American skier days for the 2011-2012 ski season were the lowest in 20 years, due to an exceptionally low snow pack, persistent warm temperatures, along with low economic growth and instability (see Exhibit 2.3). This weather pattern resulted in delayed resort openings, early closures, shorter operating hours, and less open terrain all causing operational challenges. However, North American skier days for the 2010-2011 season were at a record high since the NSAA began recording skier visitations. Based on the annual Kottke National End of Season Survey for the 2010-2011 season sponsored by the NSAA, the majority of the ski resort managers sited high levels and consistent snow fall as the main factor for the record setting skier visits. These observations support the ski resort industry adage that snow fall trumps the economy; "when it snows, you go." The Kottke National End of Season Survey has documented this trend of high skier days during periods of high snow fall and low economic growth. However, the annual survey has shown, ski resorts experience significant drops in skier days during a season of low snow fall with a good economy. Within this volatile context, it is increasingly important that skier days forecasts be as accurate as possible.

An additional idea for future research includes obtaining skier days data from additional regional ski resorts for further comparative analysis. The inclusion of additional ensemble methods such as boosting, and stackingC (a more efficient stacking algorithm) could yield additional insights.

References

http://governor.utah.gov/DEA/ERG/2010ERG.pdf, (last accessed February, 2013).

http://www.rita.dot.gov/, (last accessed July, 2012).

http://www.eia.gov/, (last accessed July, 2012).

http://rapid-i.com/content/view/181/190/, (last accessed February, 2013).

http://www.wrh.noaa.gov/slc/snow/mtnwx/mtnForecast.php, (last accessed January 2013)

http://weather.utah.edu/index.php?runcode=2012110312&t=nam212&r=WE&d=CN, (last accessed January, 2013).

Breiman, L., J. H. Friedman, R. A. Olsen and C. J. Stone (1984). Classification and Regression Trees. Belmont, Calif., Wadsworth International Group.

Breiman, L. (1996). "Bagging Predictors." Machine Learning. 24 (2): 123–140.

Burger, C. J. S. C., M. Dohnal, M. Kathrada and R. Law (2001). "A Practitioners Guide to Time-series Methods for Tourism Demand Forecasting — a Case Study of Durban, South Africa." Tourism Management 22(4): 403-409.

Chen, K. Y. (2011). "Combining Linear and Nonlinear Model in Forecasting Tourism Demand." Expert Systems with Applications 38(8): 10368-10376.

Cho, V. (2003). "A Comparison of Three Different Approaches to Tourist Arrival Forecasting." Tourism Management 24(3): 323-330.

Clifford, H. (2002). Downhill Slide : Why the Corporate Ski Industry is Bad for Skiing, Ski Towns, and the Environment. San Francisco, Sierra Club Books.

Das, R., I. Turkoglu and A. Sengur (2009). "Effective Diagnosis of Heart Disease Through Neural Networks Ensembles." Expert Systems with Applications 36(4): 7675-7680.

Dietterich, T. G. and E. B. Kong (1995). Error-correcting Output Coding Corrects Bias and Variance. International Conference on Machine Learning, Tahoe City, CA, Morgan Kaufmann.

Fausett, L. V. (1994). Fundamentals of Neural Networks : Architectures, Algorithms, and Applications. Englewood Cliffs, NJ, Prentice-Hall.

Freund, Y. and R. E. Schapire (1996). Experiments With a New Boosting Algoritm. International Conference of Machine Learning, San Francisco, CA, Morgan Kaufmann Publishers.

Hamilton, L. C., C. Brown and B. D. Keim (2007). "Ski Areas, Weather and Climate: Time Series Models for New England Case Studies." International Journal of Climatology 27(15): 2113-2124.

Han, J. and M. Kamber (2006). Data Mining : Concepts and Techniques. Amsterdam; Boston; San Francisco, CA, Elsevier ; Morgan Kaufmann.

Ho, T. K. (1998). "The Random Subspace Method for Constructing Decision forests." IEEE Transactions on Pattern Analysis and Machine Intelligence 20(8): 832-844.

Hu, C., M. Chen and S.-L. C. McCain (2004). "Forecasting in Short-Term Planning and Management for a Casino Buffet Restaurant." Journal of Travel & Tourism Marketing 16(2-3): 79-98.

Hu, Q. S. and K. Skaggs (2009). Accuracy of 6-10 Day Precipitation Forecasts and Its Improvement in the Past Six Years, 7th NOAA Annual Climate Prediction Application Science Workshop

Hunt, E. B., J. Marin and P. J. Stone (1966). Experiments in Induction. New York, Academic Press.

Kantardzic, M. (2011). Data Mining: Concepts, Models, Methods, and Algorithms. Hoboken, N.J., John Wiley : IEEE Press.

Kim, M. J. and D. K. Kang (2010). "Ensemble with Neural Networks for Bankruptcy Prediction." Expert Systems with Applications 37(4): 3373-3379.

Kim, Y. (2009). "Boosting and Measuring the Performance of Ensembles for a Successful Database Marketing." Expert Systems with Applications 36(2, Part 1): 2161-2176.

King, M. A. (2010). Modeling Consumer Demand for Vail Resorts: A Leisure and Recreational Industry Case Study. Southeast Conference of the Decision Sciences Institute. Wilmington, NC, Decision Science Institute.

Kon, S. C. and L. W. Turner (2005). "Neural Network Forecasting of Tourism Demand." Tourism Economics 11(3): 301-328.

Kuncheva, L. I. (2004). Combining Pattern Classifiers Methods and Algorithms, Wiley.

Law, R. (2000). "Back-propagation Learning in Improving the Accuracy of Neural Network-based Tourism Demand Forecasting." Tourism Management 21(4): 331-340.

Li, G., H. Song and S. F. Witt (2005). "Recent Developments in Econometric Modeling and Forecasting." Journal of Travel Research 44(1): 82-99.

Lim, C. (1999). "A Meta-Analytic Review of International Tourism Demand." Journal of Travel Research 37(3): 273-284.

Lozano, E. and E. Acuña (2011). Comparing Classifiers and Metaclassifiers
Advances in Data Mining. Applications and Theoretical Aspects. P. Perner, Springer Berlin / Heidelberg. 6870: 56-65.

Luzadder, D. (2005). "The Incredible Vanishing Vacation: No Time for Travel." Travel Weekly 64(30): 18-20.

Major, R. L. and C. T. Ragsdale (2000). "An Aggregation Approach to the Classification Problem Using Multiple Prediction Experts." Information Processing and Management 36(4): 683-696.

Major, R. L. and C. T. Ragsdale (2001). "Aggregating Expert Predictions in a Networked Environment." Computers and Operations Research 28(12): 1231-1244.

Marqués, A. I., V. García and J. S. Sánchez (2012). "Exploring the Behaviour of Base Classifiers in Credit Scoring Ensembles." Expert Systems with Applications 39(11): 10244-10250.

McGuigan, J. R., M. R. Charles and F. H. d. Harris (2008). Managerial Economics : Applications, Strategy, and Tactics. Mason, Ohio, Thomson/South-Western.

Mill, R. C. (2008). Resorts : Management and Operation, Wiley.

Montgomery, John. (2012, October 31). Telephone interview.

Naude, W. A. and A. Saayman (2005). "Determinants of Tourist Arrivals in Africa: a Panel Data Regression Analysis." Tourism Economics 11(3): 365-391.

Nisbet, R., J. F. Elder and G. Miner (2009). Handbook of Statistical Analysis and Data Mining Applications. Amsterdam; Boston, Academic Press/Elsevier.

Oh, C.-O. and B. J. Morzuch (2005). "Evaluating Time-Series Models to Forecast the Demand for Tourism in Singapore." Journal of Travel Research 43(4): 404-413.

Ott, R. L. and M. Longnecker (2001). An Introduction to Statistical Methods and Data Analysis. Pacific Grove, Calif., Duxbury - Thomson Learning.

Pai, P. F. and W. C. Hong (2005). "An Improved Neural Network Model in Forecasting Arrivals." Annals of Tourism Research 32(4): 1138-1141.

Paliwal, M. and U. A. Kumar (2009). "Neural Networks and Statistical Techniques: A Review of Applications." Expert Systems with Applications 36(1): 2-17.

Palmer, A., J. José Montaño and A. Sesé (2006). "Designing an Artificial Neural Network for Forecasting Tourism Time Series." Tourism Management 27(5): 781-790.

Patsouratis, V., Z. Frangouli and G. Anastasopoulos (2005). "Competition in Tourism Among the Mediterranean Countries." Applied Economics 37(16): 1865-1870.

Pullman, M. E. and G. M. Thompson (2002). "Evaluating Capacity and Demand Management Decisions at a Ski Resort." Cornell Hotel & Restaurant Administration Quarterly 43(6): 25.

Perdue, R. R. (2002). "Perishability, Yield Management, and Cross-Product Elasticity: A Case Study of Deep Discount Season Passes in the Colorado Ski Industry." Journal of Travel Research 41(1): 15.

Riddington, G. L. (2002). "Learning and Ability to Pay: Developing a Model to Forecast Ski Tourism." Journal of Travel & Tourism Marketing 13(1-2): 109-124.

Rokach, L. (2009). "Taxonomy for Characterizing Ensemble Methods in Classification Tasks: A Review and Annotated Bibliography." Computational Statistics and Data Analysis 53(12): 4046-4072.

Schapire, R. E. (1990). "The Strength of Weak Learnability." Machine Learning 5(2): 197-227.

Shih, C., S. Nicholls and D. F. Holecek (2009). "Impact of Weather on Downhill Ski Lift Ticket Sales." Journal of Travel Research 47(3): 359-372.

Song, H. and G. Li (2008). "Tourism Demand Modelling and Forecasting: A Review of Recent Research." Tourism Management 29(2): 203-220.

Song, H., S. F. Witt, K. F. Wong and D. C. Wu (2009). "An Empirical Study of Forecast Combination in Tourism." Journal of Hospitality & Tourism Research 33(1): 3-29.

Stevenson, W. J. (2012). Operations Management. New York, McGraw-Hill/Irwin.

Sun, J., M. y. Jia and H. Li (2011). "AdaBoost Ensemble for Financial Distress Prediction: An Empirical Comparison With Data From Chinese Listed Companies." Expert Systems with Applications 38(8): 9305-9312.

Tabachnick, B. G. and L. S. Fidell (2000). Using Multivariate Statistics. Boston, MA, Allyn and Bacon.

Tan, A. Y. F., C. McCahon and J. Miller (2002). "Modeling Tourist Flows to Indonesia and Malaysia." Journal of Travel & Tourism Marketing 13(1-2): 61-82.

Ting, K. M. and I. H. Witten (1999). "Issues in Stacked Generalization." Journal of Artificial Intelligence Research 10: 271-289.

Turban, E., R. Sharda, J. E. Arson and T.-P. Liang (2007). Decision Support and Business Intelligence Systems, Pearson Prentice Hall.

Uysal, M. and M. S. E. Roubi (1999). "Artificial Neural Networks Versus Multiple Regression in Tourism Demand Analysis." Journal of Travel Research 38(2): 111.

Witten, I. H., E. Frank and M. A. Hall (2011). Data Mining : Practical Machine Learning Tools and Techniques, Elsevier.

Wolpert, D. H. (1992). "Stacked Generalization." Neural Networks 5(2): 241-259.

Wolpert, D. H. and W. G. Macready (1997). "No Free Lunch Theorems for Optimization." Evolutionary Computation, IEEE Transactions on 1(1): 67-82.

Yi, J. and V. R. Prybutok (2001). "Neural Network Forecasting Model As An Alternative to OLS Regression Model for Handling Messy Data." Commerce and Economic Review Vol. 17(No. 1): 137-158.

Zalatan, A. (1996). "The Determinants of Planning Time in Vacation Travel." Tourism Management 17(2): 123-131.

Chapter 3

Ensemble Learning Methods for Pay-Per-Click Campaign Management

"Search, a marketing method that didn't exist a decade ago, provides the most efficient and inexpensive way for businesses to find leads."
John Battelle

Abstract
Sponsored search advertising has become a successful channel for advertisers as well as a profitable business model for the leading commercial search engines. There is an extensive sponsored search research stream regarding the classification and prediction of performance metrics such as clickthrough rate, impression rate, average results page position and conversion rate. However, there is limited research on the application of advanced data mining techniques, such as ensemble learning, to pay per click campaign classification. This research presents an in- depth analysis of sponsored search advertising campaigns by comparing the classification results from four traditional classification models (Naïve Bayes, logistic regression, decision trees, and Support Vector Machines) with four popular ensemble learning techniques (Voting, Boot Strap Aggregation, Stacked Generalization, and MetaCost). The goal of our research is to determine whether ensemble learning techniques can predict profitable pay-per-click campaigns and hence increase the profitability of the overall portfolio of campaigns when compared to standard classifiers. We found that the ensemble learning methods were superior to the base classifiers when evaluated by profit per campaign criterion. This paper extends the research on applied ensemble methods with respect to sponsored search advertising.

Keywords: Sponsored search, pay-per-click advertising, classification, ensemble modeling.

1. Introduction

Search engines are an indispensable tool for interacting with the World Wide Web (WWW) and have long provided user value, from the earliest universal resource locator (URL) directories to present day highly optimized query results. Companies competing in the search engine industry have slowly monetized their early search innovations and have created sustainable business models by providing the business community with an advertising channel called sponsored search. Internet advertising is nearing a $42.78 billion dollar industry as reported by the Interactive Advertising Bureau (IAB Internet Revenue Report, 2013). The three search engine industry leaders, Google, Yahoo!, and Bing, who hold a combined market share of over 96% (comScore, 2014), each offer a competitive sponsored search platform. Search engine providers are acutely aware of user search behavior and the associated marketing value of the page location of search results (Haans, H., N., et

al., 2013). Search engines allow any individual or company to submit a URL for advertising purposes so it can be indexed and then made available for retrieval. Search engines call this submission process organic search and provide the service free. However, the probability of a search engine listing a specific URL for an advertiser's landing page in the top display section is quite low, even with a search engine optimized landing page (Jansen and Mullen, 2008). The statistics in Exhibit 3.1 are often quoted as support for sponsored search.

93% of directed traffic to websites is referred by search engines.	Forrester Research, 2006
99% of Internet searchers do not look past the first 30 search results.	Forrester Research, 2006
97% of Internet searchers do not look past the top three results.	Forrester Research, 2006
65% of online revenue is generated by holders of the top three results.	Forrester Research, 2006
Approximately 131 billion searches performed each month, globally.	comScore, 2010

Exhibit 3.1. Search Engine Usage Statistics

In contrast to organic search, sponsored search advertising is more complex, but offers the potential of a higher return on investment (Moran and Hunt, 2009). In sponsored search, advertisers first bid on keywords offered by the search engines and after a keyword is acquired, their advertisements associated with the keyword are displayed using proprietary ranking algorithms. Ad rankings typically take into account relevance to users' search and keyword bid amount offered by advertiser (Fain and Pedersen, 2005; Jansen and Mullen, 2008). The three predominant advertisement billing schemes used by search engines are pay-per-impression (PPM), pay-per-click (PPC), and pay-per-action (PPA) (Mahdian and Tomak, 2008; Moran and Hunt, 2009). When using a PPM billing scheme, advertisers are charged each time their ad is displayed, regardless of whether the user clicks on the ad. Under PPC billing, the advertiser is charged only when their ad or URL is clicked on, and with PPA billing, the advertiser pays only when a user action such as a sign-up or purchase occurs.

The research conducted for this article analyzed a large data set of PPC advertisements placed on Google. Our data set was provided by an industry leading Internet marketing company (NAICS 518210) specializing in PPC campaign services, that managed a campaign portfolio containing 8,499 PPC campaigns for a multi-billion dollar home security provider (NAICS 561612). The Internet marketing company sells its online marketing services to its clients and assumes all PPC related costs, while the advertising clients pay on a PPA basis when purchases are made.

The objective for this study was to construct a set of classification models, using a combination of

data and text mining techniques and ensemble learning techniques, that are capable of classifying a *new* PPC advertisement campaign as either sufficiently profitable or not. Profit is based on whether clicks per acquisition are lower than the breakeven threshold for the advertised item with an overall objective of maximizing total profit of the full portfolio of initiated campaigns. An "acquisition" refers to the event of the advertiser successfully selling a multi-year home security contract to the user who clicked on the advertisement.

The remainder of this paper is organized as follows. Section 2 provides a literature review of related work. Section 3 discusses our specific research questions and contributions. Methodology and research design implementations are described in section 4. Section 5 provides a detailed discussion of the research results while section 6 presents managerial implications, research limitations, and conclusion.

2. Related work

This section describes related work in the fields of sponsored search, advertisement content modeling, and data mining.

2.1 Sponsored Search and Search Engine Marketing

Advertisers are anxious to improve the success of sponsored search listings (D'Avanzo, E., et al., 2011). Various authors have studied prediction of sponsored search campaign success from text features created from keywords or advertisement text. For example, Jansen and Schuster (2011) investigate whether keywords associated with different phases of the buying funnel (Consumer Awareness, Research, Decision, and Purchase) have different success rates. A number of authors have attempted to determine whether semantic features of keywords impact sponsored search success (Rutz and Bucklin, 2007; Shaparenko, B., et al., 2009; Rutz, O., et al., 2011).

2.2 Advertisement Content Modeling

Textual content, including that found in advertisements, can be analyzed within a data or text mining context based on the numeric representations of stylometric, sentiment, and semantic features of the text (Abrahams, A.S., et al., 2013; Aggarwal and Zhai, 2012; Nielsen, J. H., et al., 2010; Haans, H., et al., 2013).

Stylometrics describes the readability and stylistic variations of a portion of text using numerous metrics such as characters per word, syllables per word, words per sentence, number of word repetitions, and Flesch Reading Ease (Sidorov, G., et al., 2014). Tweedie, F. J., et al., (1996) and Ghose, A., et al., (2011) describe the value of stylometric modeling of text using several machine learning techniques.

Sentiment content refers to the emotional or affective communication embedded in text. Feldman (2013) provides an overview of business applications of sentiment analysis in areas such as consumer reviews, financial market blogs, political campaigns, and social media advertising. When studying the characteristics of an advertisement, the representation and interpretation of its sentiment content is important because advertisements not only deliver objective descriptions about branded products or services, but also can induce measureable emotional reactions from current or potential customers. As proposed by Heath (2005), advertisements that approach readers' feelings rather than knowledge can be processed with low attention and can result in increased buying behavior.

Semantics refers to the meaning of the text, such as the categories of items referred to in the text. Stone, Dunphy, and Smith (1966) describe a semantic tagging method that extracts word senses from text and classifies the words into concept categories. Abrahams, A.S., et al. (2013) illustrates how this type of sematic tagging of advertisement content can be useful for audience targeting.

2.3 Data mining

Numerous data mining and machine learning techniques have been used to create models that predict important sponsored search performance metrics such as clickthrough rate, conversion rate, and bounce rate. Logistic regression, Support Vector Machines (SVM) and Bayesian models, as well as other techniques, have been applied to these predictive problems. Search engines are primarily concerned with modeling the click-through rate for both new and ongoing ads, because revenue depends on their ability to rank PPC ads relevant to searchers with as high a click-through rate as possible. In contrast, advertisers are more likely to focus on the conversion rates associated with their PPC campaigns.

Richardson, et al., (2007) argues that modeling click-through rates for ads with known run times is a straight-forward process, while modeling the click-through rate for a new ad presents unique challenges. Because of the rapid growth in the inventory of new PPC ads, along with the fast turnover of these ads, the authors indicate that it has become increasingly difficult to estimate plausible click-through rates based on historical data. The authors present a logistic regression model fit using a feature set created from the actual ad text and numeric attributes (e.g., landing page, keywords, title, body, clicks and views). The authors discuss how their base logistic regression model was more accurate than an ensemble of boosted regression trees.

Wang, et al., (2012) add to the research stream of ensemble learning within a sponsored search context by introducing an ensemble for click-through prediction. Wang, et al., were team members who participated in the 2012 KDD Cup (http://www.kddcup2012.org/) and built an ensemble consisting of four base classifiers: maximum likelihood estimation, online Bayesian probit regression, Support Vector Machines, and latent factor modeling. Feature creation is a reoccurring research theme within the context of sponsored search because the unit of analysis, the PPC advertisement, typically provides relatively few independent variables. The authors' main contribution is their novel approach for combining the prediction results from the four base classifiers by using a ranking-based ensemble algorithm.

Ciaramita, et al., (2008) maintain that commercial search engines must be able to accurately predict if an ad is likely to be clicked. These ongoing predictions help determine an estimated click-through rate, which in turn, drives revenue for the search engine. Using click stream data from a commercial search engine, the author created three variants of the perceptron algorithm: a binary perceptron, a ranking perceptron, and a multilayer perceptron. Based on different feature sets developed from keyword query logs and numeric representations of text attributes, the authors show that the multilayer perceptron outperforms both the binary and ranking perceptrons. Adding to the feature creation research stream, the authors provide a discussion of feature creation from several novel numeric representations of ad text.

Graepel, et al., (2010) present a new online machine learning algorithm for predicting click-through rates for Microsoft's commercial search engine Bing, called adPredictor. At the time of their

publication, adPredictor was supporting 100% of Bing's sponsored search traffic. AdPredictor is based on a general linear regression model with a probit link function that predicts a binary outcome. The authors indicate that their primary reasons for implementing adPredictor were for speed and the scaling improvements.

Ghose and Yang (2008) take an innovative research approach by developing a predictive model from the perspective of the advertiser. The fundamental issue that Ghose and Yang attempt to address is how sponsored search conversion rates, order values, and profits compare to the same performance metrics for organic search. The authors argue that predicting the conversion rate of a specific advertisement is quite useful for most retail advertisers because it gives the advertiser the ability to better plan campaign budgets, improve ad content, select a better range of keywords, and improve the pairing of ad with the keyword, all in advance. Using a data set containing paid search advertising information from a major retail chain, the authors developed a complex hierarchical Bayesian Network model for conversion rate prediction. Their results show a strong positive overall association between the advertisement conversion rate and the feature set, retailer name, brand name, and keyword length. Ghose and Yang (2009) also provide an in-depth discussion of variable sets and metrics as related to sponsored search modeling.

Sculley, et al., (2009), Becker, et al., (2009), Attenberg, et al., (2009) discuss sponsored search performance metrics such as the bounce rate, landing page impressions, and navigational trails. Each of the authors modified several classification models that enhanced the performance of the related metrics.

3. Research contributions

The objective of our research is the development of a more effective method of classifying PPC campaign success, thus maximizing total campaign portfolio profitability. Sophisticated data mining techniques, such as ensemble learning, are needed to assist campaign managers with predicting PPC campaign success (Abrahams, A.S., et. al., 2014). Sponsored search advertising can be profitable, having a clear and measurable return on investment. However, to be successful, advertisers must understand the main assumption of search engine marketing; that is, a high level of relevant traffic plus a good conversion rate equals more sales, where quality advertisement content drives the level

of relevant traffic (Moran and Hunt, 2009). We define a PPC advertisement as including a title, a description text, and the URL. We hypothesized that ensemble learning techniques could achieve higher classification accuracies and more profitable campaigns, when compared to standard classification methods.

Ensemble analysis is currently quite popular; however, numerous researchers continue to rely on conventional data sets from the University of California Machine Learning Repository as well as other data repositories. We make a significant research contribution by using actual data provided to us based on a collaborative relationship with an industry leading PPC campaign management company. PPC campaign data set features are relatively restricted due to the length limit imposed on PPC ads by Google. We make additional research contributions by outlining the creation of derived features based on the ad text related to the title and ad content.

We preprocessed the textual content of each PPC advertisement by tokenizing them into stylometric, sentiment, and semantic features. These text preprocessing techniques are a more sophisticated approach for representing data, than the standard bag of words model (Witten, I. H., et al., 2011) and are discussed in detail in Section 4. We then analyzed the feature set, against a series of four base classifiers and ten ensemble algorithms. We believe this is the first work to use advertisement content modeling techniques to extract a broad set of stylistic, sentiment, and semantic attributes from PPC marketing campaigns, and to then apply ensemble learning techniques to PPC campaign success prediction using these attributes. We assess the level and robustness of profit produced by the portfolio of advertising campaigns chosen by each classifier.

4. Methodology

The case study method of theory building is widely accepted (Benbasat, I., et al., 1987; Eisenhardt, 1989; Yin, 2009). This research follows a design consistent with earlier studies of sponsored search advertisements (Rutz, O., et al., 2011), and adheres to the guidelines of content analysis research suggested by Neuendorf (2002).

4.1 Data Set

The data set contains 8,499 PPC campaign marketing records over a six week period obtained from a

major PPC campaign management company who administers Google ads on the behalf of a large home security systems and service advertiser. The target class, defined as a successful (i.e. profitable) campaign, contains 1,094 records, which represents 12.87% of the data set. We developed 271 numeric features from the stylistic, sentiment, and semantic analysis of the keywords available to the advertiser, the advertising text, and the campaign intention parameters. Exhibit 3.2 illustrates a partial record and feature set.

AD_ID	Ad Success (0-.5% = FAIL, >.5% is SUCCESS)	Impressions	Clicks	Click Through Rate	Conversion	Purchase Rate	Spend	Average Position	BID Average of ANEW Valence	BID Average of ANEW Arousal	BID Average of ANEW Dominance	AD Words	AD Characters	AD Sentences	AD Economic Value	AD Quality	AD Trade name	AD Product Description	
10143	SUCCESS	2805	267	0.10	20	0.07	5373.36	1.99	0.00	0.00	0.00	9	60	2	2	0	2	4	...
10144	FAIL	0	0	0.00	0	0.00	2244.43	1.99	3.64	2.11	2.77	9	60	2	2	0	2	4	...
10145	SUCCESS	5278	507	0.10	28	0.06	2244.43	1.99	3.64	2.11	2.77	10	59	3	3	0	2	4	...
10146	SUCCESS	10317	937	0.09	35	0.04	5373.36	1.99	0.00	0.00	0.00	9	61	2	2	0	2	4	...
10147	SUCCESS	3834	348	0.09	21	0.06	2244.43	1.99	3.64	2.11	2.77	9	61	2	2	0	2	4	...
10148	FAIL	588	45	0.08	0	0.00	5373.36	1.99	0.00	0.00	0.00	9	61	2	2	0	2	4	...
10149	SUCCESS	7263	576	0.08	22	0.04	5373.36	1.99	0.00	0.00	0.00	9	61	2	2	0	2	4	...
10150	FAIL	201	19	0.09	0	0.00	2244.43	1.99	3.64	2.11	2.77	9	56	1	1	1	1	2	...
10151	FAIL	12	1	0.08	0	0.00	7.54	2.00	5.06	2.81	3.81	9	59	2	2	0	2	4	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

Exhibit 3.2. Data Set Example

4.1.1 Dependent Variables

Our data set contains four traditional sponsored search dependent variables: ad impressions, the clickthrough rate, the average page position, and the purchase rate. An impression is a metric with roots from traditional advertising that counts the times a specific advertisement is shown. The clickthrough rate is the number of clicks received for an ad divided by the number of impressions provided. The average position metric indicates the average ranked order of the advertisement in relation to other competing ads displayed on a search engine's results page. The purchase rate, also known as the conversion rate or acquisition rate, represents the percentage of clicks that result in a purchase. However, we use only one of these dependent variables for this research paper. We derived a categorical dependent variable by using the purchase rate. We coded this dependent variable with a label of *Success* if the purchase rate for a campaign was above 0.5%, otherwise *Fail*. We chose 0.5% as the breakeven purchase rate, because we determined, for the target class, that at least one purchase was necessary for every 200 paid clicks in order for the campaign to be profitable. In other words, the break even ratio is the average cost-per-click divided by the marginal profit, which we found to be approximately $1/$200.

4.1.2 Independent Variables

An example of a paid search ad is illustrated in Exhibit 3.3. See Appendix A for an example of sponsored search listings. Our data set contains four complex independent variables: the

advertisement title, the advertisement textual content, the available keywords for bidding, and a set of campaign intentions developed by the campaign management company. The advertisement is the text displayed in the user's web browser after a keyword query. The advertisement contains three sections, a title headline, descriptive text and a display URL. Google limits the length of an entire ad to 70 characters.

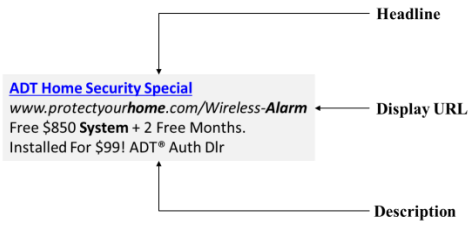


Exhibit 3.3. Structure of a Sponsored Ad

We tokenized the textual content of each advertisement in the data set for content modeling. The Affective Norms for English Words [ANEW] (Bradley and Lang, 2010), and AFFIN lexicons (Nielsen, 2012) and SentiStrength sentiment application (Thelwall, M., et al., 2010, 2012) were used for deriving the sentiment content in the advertisements contained in the data set. In each case, sentiment scores using each method (ANEW, AFINN, SentiStrength) were summed for the advertisement text. We extracted the semantic content of each PPC advertisement using General Inquirer, which includes approximately 11,780 word entries from the Harvard-IV-4 and the Lasswell lexicons (Stone, P. J., et al., 1966). We labeled each word in each PPC advertisement with one or more tags; each tag corresponding to one category defined by the General Inquirer lexicon. The counts of the tags were used as the numeric representation for the campaign. After stylometric, sentiment and semantic text preprocessing, and feature reduction, we obtained 271 content related features.

Bid words are the keywords supplied by a search engine that advertisers bid on in an auction setting. The bid amount determines the ranked position of the ad on the search results page and the eventual price per click charged to an advertiser by the search engine company. Google displays the advertisement if the bid amount is among the highest and the associated landing page the most relevant based on Google's propriety Page Rank algorithm. Each bid word set is represented as tokens for content modeling.

In a similar fashion as traditional impression-based advertising, PPC campaign managers seek to segment their market and target customer groups using specific campaigns with the hope of matching it with customer motivation. Our data set contains nine campaign intentions (e.g., time targeting, device targeting, brand targeting, and place targeting) that we coded numerically.

4.2 Evaluation Criteria

A common practice in classification analysis is to assume that the misclassification costs are symmetric and that the classes contained in a data set are equally balanced. However, researchers making these assumptions should use caution as classification techniques have numerous evaluations metrics that may lead to very different performance conclusions given a specific research context (Moraes, R., et al., 2013). We base this concern on the observation that many real world data sets are imbalanced with respect to the target class and exhibit asymmetric misclassification costs. Weiss and Provost (2003) ran a large empirical study with imbalanced data sets and concluded that relying solely on overall accuracy led to biased decisions with respect to the target class.

Although we evaluate several traditional confusion matrix metrics, shown in Exhibit 3.4, for comparing the performance of our base and ensembles models, the main goal of our research is assessing the total profitability of the full portfolio of selected PPC marketing campaigns, applying ensemble learning techniques. The true positive rate (TP) represents the number of true instances correctly classified. The true negative rate (TN) represents the false instances correctly classified. The false positive rate (FP) and false negative rate (FN) indicate the converse. Due to the search engine marketing context of our research, we suggest two specific financial metrics, Profit per PPC Campaign and Total Campaign Portfolio Profit, as more applicable performance metrics for our analysis (Moran and Hunt, 2009). We define Profit per PPC Campaign as:

$$(Campaign\ Purchases * Revenue\ per\ Conversion) - Campaign\ Spend$$

Campaign Spend is the sum of the cost-per-click, for all clicks in the campaign. We define Total Campaign Portfolio Profit as the sum of the profit or loss for all PPC campaigns. As discussed later in detail, we calculate all performance metrics, including the profit analysis, from 30 different unseen hold out partitions.

$$\textbf{Average Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

Exhibit 3.4. Performance Metrics

4.3 Base Classifiers

We utilized four popular classifier algorithms based on their ability to predict a categorical dependent variable, their extensive research streams, and their diversity of classification mechanisms, e.g., probabilistic, statistical, structural or logical (Kotsiantis, 2007). We feel that our classifier selection process helps reduce model bias and facilitates the comparative assessments of model performance.

4.3.1 Naïve Bayes

The Naïve Bayes technique is a simple probabilistic classifier patterned after the Bayes Theorem. Despite its independence assumption simplification, the Naïve Bayes classifier often exhibits excellent predictive accuracy with the additional benefit of being very computationally efficient (Domingos and Pazzani, 1997). Researchers have frequently applied the Naïve Bayes algorithm to classification applications related to text mining (Li, D., et al., 2013). In a PPC context, the independent variables are the attributes we derived. The naïve independence assumption states that attributes $A = \{a_1, a_2, a_3, \ldots a_n\}$ representing a campaign ppc_i that is classified as label C_j are all statistically independent. In our binary class environment where C_0 represents *Fail* and C_1 represents *Success*, we calculated the predicted class C_j for campaign ppc_i as the maximization of

$$P(C_j|ppc_i) = argmax\left(\frac{P(C_j)\prod_{i=1}^{n} P(a_i|C_j)}{P(C_0)\prod_{i=1}^{n} P(a_i|C_0) + P(C_1)\prod_{i=1}^{n} P(a_i|C_1)}\right)$$

4.3.2 Logistic Regression

Logistic regression is a popular and powerful statistics-based classification technique used in numerous business applications (Richardson, M., et al., 2007). Logistic regression is similar to multiple linear regression, but is appropriate for situations involving a binary dependent variable.

The logistic response function ensures the estimated value of the dependent variable will always fall between zero and one, making it a suitable technique for binary classification problems. The logistic response function estimates the probability of an observation $A = \{a_1, a_2, a_3, \ldots a_n\}$ belonging to class C_1 as follows:

$$p(y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_n X_n)}}$$

The parameters for the logistic regression function are typically determined using maximum likelihood estimation.

4.3.3 Classification Trees

Classification trees (Breiman, L.F., et al., 1984; Quinlan, J., 1986) are one of the most intuitive and transparent classification algorithms, when compared to other supervised learning techniques, which helps explain its popularity in application as well as its extensive research stream. Classification trees require few model assumptions, no domain knowledge and minimal parameter settings, which makes the flexible technique attractive in numerous business applications. The goal of a classification tree is to recursively partition a training data set, typically using binary splits, into increasingly homogenous subsets. The level of homogeneity is based on the concept of minimizing membership diversity within each new partition (Shmueli, G., et al., 2010).

4.3.4 Support Vector Machines

Support Vector Machines are a supervised classification algorithm that is growing in popularity due to strong theoretical support provided by Statistical Learning Theory and its superior accuracy in numerous applications. SVM have been used extensively in text mining applications and research (Taboada, M., et al., 2011; Miner, G., et al., 2012; Sebastiani, F., 2002). Vapnik, Boser and Guyon developed SVM in the early 1990s, and subsequently patented their algorithm (Vapnik, 1995). The main idea behind Support Vector Machines is classifying by applying a linear hyperplane regardless of whether the data is linear separable. The algorithm transforms the original data inputs into a higher dimension feature space using one of several popular kernel functions. The main theoretical advantage of using a SVM is that the problem can be written as a constrained quadratic optimization problem, which assures the algorithm finds a global maximum (Burges, C. C., 1998). The optimization problem can be expressed by

$$max \sum_{i=1}^{n} \alpha_i - .5 \sum_{ij} \alpha_i \alpha_j c_i c_j K(A_i, A_j)$$

where A_i is a PPC campaign attribute vector such that $A_i=\{a_i,a_2,a_3,\ldots a_n\}$, A_j is a testing instance, c_i represents the known classes, c_j represents the testing instance class, α_i and α_j are parameters for the model, and K denotes the incorporated kernel function. We used the linear kernel, which is the default kernel setting in RapidMiner v5.3.013

4.4 Ensemble Learning

Ensemble learning, involves building a classification model from a diverse set of base classifiers where accuracy can come to the forefront and the errors tend to cancel out. Exhibit 3.5 provides a classic conceptual illustration of classification error cancelation by averaging decision tree hyperplanes. The first pane shows that the two classes are separated by a diagonal decision boundary. The second pane shows the classification errors for three separate decision trees. The last pane shows the error cancelation effect.

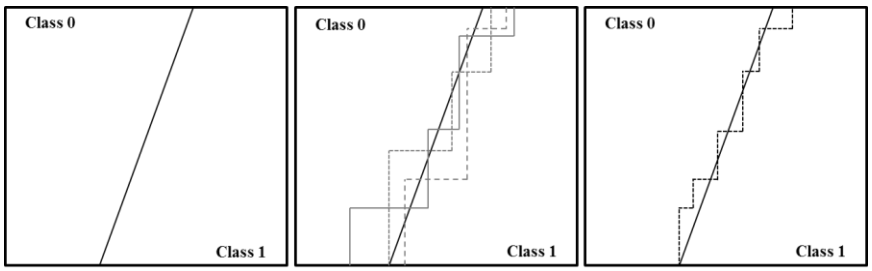


Exhibit 3.5. Decision Tree Averaging (after Dietterich, 1997)

Ensembles, also known as meta-classifiers, are generally more accurate than a standalone classifier when the ensemble contains a diverse set of base models, created from a moderately independent set of training partitions, and then combined by an algebraic method (Kim, 2009; Kuncheva, 2004). Our research compares the classification accuracy of four well-known ensemble methods, Voting, Boot Strap Aggregation, Stacked Generalization, and MetaCost, with the four previously discussed individual classifiers. We selected these ensembles based on their extensive research streams as well as being representative of common themes synthesized from several taxonomies reflected in the literature (Kuncheva, 2004; Kantardzic, 2011; Witten, et al., 2011; Rokach, 2009).

4.4.1 Voting

Voting is the simplest ensemble learning algorithm for combining several different types of classification models. Because of this simplicity, researchers commonly use Voting as a base line measure when comparing the performance of multiple ensemble models (Sigletos, G., et al., 2005). For a standard two-class problem, Voting determines the final classification result based on a majority vote with respect to the class label. For a multi-class problem, plurality Voting determines the class label. In the base case, Voting gives each classification model equal weighting, and represents a special case of algebraic combination rules. Weighted voting, a more general voting rule, allows a modeler to place more weight on models that have higher accuracy, based on prior domain knowledge, which lessens the chance of selecting an inaccurate model (Major and Ragsdale, 2001). For a continuous dependent variable, additional algebraic combination rules are applied; such as the maximum rule, the minimum rule, and the product rule (Kantardzic, 2011).

4.4.2 Boot Strap Aggregation

Boot Strap Aggregation, abbreviated as Bagging, was proposed by Breiman (1996) and has become one of the most extensively researched and applied ensemble learning techniques to date due to its accuracy, simplicity and computation efficiency (Polikar, 2012). The basic process behind Bagging is to take a training data set and create a set of new training data partitions, each containing the same number of observations, by randomly sampling with replacement from the original data set. The next step is to train a set of classifiers using the training data set replicates and then to combine the predicted class of each observation from each model by majority vote. Several research streams indicate that the Bagging ensemble achieves higher average accuracy when it contains unstable classifiers, such as decision trees and artificial neural networks (Kim, 2009).

4.4.3 Stacked Generalization

Stacked Generalization (or Stacking), created by Wolpert (1992), is a hierarchical structured ensemble that contains a set of level 0 classifiers and one level 1 classifier, where the level 1 classifier acts as an arbiter by combining the results of each level 0 classifier. In Stacking, unlike Bagging or other ensemble techniques, the level 0 set of classifiers may contain different types of algorithms, e.g., decision trees, logistic regression, Naïve Bayes, and Support Vector Machines. Stacking is a popular ensemble technique; however, there is a lack of consensus on how to choose a

level 1 classifier (Wang, G., et al., 2011). Stacking is very similar to Voting, but instead being restricted to a simple majority mechanism for combination, Stacking allows the researcher to embed any classifier as the level 1 classifier.

Stacking first divides the data into two disjoint partitions where the training partition contains 67% of the observations data and the testing partition contains the remainder. Next, each level 0 classifier is trained on the training partition. Next, the level 0 classifiers are applied to the testing partition to output a predicted class. Lastly, to train the level 1 classifier, the predicted classes from each level 0 classifiers are used as inputs along with the correct class labels from the test set as outputs.

4.4.4 MetaCost

Most classification algorithms make the implicit assumptions that the data set is balanced with respect to the class label and, most importantly, that the cost of misclassification errors are symmetric (Elkan, 2001). However, the assumption of symmetric misclassification cost is seldom true when working with realistic data sets because the class label of interest usually represents a small portion of the data set, and has a relatively high cost of misclassification.

MetaCost is an ensemble wrapping technique that allows a researcher to embed a cost agnostic classifier within a cost minimization procedure, which makes the base classifier cost sensitive by assigning a higher cost to false negatives than to false positives. The wrapper does not make any changes to the embedded classifier and does not require any functional knowledge of the embedded classifier. The MetaCost ensemble is quite flexible as it is applicable to multi-class data sets as well as to arbitrary cost matrices (Sammut and Webb, 2011). In his seminal paper, Domingos (1999), noted that it is common to see large misclassification cost reductions when using MetaCost when compared to cost agnostic classifiers. In a sense, MetaCost is a modified form of Bootstrap Aggregation where the known class labels are relabeled based on minimizing a conditional risk function. After the training data sets are re-labeled, a single new classifier is trained (Witten and et al., 2011).

In addition to the ability to make a classifier cost sensitive, MetaCost has the added benefit of interpretability, because once the ensemble work of class relabeling is completed, it trains a single

classifier. MetaCost requires a longer runtime and thus higher computational utilization when compared to other common ensembles. In addition, MetaCost assumes that misclassification costs are known in advance and are the same for each observation (Witten and et al., 2011).

4.5 Experimental Design

We constructed 14 different classification models, utilizing RapidMiner v5.3.013 as our software platform. RapidMiner is an open source data mining workbench, with both a community edition and commercial version, which contains an extensive set of ensemble operators. We created four standalone classification models, as previously discussed, to establish baseline performance metrics. Then we created one Voting ensemble containing all base models, four Bagging ensembles, one for each base model, one Stacking ensemble containing all base models and four MetaCost ensembles, one for each base model. Exhibit 3.6 provides a conceptual diagram of the process logic for our complete experimental procedure.

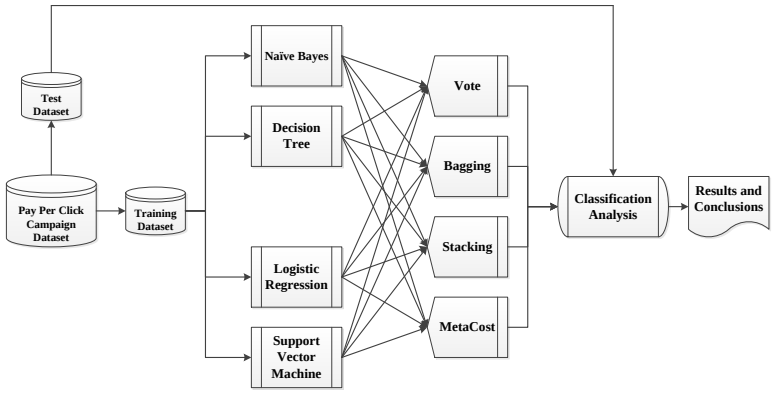


Exhibit 3.6. Conceptual Ensemble Experiment

To implement the full experiment, we followed a repeated measures experimental design where the same *subject* is measured multiple times under different *conditions*. In our current research context, the subjects are the training and testing partitions and the conditions represent the 14 different classification model configurations. The primary advantage of using a repeated measures design is that the structure gives us the ability to control variations of the training and testing partitions across all 14 models, which typically creates a smaller standard error and thus a more powerful *t* test statistic (Ott and Longnecker, 2001). Using a stratified random sampling approach, which we replicated using 30 different random seeds, we created 30 different training and testing partitions, using the generally accepted partitioning ratio of 67% for training and 33% for

testing (Witten, I. H., et al., 2011; Kantardzic, M., 2011). Every model was tested on the same 30 random training and test set pairs, allowing us to use the matched pair *t* test method to determine whether differences were statistically significant. Each of the 30 training and testing partitions, considered a replication, was used as input across the 14 classification models as illustrated in Exhibit 3.7.

	Base Model or Ensemble				
Training/Testing Partitions	1	2	3	…	14
1	$m_{1,1}$	$m_{1,2}$	$m_{1,3}$	…	$m_{1,14}$
2	$m_{2,1}$	$m_{2,2}$	$m_{2,3}$	…	$m_{2,14}$
3	$m_{3,1}$	$m_{3,2}$	$m_{3,3}$	…	$m_{3,14}$
⋮	…	…	…	…	…
30	$m_{30,1}$	$m_{30,2}$	$m_{30,3}$	…	$m_{30,14}$

Exhibit 3.7. Repeated Measure Experimental Design

We selected the default parameter settings for most RapidMiner operators; however, we did make several exceptions. We selected the Laplace Correction parameter for the Naïve Bayes operator. We increased the ensemble sizes of both the Bagging and MetaCost ensembles to 25 members (Kim, 2009; King, M.A., et al., 2014). We controlled our random seed values for partitioning the data set with the simple process of enumerating them 1 through 30. The MetaCost ensemble requires a parameter input of false negative and false positive misclassification costs. We derived these values based on our revenue per acquisition and cost per acquisition. Estimating the actual values used for these two misclassification cost are important, however the size of the ratio of the two costs has the most effect on MetaCost accuracy (Kim, J., et al., 2012).

5. Results and evaluation

In this research study, there were two advertisement campaign classes, Success and Failure. As previously discussed, we defined an advertisement campaign as a Success where the Purchase Rate was above 0.5%. As is common with most binary classification research, a discussion of overall accuracy, precision, recall and model lift will follow. Moreover, we also present results describing model profit, as this metric is the most applicable for our research due to asymmetric classification costs.

Exhibit 3.8 shows the average model accuracies, as previously defined in Exhibit 3.4. Naïve Bayes base model and the Bagging and MetaCost ensemble models performed poorly as compared to the remaining base and ensemble models. We did not expect the degree of performance difference to be this large, as the Naïve Bayes classification typically performs quite well with large feature sets. There are several possible reasons why the Naïve Bayes classifiers performed so poorly. Researchers commonly use the Naïve Bayes classifier for baseline classification analysis, due to its computational speed and simplicity. The Voting ensemble is implemented with the same reasoning. One possible reason for poor performance of the Naïve Bayes classifier is the conditional independence that the Naïve Bayes classifier assumes is severely inappropriate for our data set. When features have abnormally high levels of multicollinearity, the Naïve Bayes algorithm performs poorly. An additional possible reason for the poor performance is that our hypothesis space may not be linearly separable.

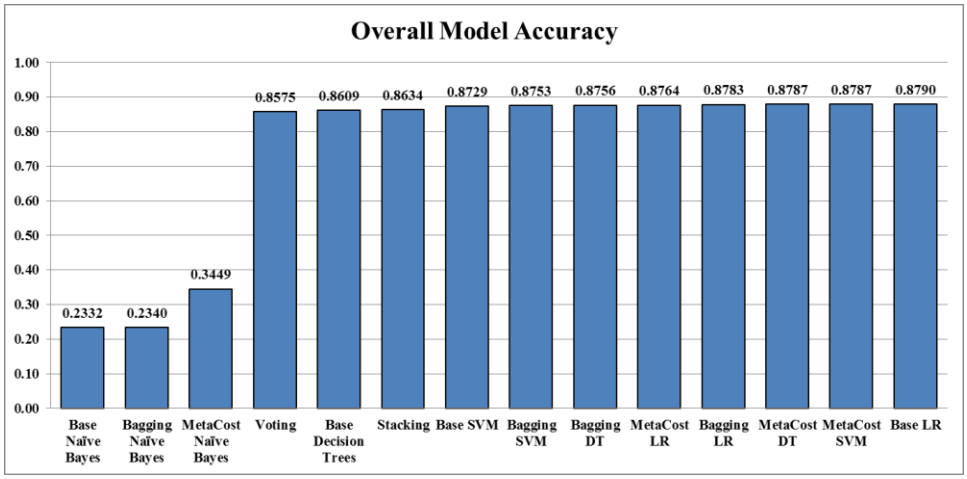


Exhibit 3.8. Overall Model Accuracy

A final possible reason that Naïve Bayes models perform poorly when measured in terms of overall campaign portfolio profit is that Naïve Bayes models have very high recall and low precision as illustrated in Exhibit 3.10. The Naïve Bayes models are costly in terms of ad spend due to the large number of false positives, i.e. failed campaigns that Naïve Bayes invests in as a result of their low precision. In addition, while the Naïve Bayes models have high recall across campaigns in general, it should be noted that campaigns have vastly different profitability, and the resulting distribution of campaign profit has a strong right skew. This means there are a small number of extremely high profit campaigns. Evidently, our Naïve Bayes models do not select the highest profit campaigns,

and leave a "lot of money on the table." However, the MetaCost Naïve Bayes model clearly compensates to a degree for this weakness, as illustrated in Exhibit 3.11. Therefore, in terms of Recall, while the Naïve Bayes models are correct more often than other classifiers, they are unfortunately incorrect "when it counts," suffering from false negatives on the highest profit campaigns.

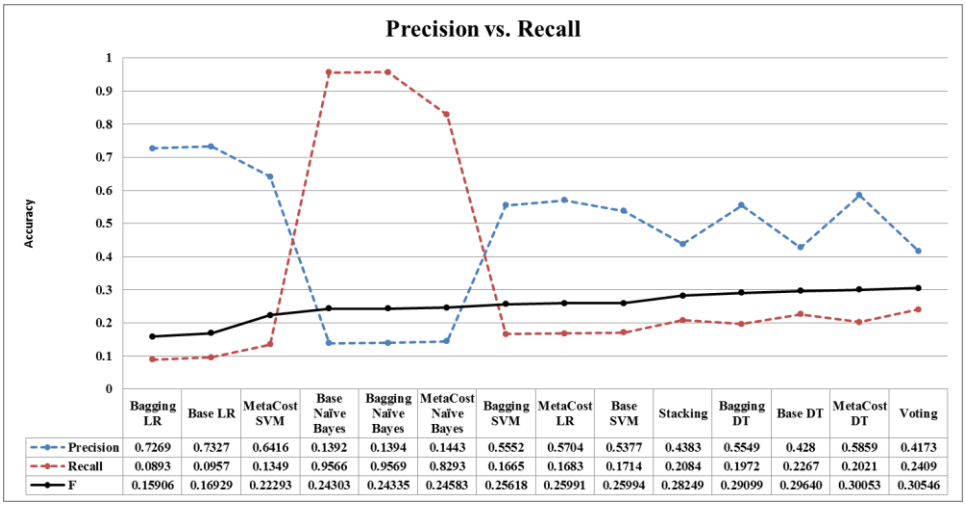
While the remaining 11 models have similar overall accuracies, the majority of accuracy differences are statistically significant at the .05, and the .01 significance level, based on a matched pair *t* test with the Bonferroni correction (Bonferroni, 1936) depicted by shading, as detailed in Exhibit 3.9. Given the larger number of pair-wise model accuracy comparisons, the experiment-wise error rate (Familywise Error Rate) is elevated. Thus, we applied the Bonferroni correction to determine whether each pair-wise model comparison was actually statistically significant. We divided both alpha levels of 0.05 and 0.01 by 14 (the number of models being compared) to determine our significance thresholds. It is interesting to note that the base logistic regression model achieved the highest overall accuracy, although *not* statistically different from the MetaCost Decision Trees and MetaCost Support Vector Machines ensemble models. As a reminder, overall model accuracy is not an appropriate evaluation metric for classification when there are asymmetric costs associated with misclassification error. Relying on overall accuracy in a classification context as this can be misleading; however, the metric can provide a comparative baseline for additional analysis.

	Naïve Bayes	Bagging Naïve Bayes	MetaCost Naïve Bayes	Voting	Decision Trees	Stacking	SVM	Bagging SVM	Bagging Decision Trees	MetaCost Logistic Regression	Bagging Logistic Regression	MetaCost Decision Trees	MetaCost SVM	Logistic Regression
Naïve Bayes														
Bagging Naïve Bayes	0.1722													
MetaCost Naïve Bayes	0.0000	0.0000												
Voting	0.0000	0.0000	0.0000											
Decision Trees	0.0000	0.0000	0.0000	0.0348										
Stacking	0.0000	0.0000	0.0000	0.0050	0.0152									
SVM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000								
Bagging SVM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0833							
Bagging Decision Trees	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0755	0.4068						
MetaCost Logistic Regression	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0271	0.0698	0.1927					
Bagging Logistic Regression	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0030	0.0005	0.0231	0.0094				
MetaCost Decision Trees	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0009	0.0000	0.0017	0.0003	0.3211			
MetaCost SVM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0011	0.0000	0.0043	0.0001	0.2214	0.4441		
Logistic Regression	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0007	0.0000	0.0044	0.0003	0.0038	0.3172	0.3052	

Exhibit 3.9. Overall Accuracy: P-values < .05 and .01 Level, Bonferroni Adjusted

The classification evaluation metrics of precision and recall can provide more granular performance details when required. Precision and recall are calculated from the results contained in a confusion matrix and were previously defined in Exhibit 3.4. Precision is a measure of accuracy provided a specific class for an instance was predicted. In our research, precision describes the percentage of correct Success predictions. As shown in Exhibit 3.10, once again the base model

logistic regression has the highest precision. The ranking of model precision is very similar to overall accuracy performance. Recall is a measure of accuracy or ability of a classification model to select instances from a specific class. Recall is also known as the True Positive Rate. In our research, recall describes the percentage of Success cases identified as such by the classifier. Exhibit 3.10 displays the Recall metrics for our research. Note the tradeoff between our precision and recall metrics. This tradeoff pattern is found in most classification research. The harmonic mean F, also shown in Exhibit 3.10, might be a more appropriate metric in other research contexts because it emphasizes the importance of small values, whereas outliers have a more significant influence on the arithmetic mean. However, the harmonic mean, which is essentially the same over the majority of our models, did not provide much insight in this research.



	Bagging LR	Base LR	MetaCost SVM	Base Naïve Bayes	Bagging Naïve Bayes	MetaCost Naïve Bayes	Bagging SVM	MetaCost LR	Base SVM	Stacking	Bagging DT	Base DT	MetaCost DT	Voting
Precision	0.7269	0.7327	0.6416	0.1392	0.1394	0.1443	0.5552	0.5704	0.5377	0.4383	0.5549	0.428	0.5859	0.4173
Recall	0.0893	0.0957	0.1349	0.9566	0.9569	0.8293	0.1665	0.1683	0.1714	0.2084	0.1972	0.2267	0.2021	0.2409
F	0.15906	0.16929	0.22293	0.24303	0.24335	0.24583	0.25618	0.25991	0.25994	0.28249	0.29099	0.29640	0.30053	0.30546

Exhibit 3.10. Model Precision vs Recall

Although calculating the overall model accuracy, precision, and recall evaluation metrics are helpful and can provide some research direction, our primary goal is to discover which classification methods predict the most campaign profit. As previously discussed, we define Profit per PPC Campaign as the number of purchases from a campaign multiplied by the revenue per conversion minus the campaign costs. The total campaign portfolio profit is the sum of profits from all campaigns selected by the classifier. A profit evaluation is much more applicable to our research because the main business objective is maximizing the overall campaign portfolio profit and not overall model accuracy. In numerous business applications such as fraud detection, medical diagnosis and marketing campaigns such as ours, the primary focus for the researcher, is in fact, the minority class. In these applications, not only are the data distributions skewed, but also the

misclassification costs associated with confusion matrix elements. A majority of classification algorithms assume equal misclassification costs and ignore the different types of misclassification errors. One practical method to minimize this problem is to utilize a cost sensitive classification algorithm (Viaene and Dedene, 2005). As discussed previously, cost sensitive classification techniques, such as MetaCost, include misclassification costs during model training and create a classifier that has the lowest cost associated with its confusion matrix. In our case, the MetaCost ensemble assigns higher cost to false negatives than to false positive which improves the classifier performance with respect to the positive class of Success.

Exhibit 3.11 summarizes the campaign portfolio profit, based on 30 replications, for each base and ensemble model. Not surprising, the Naïve Bayes models produced the lowest average profits, which follows the same overall accuracy pattern seen in Exhibit 3.8. The three Naïve Bayes model profits also show a much larger variation when compared to the other base and ensemble models. However, the ranking of the remaining models are quite different when compared to the ranking in Exhibit 3.8. The MetaCost Logistic Regression ensemble performed the best with an average campaign portfolio profit of $14,962, considerably higher than the second and third place ensembles.

As detailed in Exhibit 3.12, the Optimal model profit results are statistically different, at both the .05 and .01 significance level, from all models, using the Bonferroni correction. Although it was not surprising that the MetaCost ensemble models and the logistic regression models performed well, what is interesting is the stacking ensemble ranked second in profitability. Stacking is not known for strong performance when applied in a text mining context and since this research is based on a hybrid data set containing both standard numeric attributes and numeric attributes created from text attributes that produce a very sparse data set, its excellent performance was unexpected. We also note that all of the SVM models performed well, as expected, due to its strong text classification capabilities (Joachims, 2002).

Exhibit 3.11 illustrates how using a different evaluation metric, such as campaign portfolio profit, can provide valuable, as well as conflicting results in comparison to more traditional measures of classification accuracy. Our model results are an excellent example of why researchers should use several evaluation metrics when comparing classifier models. Classifier performance metrics

such as profits or costs are of critical importance, but are typically less robust when compared to a more stable metric such as overall accuracy. Pay-per-click campaigns can have multiple goals. As an example, a risk averse campaign analyst may wish to choose a model with a combination of stable overall accuracy and a minimum level of campaign profit, while a risk seeking campaign analyst may choose to the classifier that produces the highest campaign profit.

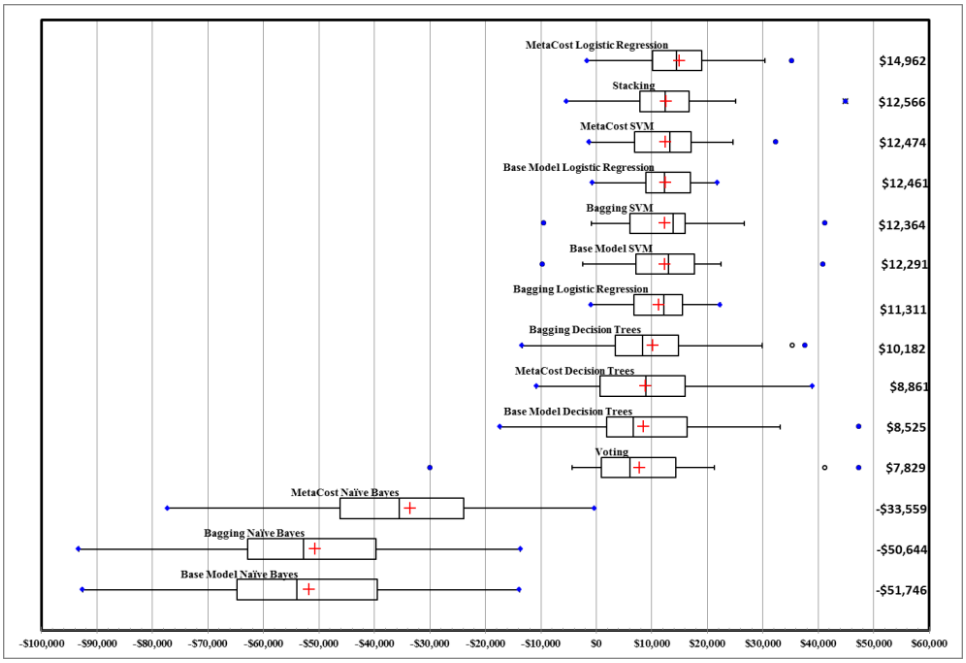


Exhibit 3.11. Model Profit per Campaign Portfollio

From a technical standpoint, our PPC campaigns have unequal profits. We have very few high profit campaigns and many lower profit campaigns. A classifier with a lower overall accuracy may still achieve a higher profit when compared to a classifier with a higher overall accuracy, because a higher overall accuracy classifier may yield its few false negatives on high profit

	Baseline Model Profit	Base Model Naïve Bayes	Bagging Naive Bayes	MetaCost Naive Bayes	Voting	Base Model Decision Trees	MetaCost Decision Trees	Bagging Decision Trees	Bagging Logistic Regression	Base Model SVM	Bagging SVM	Base Model Logistic Regression	MetaCost SVM	Stacking	MetaCost Logistic Regression	Optimum Model Profit
Baseline Model Profit																
Naive Bayes	0.0000															
Bagging Naive Bayes	0.0000	0.0113														
MetaCost Naive Bayes	0.0000	0.0000	0.0000													
Voting	0.0000	0.0000	0.0000	0.0000												
Decision Trees	0.0000	0.0000	0.0000	0.0000	0.2442											
MetaCost Decision Trees	0.0000	0.0000	0.0000	0.0000	0.3048	0.4305										
Bagging Decision Trees	0.0000	0.0000	0.0000	0.0000	0.1024	0.1628	0.1147									
Bagging Logistic Regression	0.0000	0.0000	0.0000	0.0000	0.0789	0.1169	0.1183	0.3041								
Base Model SVM	0.0000	0.0000	0.0000	0.0000	0.0084	0.0445	0.0318	0.1483	0.2606							
Bagging SVM	0.0000	0.0000	0.0000	0.0000	0.0102	0.0424	0.0410	0.1563	0.2366	0.4579						
Logistic Regression	0.0000	0.0000	0.0000	0.0000	0.0267	0.0406	0.0309	0.1271	0.0225	0.4517	0.4712					
MetaCost SVM	0.0000	0.0000	0.0000	0.0000	0.0133	0.0324	0.0139	0.1126	0.1722	0.4331	0.4629	0.4957				
Stacking	0.0000	0.0000	0.0000	0.0000	0.0130	0.0411	0.0281	0.1117	0.2230	0.3728	0.4221	0.4722	0.4740			
MetaCost Logistic Regression	0.0000	0.0000	0.0000	0.0000	0.0009	0.0012	0.0001	0.0039	0.0042	0.0281	0.0350	0.0216	0.0194	0.0526		
Optimum Model Profit	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	

Exhibit 3.12. Model Profits: P-values < .05 and .01 Level, Bonferroni Adjusted

campaigns, while a lower overall accuracy classifier may yield slightly more false negatives with respect to predominately low profit campaigns. The lower overall accuracy classifier may be correct

when it matters and thus correctly predict the very high profit campaigns. Whereas Exhibit 3.9 shows that the majority of the models statistically different in terms of overall accuracy, Exhibit 3.12, with statistical significance after Bonferroni correction depicted by shading, displays a contrasting pattern of non-significant differences in campaign profits.

6. Conclusion and future work

Our experimental results support our hypothesis that applying ensemble learning techniques in PPC marketing campaigns can achieve higher profits for our home security service and equipment provider. We introduced the evaluation metric of total campaign portfolio profit and illustrated how relying on overall model accuracy can be misleading. As additional metrics for model comparison, Exhibit 3.13 shows box plots of optimum and baseline campaign portfolio profits. Optimum

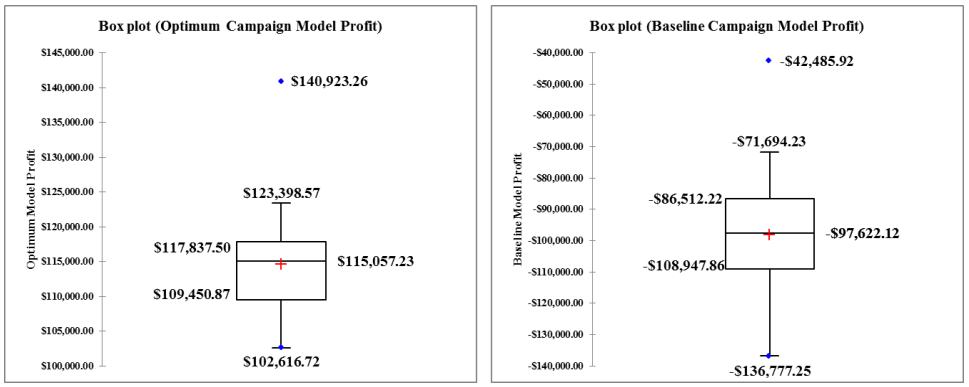


Exhibit 3.13. Optimum and Baseline Campaign Portfolio Profits

campaign portfolio profit represents the sum of the scenarios where only profitable PPC campaigns were identified and deployed. To clarify, Exhibit 3.13 illustrates the aggregate assessment of an optimum or perfect model versus a baseline or random model, on each of the holdout samples. The baseline campaign portfolio profit represents the sum of the scenarios where all PPC campaigns whether profitable or not were deployed. As can be seen, the two means of +$115,057.23 and -$97,622.12, convey a high level of variance which is a result of the marketing risk associated the PPC campaigns. Thus, any positive profit advantage achieved through improved classification techniques is beneficial. We are encouraged that all of our models outperformed the baseline campaign portfolio profit, and that the top 11 models produced positive profits. However, none of our model profits came close to the optimum campaign model profit, although all models performed consistently well, with the exception of the Naïve Bayes models. Given the difficulty of feature creation due the limited amount of words contained in an advertisement, we feel that our ensemble

framework for sponsored search advertising makes a valuable contribution to industry.

One interesting observation we noted from Exhibit 3.10 is Voting (typically used for a baseline ensemble performance) has the highest F metric based on the harmonic mean between precision and recall. Different configurations of the Voting ensemble could be a possible area for future research. Additionally, using each of the remaining dependent variables discussed in Section 4.1.1 and applying the same experiment method introduced in this article could be useful for the business community. Naïve Bayes classifiers typically perform well, making our results somewhat unexpected and a subject for future research. Additionally, feature selection methods and ensemble learning techniques could be compared.

References

http://www.iab.net/research/industry_data_and_landscape/adrevenuereport. (Last accessed September, 2014).

http://www.comscore.com/Insights/Press-Releases/2014/4/comScore-Releases-March-2014-U.S.-Search-Engine-Rankings. (Last accessed September, 2014).

Abernethy, A. M. and G. R. Franke (1996). "The Information Content of Advertising: A Meta-Analysis." Journal of Advertising 25(2): 1-17.

Abrahams, A. S., E. Coupey, E. X. Zhong, R. Barkhi and P. S. Manasantivongs (2013). "Audience Targeting by B-to-B Advertisement Classification: A Neural Network Approach." Expert Systems with Applications 40(8): 2777-2791.

Abrahams, A., R. Barkhi, E. Coupey, C. Ragsdale and L. Wallace (2014). "Converting Browsers into Recurring Customers: An Analysis of the Determinants of Sponsored Search Success for Monthly Subscription Services." Information Technology and Management 15(3): 177-197.

Adeva, J. J., García , J. M. Atxa, Pikatza, M. Carrillo, Ubeda and E. Zengotitabengoa, Ansuategi (2014). "Automatic Text Classification to Support Systematic Reviews in Medicine." Expert Systems with Applications 41(4, Part 1): 1498-1508.

Aggarwal, C. C. and C. Zhai (2012). Mining Text Data. New York, Springer.

Attenberg, J., S. Pandey and T. Suel (2009). Modeling and Predicting User Behavior in Sponsored Search. KDD-09: 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Paris, France, Association of Computing Machinery.

Benbasat, I., D. K. Goldstein and M. Mead (1987). "The Case Research Strategy in Studies of Information Systems." MIS Quarterly 11(3): 369-386.

Bradley, M.M. and P.J. Lang (2010). Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings. Technical Report C-2, The Center for Research in Psychophysiology, University of Florida.

Breiman, L., J. H. Friedman, R. A. Olsen and C. J. Stone (1984). Classification and Regression Trees. Belmont, Calif., Wadsworth International Group.

Breiman, L. (1996). "Bagging Predictors." Machine Learning. 24 (2): 123–140.

Bonferroni, C.E. (1936). Teoria statistica delle classi e calcolo delle probabilità, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8 (1936) 3-62.

Burges, C. C. (1998). "A Tutorial on Support Vector Machines for Pattern Recognition." Data Mining and Knowledge Discovery 2(2): 121-167.

Ciaramita, M., V. Murdock and V. Plachouras (2008). Online Learning from Click Data for Sponsored search. Proceedings of the 17th International Conference on World Wide Web. Beijing, China, ACM: 227-236.

D'Avanzo, E., T. Kuflik and A. Elia (2011). Online Advertising Using Linguistic Knowledge Information Technology and Innovation Trends in Organizations. Information Technology and Innovation Trends in Organizations. A. D'Atri, M. Ferrara, J. F. George and P. Spagnoletti, Physica-Verlag HD: 143-150.

Dietterich, T. G. (1997). "Machine-learning Research - Four current directions." AI Magazine 18(4): 97-136.

Domingos, P., and M. J. Pazzani (1997). "Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier." Machine Learning 29, 103-130.

Domingos, P. (1999). MetaCost: A General Method for Making Classifiers Cost-sensitive. Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, California, USA, ACM**:** 155-164.

Eisenhardt, K. M. (1989). "Building Theories from Case Study Research." The Academy of Management Review 14(4): 532-550.

Elkan C. (2001). The Foundations of Cost-Sensitive Learning. Proceedings of the Seventeenth International Conference on Artificial Intelligence. Seattle 2001, Washington, USA, August 2001, 973-978.

Fain, D., C. and J. Pedersen, O. (2005). "Sponsored Search: A Brief History." Bulletin of the American Society for Information Science and Technology 32(2): 12.

Feldman, R. (2013). "Techniques and Applications for Sentiment Analysis." Communications of the ACM 56(4): 82-89.

Ghose, A. and S. Yang (2008). Comparing Performance Metrics in Organic Search With Sponsored Search Advertising. Proceedings of the 2nd International Workshop on Data Mining and Audience

Intelligence for Advertising. Las Vegas, Nevada, ACM**:** 18-26.

Ghose, A. and S. Yang (2009). "An Empirical Analysis of Search Engine Advertising: Sponsored Search in Electronic Markets." Management Science 55(10): 1605-1622.

Ghose, A. and P. G. Ipeirotis (2011). "Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics." IEEE Transactions on Knowledge & Data Engineering 23(10): 1498-1512.

Graepel, T., J. Candela, T. Borchert and R. Herbrich (2010). Web-Scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine. 27th International Conference on Machine Learning, Haifa, Israel.

Haans, H., N. Raassens and R. Hout (2013). "Search Engine Advertisements: The Impact of Advertising Statements on Click-through and Conversion Rates." Marketing Letters 24(2): 151-163.

Heath, R. and A. Nairn (2005). "Measuring Affective Advertising: Implications of Low Attention Processing on Recall." Journal of Advertising Research 45(2): 269-281.

Jansen, B. and S. Schuster (2011). "Bidding on the Buying Funnel for Sponsored Search Campaigns." Journal of Electronic Commerce Research 12(1): 1.

Jansen, B. J. and T. Mullen (2008). "Sponsored Search: an Overview of the Concept, History, and Technology." International Journal of Electronic Business 6(2): 114-131.

Jansen, B. J., K. Sobel and M. Zhang (2011). "The Brand Effect of Key Phrases and Advertisements in Sponsored Search." International Journal of Electronic Commerce 16(1): 77-106.

Joachims, T. (2002). Learning to classify text using support vector machines. Scitech Book News. Portland, Ringgold Inc. 26.

Kantardzic, M. (2011). Data Mining: Concepts, Models, Methods, and Algorithms. Hoboken, N.J., John Wiley: IEEE Press.

Kim, J., K. Choi, G. Kim and Y. Suh (2012). "Classification Cost: An Empirical Comparison Among Traditional Classifier, Cost-Sensitive Classifier, and MetaCost." Expert Systems with Applications 39(4): 4013-4019.

Kim, Y., W. N. Street, G. J. Russell and F. Menczer (2005). "Customer Targeting: A Neural Network Approach Guided by Genetic Algorithms." Management Science 51(2): 264-276.

Kim, YongSeog. (2009). “Boosting and Measuring the Performance of Ensembles for a Successful Database Marketing.” Expert Systems with Applications 36(2): 2161-2176.

King, M. A., A. S. Abrahams and C. T. Ragsdale (2014). "Ensemble Methods for Advanced Skier Days Prediction." Expert Systems with Applications 41(4, Part 1): 1176-1188.

Kotsiantis, S. B. (2007). "Supervised Machine Learning: A Review of Classification Techniques." Informatica 31(3): 249-268.

Kuncheva, L. I. (2004). Combining Pattern Classifiers Methods and Algorithms, Wiley.

Li, D., L. H. Liu and Z. X. Zhang (2013). Research of Text Categorization on WEKA. 2013 Third International Conference on Intelligent System Design and Engineering Applications, New York, IEEE.

Mahdian, M. and K. Tomak (2008). "Pay-Per-Action Model for On-line Advertising." International Journal of Electronic Commerce 13(2): 113-128.

Major, R. L. and C. T. Ragsdale (2001). "Aggregating Expert Predictions in a Networked Environment." Computers and Operations Research 28(12): 1231-1244.

Miner, G., D. Delen, J. F. Elder, A. Fast, T. Hill and R. A. Nisbet (2012). Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. Waltham, MA, Academic Press.

Moraes, R., J. F. Valiati and W. P. Gavião Neto (2013). "Document-level Sentiment Classification: An Empirical Comparison Between SVM and ANN." Expert Systems with Applications 40(2): 621-633.

Moran, M. and B. Hunt (2009). Search Engine Marketing, Inc. : Driving Search Traffic to Your Company's Web Site, IBM Press/Pearson.

Neuendorf, K. A. (2002). The Content Analysis Guidebook. Thousand Oaks, Calif., Sage Publications.

Nielsen, F. Å. (2011). "A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs." from http://arxiv.org/abs/1103.2903.

Nielsen, J. H., S. A. Shapiro and C. H. Mason (2010). "Emotionally and Semantic Onsets: Exploring Orienting Attention Responses in Advertising." Journal of Marketing Research 47(6): 1138-1150.

Ott, R. L. and M. Longnecker (2001). An Introduction to Statistical Methods and Data Analysis. Pacific Grove, Calif., Duxbury - Thomson Learning.

Polikar, R. (2012). Ensemble Learning. Ensemble Machine Learning. C. Zhang and Y. Ma. New York, Springer US**:** 1-34.

Quinlan, J. (1986). "Introduction of Decision Trees." Machine Learning 81-106.

Richardson, M., E. Dominowska and R. Ragno (2007). Predicting Clicks: Estimating the Click-through Rate for New Ads. Proceedings of the 16th International Conference on World Wide Web. Banff, Alberta, Canada, ACM**:** 521-530.

Rokach, L. (2009). "Taxonomy for Characterizing Ensemble Methods in Classification Tasks: A Review and Annotated Bibliography." Computational Statistics and Data Analysis 53(12): 4046-4072.

Rutz, O. J. and R. E. Bucklin. (2007). "A Model of Individual Keyword Performance in Paid Search Advertising." from http://ssrn.com/abstract=1024765 or http://dx.doi.org/10.2139/ssrn.1024765.

Rutz, O. J., M. Trusov and R. E. Bucklin (2011). "Modeling Indirect Effects of Paid Search Advertising: Which Keywords Lead to More Future Visits?" Marketing Science 30(4): 646-665.

Sammut, C. and G. I. Webb (2011). Encyclopedia of Machine Learning, New York: Springer.

Sculley, D., R. G. Malkin, S. Basu and R. J. Bayardo (2009). Predicting Bounce Rates in Sponsored Search Advertisements. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, ACM**:** 1325-1334.

Sebastiani, F. (2002). "Machine Learning in Automated Text Categorization." ACM Computing Surveys 34(1): 1-47.

Shaparenko, B., Ö. Çetin and R. Iyer (2009). Data-driven Text Features for Sponsored Search Click Prediction. Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising. Paris, France, ACM: 46-54.

Shmueli, G., N. R. Patel and P. C. Bruce (2010). Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner. Hoboken, N.J., Wiley.

Sidorov, G., F. Velasquez, E. Stamatatos, A. Gelbukh and L. Chanona-Hernández (2014). "Syntactic N-grams as Machine Learning Features for Natural Language Processing." Expert Systems with Applications 41(3): 853-860.

Sigletos, G., G. Paliouras, C. D. Spyropoulos and M. Hatzopoulos (2005). "Combining Information Extraction Systems Using Voting and Stacked Generalization." Journal of Machine Learning Research 6: 1751-1782.

Stone, P. J., D. C. Dunphy and M. S. Smith (1966). The General Inquirer: A Computer Approach to Content Analysis. Oxford, England, M.I.T. Press.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll and M. Stede (2011). "Lexicon-based Methods for Sentiment Analysis." Journal of Computational Linguistics 37(2): 267-307.

Thelwall, M., K. Buckley and G. Paltoglou (2012). "Sentiment Strength Detection for the Social Web." Journal of the American Society for Information Science and Technology 63(1): 163-173.

Thelwall, M., K. Buckley, G. Paltoglou, D. Cai and A. Kappas (2010). "Sentiment Strength Detection in Short Informal Text." Journal of the American Society for Information Science and Technology 61(12): 2544-2558.

Tweedie, F. J., S. Singh and D. I. Holmes (1996). "Neural Network Applications in Sylometry: The Federalist Papers." Computers and the Humanities 30(1): 1-10.

Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. New York: Springer-Verlag.

Viaene, S. and G. Dedene (2005). "Cost-sensitive learning and decision making revisited." European Journal of Operational Research 166(1): 212-220.

Wang, G., J. Hao, J. Ma and H. Jiang (2011). "A Comparative Assessment of Ensemble Learning for Credit Scoring." Expert Systems with Applications 38(1): 223-230.

Wang, X., S. Lin, D. Kong, L. Xu, Q. Yan, S. Lai, et al. (2012). Click-Through Prediction for Sponsored Search Advertising with Hybrid Models. 2012 KDD Cup, Beijing, China.

Weiss, G., F. Provost. (2003). "Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction." The Journal of artificial intelligence research 3(19): 315-354.

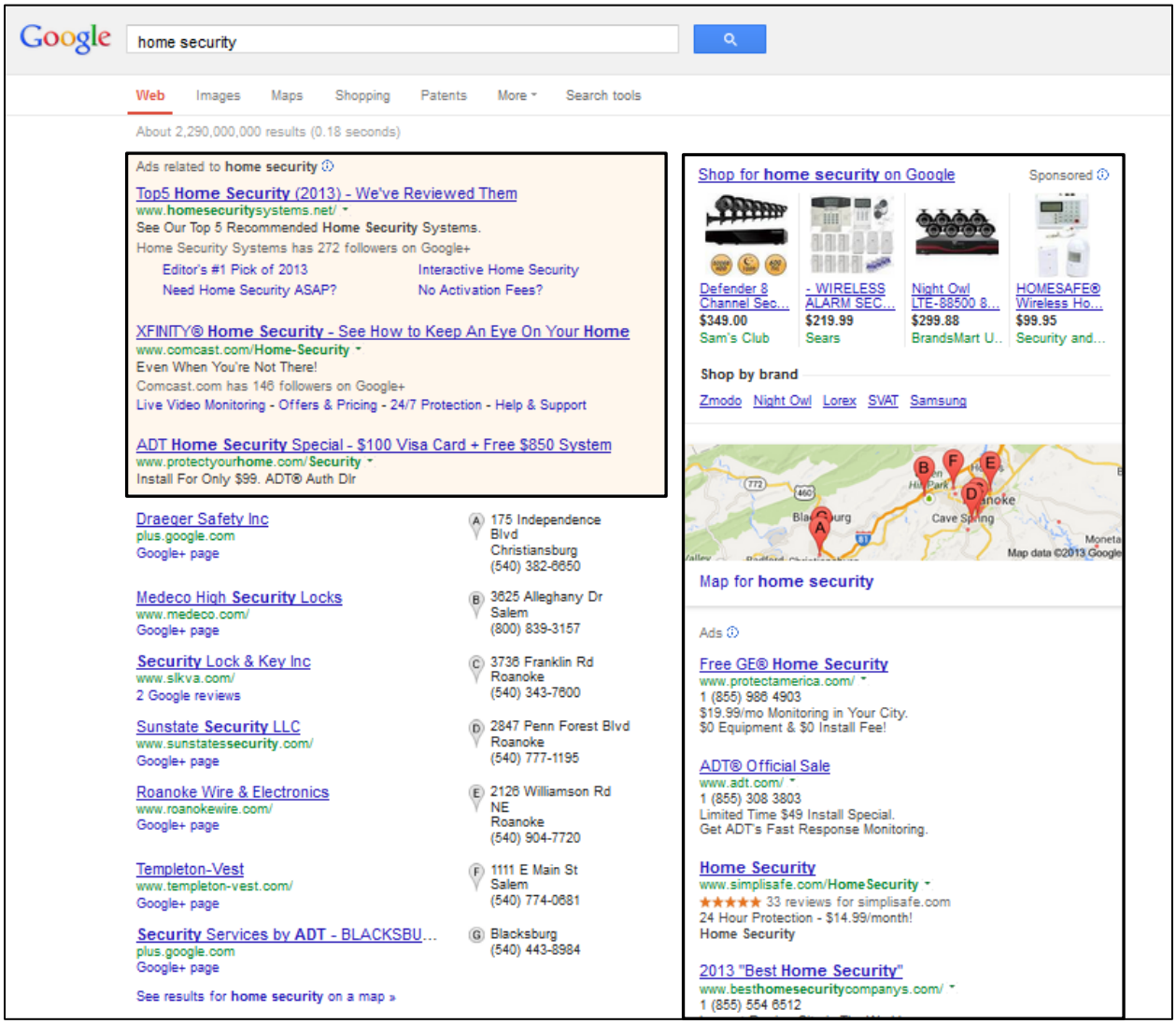
Witten, I. H., E. Frank and M. A. Hall (2011). Data Mining : Practical Machine Learning Tools and Techniques, Elsevier.

Wolpert, D. H. (1992). "Stacked Generalization." Neural Networks **5**(2): 241-259.

Yin, R. K. (2009). Case Study Research : Design and Methods. Los Angeles, Calif., Sage Publications.

Zhang, G. P. (2000). "Neural Networks for Classification: A Survey." Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 30(4): 451-462.

Appendix A: Paid Search Examples



Chapter 4

Service Improvement Using Text Analytics with Big Data

"Great minds think alike, clever minds think together."
Lior Zoref

The Internet has transformed the delivery of word of mouth communications into a more symmetric, assessable, and time sensitive form of digital information exchange. Both consumers and management now enjoy the benefits of product and service related online consumer review websites. However, unstructured consumer reviews can be emotive in nature and opaque with respect to the actual objective meaning of the review. We present the two general processes of defect discovery and remediation, and accolade discovery and aspiration, for service improvement. We believe that this research is the first to discuss the integration of GLOW and SMOKE words, as well as GLOW and SMOKE scores, derived from unstructured text, with innovative ensemble learning methods with the overall goal of service improvement.

Keywords: consumer reviews, derived features, ensemble learning, service improvement, text analytics.

1. Introduction

Word of mouth (WOM) communication has long been an important source of service information, related to the cost, perceived value, and overall satisfaction associated with a specific service (Looker, A., et al., 2007). Personal contacts, posted customer comment cards, customer phone surveys, and support personnel have provided the traditional sources of service WOM information (Peterson and Merino, 2003). Due to continuous advances and improvements associated with computers, the Internet, social media, and mobile technologies, the delivery of WOM communications is rapidly shifting away from an asymmetric form of information exchange to a more symmetric eWOM process (Clerides, S., et al., 2005; Litvin, S. W., et al., 2008; O'Connor, P., 2010). Many organizations and their customers, now enjoy numerous advantages of online consumer reviews associated with products and services (Puri, 2007; Dellarocas, C., 2003).

With the incentive to achieve and sustain a possible competitive advantage, product defect detection and prioritization by data mining online consumer reviews is becoming a management priority. Prior research used online consumer reviews to establish the nature and severity of product defects and identify SMOKE words, i.e., words typically indicative of consumer dissatisfaction with a product (Abrahams, A. S., et al., 2013; Abrahams, A. S., et al., 2012). However, using unstructured

data from online consumer reviews related to consumer services to discover SMOKE words and the associated service defects has not been analyzed. In this article, we introduce a service improvement framework, which includes GLOW and SMOKE words, that integrates traditional text mining methods with innovative ensemble learning techniques. To the best of our knowledge, this article is the first to discuss the application of ensemble learning techniques with the goal of service improvement.

We propose that two general mechanisms are available for service improvement: (1) defect discovery and remediation, and (2) accolade discovery and aspiration. When using the defect discovery and remediation process, the service provider discovers and remediates service defects. With accolade discovery and aspiration, the service provider discovers service accolades received by competitors and aspires to implement comparable improvements. We illustrate these mechanisms using representative cases from the hospitality service industry and the education service industry.

As our first case study, we used TripAdvisor.com, the largest travel media and consumer review web site, as a hospitality industry related case study for this paper. The travel industry, and hotel management in particular, has recognized a relationship between online consumer reviews and revenue as well as other metrics, such as booking rates, occupancy rates and adjusted room rates (Vermeulen and Seegers, 2009). According to a recent Travel Weekly Consumer Trends Survey (www.travelweekly.com, 2013), 58% of respondents indicated that they regularly consult travel review web sites before making a travel related purchase. More importantly, the survey discusses the increasing readership and review postings trend and the direct impact on revenue that these consumer reviews generate, despite the evidence that indicate approximately 15% of online travel reviews are fake (Duan, W., et al., 2008; Hu, N., et al., 2011).

As a second case for this research, we utilized Koofers.com, a specialized social media web site with a focus on higher education. The web site is a collaborative platform for both students and professors that provides online services that span the academic year. Although Koofers.com is primarily an online consumer review web site, it does distinguish itself by providing a wide array of online services, such as course selection guides, course study materials, grading histories, as well as job recruitment. Direct competitors to Koofers.com include RateMyProfessors and NoteSwap. The

mission of Koofers.com is to be a disruptor of higher education by moving course and professor evaluations to a synchronous information exchange. Koofers.com markets their web site by collaborating with other social media sites such as Facebook, Twitter, and LinkedIn.

We organized this paper as follows. Section 2 describes our motivation and our research contributions. Section 3 provides an overview of the travel media industry and the higher education social media industry. Section 4 discusses related research with respect to classification and ensemble learning techniques. Data set construction, brief descriptive statistics of the data sets and the subsequent methodology we followed to support our research is detailed in Section 5. Section 6 provides an analysis of our experimental results. Section 7 summarizes this work, discusses managerial implications and suggests future research opportunities.

2. Motivation and research contribution

The Cornell School of Hotel Administration recently published research that supports the business value of active consumer review management. Researchers at the school developed a revenue model that illustrates how a one point rating change on the Travelocity five point Star Rating scale could cause the average room rate to change as much as 11% for a specific hotel (Anderson, 2012).

However, consumers often write textual reviews related to services that express more information than can be captured by the common star rating scale. These textual reviews may be posted in an unstructured format on consumer review web sites, discussion forums, direct email, submissions to "contact us" online forms, Facebook, Twitter, and other social media web sites. For consumer reviews sites that do not contain an explicit Star Rating assigned by the consumer, it would be helpful for the service provider be able to automatically prioritize keywords or infer a Star Rating from ongoing submissions.

Given the overwhelming volume of free or inexpensive textual commentary, service industry executives, particularly those in service operations management, require tools to help them discover and sort the objective criticisms from the most frequently occurring and most emotionally laden compliments (Xiang, Z., et al., 2015). In this paper, we provide a novel procedure to infer a Star Rating from unstructured textual consumer reviews, and assess the accuracy of inferences across a

range of classification and ensemble learning methods. We demonstrate how ensemble learning methods can improve the overall accuracy and Kappa agreement of base classifiers, given the difficulties associated with a text mining environment.

This paper also defines a list of GLOW and SMOKE words for the hotel industry and higher education industry that are indicative of superior service as well as poor service. Extensive research streams exist related to sentiment analysis and the sentiment polarity of words, but this paper is distinctive as it investigates opportunities for service improvements, not merely incidences of consumer emotion. Prior work has shown that sentiment words, which are emotive, are distinct from SMOKE words, which are often factual or objective (Abrahams, A., et al, 2012). Thus, sentiment words can be ineffective in identifying and prioritizing consumer review postings that may relate to defects. For example, "the room key would not open the door" or "the professor was late to class" are factual, unemotive descriptions of a service defect. The inadequacy of using sentiment words to discover and prioritize product defects has been shown for automotive industry (Abrahams, A., et al, 2011), and this research attempts to ascertain whether the same distinction between sentiment words indicative of emotion, and SMOKE words indicative of a defect holds for the service industry.

MacMillan and MacGrath's (2000) Attribute Mapping framework inspired our GLOW and SMOKE word formulation proposed in this paper. We believe our process is the first concrete and semi-automatic text analytic procedure for implementing Attribute Mapping. We feel that our SMOKE and GLOW word constructs measure how well a service offering ranks in relation to other possible service offerings that may meet a consumer's needs. SMOKE words capture the essence of MacMillan and MacGrath's "Dissatisfiers," "Enragers," "Terrifiers," and "Disgusters," while GLOW words capture the essence of their "Nonnegotiables," "Differentiaters," and "Exciters." MacMillan and MacGrath argue that the most efficient way to improve a business model is to redesign or change the value proposition of the service or product offerings. In this paper, we demonstrate how to discover and prioritize service improvement opportunities from large volumes of unstructured text postings, enabling management to map out the attributes of their service offerings. Our framework enables managers to identify opportunities for eliminating "Enragers," "Terrifiers," and "Disgusters," or aspire to higher service levels by identifying and operationalizing "Nonnegotiables," "Differentiators," and "Exciters."

3. Background

In this section, we describe the background for our two case studies.

3.1 Background on the travel research and planning industry

An entire industry providing travel information collection and delivery has grown up around Web 2.0 technologies such as social media, social networks and crowd sourcing. This industry is classified as North American Industry Classification System code 561599, *All Other Travel Arrangement and Reservation Services*, however this North American Industry Classification System category is rather vague, and is more appropriately defined by *Online Travel Research and Planning* (www.hoovers.com, 2014). Industry leaders include TripAdvisor.com, Yelp, Zagat, Priceline, Orbitz, and CityGrid.

TripAdvisor.com is one of the world's largest and most popular web sites providing online travel related media (www.Forbes.com, 2013). The mission of TripAdvisor.com is to help travelers plan and experience a seamless trip by hosting user generated content related to hotels, restaurants, and attractions. TripAdvisor.com claims 62 million monthly web site visits and over 75 million user generated reviews. Although the TripAdvisor.com web site provides numerous information sources and services, the ability to read the user generated consumer reviews is the primary reason end users browse the website. Individual reviewers are able to add their own reviews concerning previous travel experiences after they create a TripAdvisor.com account. Reviewers are required to answer several questions about their travel and assign a satisfaction rating which summarizes their review. TripAdvisor.com calls the rating a Star Rating, which can range between one and five, with five representing the highest rating. TripAdvisor.com claims, to the skepticism of some travel professionals (Hu, N., et al., 2012; O'Connor, P., 2010), to scan each new review verifying whether it meets content guidelines and norms of appropriateness. Then, using a proprietary algorithm, TripAdvisor.com ranks the new review in relation to the existing reviews for the destination, sorts the list and then publishes the new results. TripAdvisor.com rarely removes a review once published; however, hospitality managers do have the opportunity to submit a specific response that is posted below the relevant consumer review.

Appendix A shows an example of a hotel review from the TripAdvisor.com web site. The review information is broken down into three sections. The top section provides basic destination information, photos, the overall ranking, the Star Rating, and acknowledgement of any TripAdvisor.com awards. The middle section provides two important pieces of information, a histogram for the Star Rating distributions and the number of reviews. The last section, which usually runs for several web pages, contains the actual reviewer assigned Star Rating, the date reviewed, the reviewer profile and most importantly, the review text.

3.2 Background for the higher education online review industry

The United States Department of Labor recently reported that postsecondary education (North American Industry Classification System 6113) is one of the fastest growing industries in the United States. The report indicated that demand for postsecondary education had steadily increased over the past several decades and is predicted to continue (http://www.bls.gov/opub/mlr/2012/01/art4full.pdf). Several demographic as well as socio-economic factors partially explain the growing demand for higher education. One of the most obvious being the children of Baby Boomers are now enrolling in numerous forms of higher education.

Current and future students are becoming well informed consumers of higher educations and have high value expectations for tuition paid. At its core, higher education is a service and educators deliver this service in a variety of settings using an assortment of methods. The quality of the teaching and support services provided by educators was traditionally disseminated by student word of mouth and well within the limited confines of an education facility. The social network connectivity provided by the Internet has disrupted the asymmetric form of controlled structured professor and course evaluations delivered to students by an educational institution. As tuition, fees, and the resulting educational debt escalate, and the probability of graduating with an undergraduate degree within four years continues to fall (Fry, R., 2014), students welcome the transparency provided by online educational review websites.

The higher education online review industry leaders are Koofers.com, RateMyProfessor.com, and MyEdu.com, all being relatively new companies. Additional smaller industry notables are Students Review.com, and KnowYourProfessor.com. MyEdu.com and Koofers.com were both founded in

2008, while RateMyProfessor.com is almost fifteen years old. Each company has struggled to monetize their online review services and have added fee based services such as job recruitment services, textbook sales and distribution, and student advertising. Koofers.com provides a rather board range of educational services that helps it compete and distinguish itself from RateMyProfessor.com and MyEdu.com. We selected Koofers.com for this case study because it is the sole remaining industry pure play. Appendix A provides an illustration of the high level of professor and course information available on one professor. Koofers.com is organized primary by university and professor name. Once a specific professor is selected, the website provides a grading history, a Star Rating, and students commends, organized by the classes taught by the professor.

4. Related research

There are numerous research streams related to text mining, classification and ensemble learning techniques. However, a few papers combine these three topics within the context of online consumer reviews. This paper helps bridge this research gap. This section describes related work in the fields of electronic word-of-mouth (eWOM), text mining of consumer reviews, and ensemble learning methods for consumer review classification.

4.1 eWOM

Traditional word-of-mouth is face-to-face communication about a consumer event, such as a purchase, an amusement activity or service. This definition has evolved over time to include any type of informal interpersonal communication about a commercial transaction (Litvin, S. W., et al., 2008). Dellarocas (2003) provide additional details relating to the benefits and complexities of digitizing word-of-mouth. Bronner and de Hoog (2011) develop a framework for understanding the motivations behind travel review postings while Cheung and Lee (2012) discovered factors, using structural equations modeling, which motivate consumers to post online reviews.

4.2 Text mining of consumer of reviews

Lahlou, F. Z., et al. (2013) provide a detailed overview of the text mining process of hotel reviews from TripAdvisor.com within the context of classification. The authors' applied support vector machine (SVM), k-nearest neighbor, and Naïve Bayes algorithms for their classification experiments and noted that the Naïve Bayes outperformed other two classifiers based on the *F* assessment metric.

Their research, as well as work by others, provides a precedent for including the Naïve Bayes classifier within our methodology (Aggarwal and Zhai, 2012; Bermejo, P., et al., 2011; Zhang, W., et al., 2011).

Two common problems associated with text mining are high dimensionality and sparsity of the resulting vector space model (Alibeigi, M., et al., 2012; Ruiz, R., et al., 2012; Yin, L., et al., 2013). The sheer volume of available data causes these two persistent problems. As discussed in more detail later in the article, feature reduction, which lowers dimensionality and reduces sparsity, is essential for satisfactory classifier performance. However, there is no one agreed upon feature reduction method for consumer review text. Several sophisticated feature selection algorithms have been proposed such as Regressional ReliefF, Fisher's criterion, Sequential Forward Selection and several variants of Mutual Information Feature Selection (Ngo-Ye and Sinha, 2012; Liu and Schumann, 2005; Bakus and Kamel, 2006; Ruiz, R., et al., 2012). Many of these newer techniques are not included in popular data mining platforms, while several correlation based feature selection variants are common (Wasikowski and X. Chen, 2010; Liu and Schumann, 2005; Isabella and Suresh, 2011; Čehovin and Bosnić, 2010).

4.3 Ensemble methods for consumer review classification

Ensemble learning techniques are general meta algorithms that can increase the accuracy of predictive or classification machine learning models (Kim, 2009). Bagheri, et al., (2013) present an ensemble taxonomy within a multiclass context, along the same lines as Rokach (2009), which groups ensemble methods into one of four model development techniques named: subsample, subspace, classifier manipulation, and subclass. The subsample and subspace techniques resample the training partition and the feature space, respectively, using several randomized strategies. A third model development method is the classifier manipulation method. The authors discuss a continuum of classifier manipulation methods ranging from a simple set of different classifiers (such as contained in a voting ensemble) to a more sophisticated grid search method of discovering the optimal set of parameters given a set of base classifiers. Lastly, the subclass technique, reduces a multiclass target into a set of binary targets, with the idea of constructing an ensemble of simpler classification problems. Although more complex than the subsample and subspace techniques, the authors indicate that the subclass process holds significant promise for multiclass text mining

problems.

Wang, et al., (2014) and Xia, et al., (2011) discuss the application of ensemble learning methods for the classification of user-generated content, more commonly known as online consumer reviews. Both authors define sentiment, illustrate several applications of sentiment classification, and make convincing cases for why sentiment classification is valuable to the greater business community. The authors provide a detailed comparative assessment of several ensemble learning techniques developed from popular classifiers, using some of the same multiclass data sets. However, the authors' discussions diverge regarding the specific ensemble learning methods utilized by their classification experiments. While Wang, et al. used the three popular ensemble methods Bagging, Boosting and Random Subspace, Xia, et al. in contrast, used generic, and thus more cumbersome, ensemble methods of aggregation rules, meta classifier, and weighted combination, for their ensemble experiments. In fact, these three generic methods are generalizations of the well-known Voting, Stacking and Boosting ensembles.

Vinodhini and Chandrasekaran (2014) present a thorough methodology for their opinion mining experiments. They provide justifications for preprocessing consumer reviews instances at the document level into unigrams, bigrams, and trigrams and, subsequently apply principal component analysis as an innovative feature reduction method (Mehta and Nejdl, 2009). The authors discuss their comparative assessment of Bagging and Bayesian Boosting using logistic regression and support vector machines as base classifiers. Although the research contains valuable insights for consumer review classification, there are several major problems that are worth noting that prevent useful generalization. The authors stated that their imbalanced data sets were small, so within that context it could be useful to generate additional bigrams and trigrams features. However, the creation and bigrams and trigrams from larger data sets, common in current data mining research, becomes computationally expensive and often untenable (Tan, C. M., et al., 2002).

5. Methodology

In the following sections, we discuss data set creation, feature creation from text, feature selection, base classifiers, and ensembles learning methods along with the experimental design we used for our analysis.

5.1 Data sets

Our data sets consist of multiple consumer reviews and we define each review as a single text "document" or "posting." For our TripAdvisor.com data set, we defined the population as the set of all consumer reviews posted on the TripAdvisor.com website. TripAdvisor.com groups their approximate 75 million reviews into three primary categories: hotel, restaurant, and attractions. We restrict our analysis for this paper to hotels. We will consider the Star Rating as a proxy for a class label, which also presents a multinomial challenge for our classification analysis. The text portion of these reviews, as a whole, will represent the corpus for this data and the subsequent text mining analysis.

To provide an objective framework for our data set creation, we chose Forbes America's Top Ten Most-Visited Cities List (Forbes, 2010). This sampling strategy was chosen in order to include the busiest hotel properties in the largest tourist cities, which account for a majority of hotel chain overall revenue. We further defined our sample frame as the top 100 hotels, as ranked by TripAdvisor.com, from these top ten cities (Peterson and Merino, 2003; O'Mahony and Smyth, 2010). Taking over six days, we crawled the TripAdvisor.com website using Helium Scraper v2.2.2.2 supported by a custom JavaScript, and imported all of the reviews from the top 100 hotels, city by city, into a comma separated value file. This data set contains 314,424 observations, each containing a Primary Key ID, City, Hotel Name, Review Title, Date, Rating, and Review. Exhibit 4.1 provides a partial example from the data set.

PK ID	City	HotelName	Review Title	Date	Rating	Review
21139	Anaheim	Abby's Anaheimer Inn	Great people comfy stay	March 6, 2008	5	I loved my stay at this place. It is a small comfy little place w
21138	Anaheim	Abby's Anaheimer Inn	Friendly quiet hotel very central to Disneyland.	February 27, 2009	4	Quiet friendly hotel, near to Disneyland. Great to use as a b:
21137	Anaheim	Abby's Anaheimer Inn	Awesome Place!!	December 14, 2009	5	Abby's is a great place to stay. We go to Disneyland severa
21136	Anaheim	Abby's Anaheimer Inn	They Lie! Waste Of Money! Rude! Discusting! Better Places Outh	April 14, 2010	1	Front desk man was rude, and would knock on our door, a
21135	Anaheim	Abby's Anaheimer Inn	Abby's is AWESOME !!!!!!!!	July 16, 2010	4	It was a great stay..the owner manager was awesome and v
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Exhibit 4.1. TripAdvisor.com Partial Data Set

We created our Koofers.com data set from data provided to us by the executive management at Koofers.com that contains 3000 randomly selected professor review records. The Koofers.com data set also contains a Star Rating class label, identical to Star Rating systems used by TripAdvisor.com. The Star Rating relates the students' subjective assessment of the professor, which is based on their perception of teaching quality, course administration, and overall course utility. The student free text form review of the professor from each record will represent the corpus for this data set and will

be used in the subsequent text mining analysis. After preprocessing and removing records with missing data, the final data set contains 2619 records, each containing a Record ID, University ID, Star Rating, Word Count, and Review. Exhibit 4.2 illustrates a partial example of the data set.

Record ID	UniversityID	StarRating	Word Count	Review
1	1	5	58	I Was Very Worried About Statics Initially Since A Lot Of My Upperclassmen Friends Would
2	1	5	52	Although Prof. Knepp Is Not The Person That Teaches The Class The Few Times He Went O
3	1	5	85	He Knew Which Parts Of The Year To Do Experiments. Fall Weather Is Dry So We Did Exp
4	1	5	69	Loved Her! Go To Class And Contribute During Discussions And She Will Love You Too. Sh
5	1	4	83	I Started As An Hd Major And Switched Because Of How Boring This Class Was It'S Super
⋮	⋮	⋮	⋮	⋮

Exhibit 4.2. Koofers.com Partial Data Set

Appendix B provides basic description statistics for each data set. The TripAdvisor.com Cities table lists the top ten cities sorted by the number of reviews scraped. The Star Rating distribution section provides the percentage of the total number of reviews represented by each Star Rating. Our Star Rating distribution is comparable to other data sets created from TripAdvisor.com (Jeacle and Carter, 2011). These percentages also may seem reversed. It is human nature to communicate bad experiences by word of mouth or eWOM (Cheung and Lee, 2012). However, people may be more inclined to take the time to write a consumer review when they had a positive experience, and may consider writing a consumer review for a negative experience as a waste of time. We noted the high combined percentage of Five Star and Four Star Ratings. Some researchers argue that consumers should be cautious when reading reviews posted on TripAdvisor.com, Yelp, as well as other reviews site, because most display this disproportionate rating distribution (Hu, N., et al., 2012).

For the Koofers.com data set, the Star Rating distribution is skewed towards the five Star Rating, which is very similar to the TripAdvisor.com data set. However, the mean and standard deviation from the Koofers.com word counts are each very similar. In contrast, the results from the TripAdvisor.com word count analysis are rather skewed.

An imbalanced data set containing a binary class label poses difficulties for most classification algorithms; however, having data sets containing a multinomial class label escalates the modeling difficulties (Drummond and Holte, 2003). In light of these class label concerns, we conducted an exploratory analysis applying replications of decision trees, logistic regression, and Naïve Bayes, discussed in more detail in Section 5.6, to assess the average accuracy. The relatively low overall accuracies we observed from our test model replications, gave us an early indication of classification

difficulties related to the imbalanced nature of our multinomial data set.

Acknowledging the respective difficulties associated with text mining and multinomial classification, we transformed the Star Rating label contained in both data sets to a binary label (Furnkranz, J., 2002; Galar, M., A., et al., 2011; Galar, M., A., et al., 2014). For both data sets, we converted a Star Ratings of five and or four to *High*. For the TripAdvisor.com data set, we based our decision on the logic that a majority of travelers would stay at a hotel with Star Rating of five or four. We also applied the same logic to the Koofers.com data set; assuming that the majority of students would prefer to take a course from a professor assigned a five or four Star Rating. For each data set, we assigned the label *Low* to a Star Rating of one, two, or three. We assume that the Star Rating scale is not linear, and if a hotel or professor was assigned a Star Rating of one, two, or three, the consumer or student would probably search for an alternative.

5.2 Feature creation

Independent variable creation is one of the most complex aspects of text mining. In this section, we discuss the specific preprocessing steps required to quantify text. Exhibit 4.3 provides a visual for the overall text mining process we follow for our analysis.

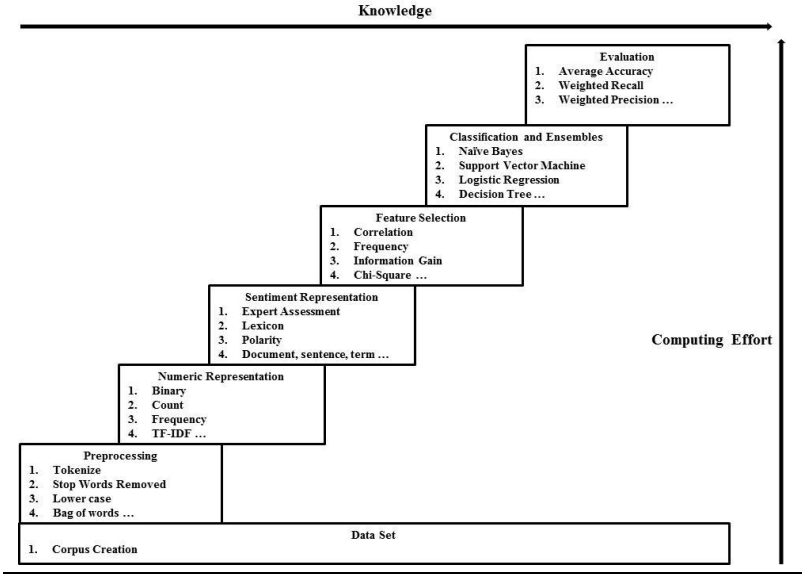


Exhibit 4.3. The Text Mining Process

One of the most challenging aspects of text mining is the curse of dimensionality (Bellman, 1961) associated with a newly created feature set. From our TripAdvisor.com data set of 314,424 records,

the document-term vector indexing process, isolated over 500,000 features. This large number of independent variables is a problem in and of itself for most supervised learning applications; however, the more problematic issue is the sparsity of the resulting document-term vector (Liu and Motoda, 2007). As an example, given one specific record where the text review contains 15 words, the values for the approximate remaining 499,985 attributes are zeroes. This sparsity causes the observations from a data set to appear almost identical, which makes it difficult to train a supervised classifier algorithm and produce an applicable predictive model. To help reduce the computing complexity associated with high dimensionality and sparsity, we chose to create a ten percent random sample from the TripAdvisor.com data set stratified by the binary Star Rating label (Čehovin and Bosnić, 2010). Although not a perfect solution, this sampling step made the data set tenable for our classification and ensemble algorithms. We used the full Koofers.com data set for the second case analysis.

5.2.1 Representing Text

The first process for quantifying unstructured text is creating a bag or multiset of words from all of the words contained in the review corpus. Each review is split up or tokenized based on white space in the text. We processed each token through the following four additional language dependent steps, before adding the token to the bag of words. First, we normalized each token to a lower case since we require, for example, Hotel and hotel to represent the same entity. This normalization process requires a tradeoff to be made since there are exceptions that may be required for named entities. For example, the tokens "Cliff" and "cliff" have two different meanings where the first token represents a named entity and the second, a rock precipice. However, when normalized to lower case, the named entity loses its specific connotation and takes the meaning of a rock precipice. Second, we removed stop words, such as, "is", "a", "to", and "the" using the RapidMiner English Stop Word operator since these words are connecting words supporting proper syntax and do not carry any domain specific meaning. Third, we stemmed each token to a root form using the Porter Stemmer for the English Language (Porter, 1980). The process removes inflectional endings from words having the same root and thus the same essential meaning, e.g., "reviews", "reviewing" and "reviewed" all are reduced to the same token "review." Lastly, we filtered out tokens that have a character size of two and less because they typically do not hold substantial differentiating power (Miner, G., et al., 2012). Once all these preprocessing steps are completed, we added each unique

token to the bag of words without regard to the original word order. We automated these preprocessing steps with text mining operators included in RapidMiner v5.3.015, an open source data mining workbench.

5.2.2 Quantifying unstructured text

The next major process in feature creation is to represent the bag of words numerically using a document-term vector. The formal name for this data representation model is the vector space model (Salton, G., et al., 1975). Each unique token is considered a feature and is represented as a column in an *n x m* size matrix. Each row by column numeric entry represents one or more occurrences of a specific token parsed from a single text observation. The numeric entry can be calculated from one of four common quantification methods. The first and simplest quantification method is to assign the binary digit of one to represent the presence of a specific token and zero otherwise. Binary representation is a computationally fast process; however, valuable information is lost when more than one occurrence of a token appears in the text. A second method is to count the number of unique occurrences of a token, which is also intuitive and computationally fast. As a third method, calculating the frequency that a term occurs in each observation relative to the total number of terms in each observation is somewhat more sophisticated. And fourth, the term frequency – inverse document frequency (TF-IDF) is possibly the most sophisticated and popular quantification method commonly utilized in text mining research and applications (Witten, 2005; Adeva, et al., 2014). The TF-IDF is calculated as:

$$TDIDF = TF * \log\left(\frac{N}{DF}\right)$$

The term frequency, *TF*, represents the frequency a specific token occurs in a specific document and indicates how important a specific token is for an observation. The document frequency, *DF*,

	Terms					Class
Documents	t_1	t_2	t_3	…	t_m	
d_1	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	…	$y_{1,14}$	High
d_2	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	…	$y_{2,14}$	High
d_3	$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	…	$y_{3,14}$	Low
⋮	…	…	…	…	…	…
d_n	$y_{n,1}$	$y_{n,2}$	$y_{n,3}$	…	$y_{n,m}$	High

Exhibit 4.4. Conceptual Document-Term Vector

represents the number of documents that contain a specific token. The token is more important the less it appears across documents, meaning it provides more differentiation between observations. *N*

is the number of documents in the corpus. The more often a token appears in one observation and the less it appears across all documents makes it a more discerning or differentiating variable for classification modeling. Exhibit 4.4 provides a conceptual illustration of a document-term vector, where d_n represents a text document, t_m represents the specific features created from the tokens and $y_{n,m}$ represents the value of one of the four numeric representation methods discussed above. The Class attribute represents the known class label required for supervised learning.

5.2.3 Second order features

As the degrees of sparsity and dimensionality, associated with a given document-term matrix increase, the ability of a classifier to discriminate between the observations within a data set becomes progressively more difficult. Researchers and practitioners working in a consumer review context often encounter the curse of dimensionality and its negative effects on overall accuracy as well as other performance metrics (Tu and Yang, 2013; Jun, S., et al., 2014). One option that may increase classifier performance, within this context, is to create a hybrid data set by adding second order or derived features to the original data set. We created a set of second order features grounded in concepts from the fields of natural language processing and information science. We added Net Polarity, Robertson's Selection Value (RSV), Document and Relevance Correlation (DRC), and Correlation Coefficient $\mathcal{C}$ to each case document-term vector and discuss them in detail below.

We performed a sentiment analysis by calculating the net polarity (a concept from natural language processing) of each review from each case data set by utilizing the AFINN sentiment lexicon. The AFINN lexicon, developed by Finn Arup Nielsen (2011), is a specialized list containing 2477 English words each having an assigned integer ranging from negative five to positive five. As an example, the word *breathtaking* is assigned a +5, while the word *bastard* is assigned a -5. This polarity scale represents the emotional communication embedded within the consumer review. Each word contained in the review is assigned a polarity score, if contained in the AFINN lexicon, and then all values are summed into a net polarity. Feldman (2013) provides an overview of the business value of sentiment analysis in areas of consumer reviews, as well as others business contexts. Wang, G., et al. (2014) provides a discussion related to ensemble learning techniques for sentiment classification utilizing net polarity as a feature.

The Robertson Selection Value (RSV) (Robertson, 1986) is an information science metric that measures the relative prevalence of a term *t* in a category of postings, such as *High* star reviews, versus a contrast category of postings from *Low* star reviews. A term may be a unigram, or sequence of words creating bigrams or trigrams. We used both tokens and stop words for the analysis. We used RSV as a proxy metric for our concept of GLOW and SMOKE word scores. As discussed in Section 2, GLOW and SMOKE words are relevant terms or tokens contained within a consumer review that are probabilistically associated with the class label, as in our case *High* or *Low*. Following Fan et al. (2005), we computed RSV as follows:

$$RSV = A * Log \frac{A * D}{B * C}$$

where *A* represents the number of postings from our *High* star reviews in which term *t* appears. *B* represents the number of postings from our *Low* star reviews in which term *t* appears. *C* represents the number of postings from our *High* star reviews in which *t* does not appear. *D* represents the number of postings from our *Low* star reviews in which *t* does not appear.

For example, consider the term "great," which is the most prevalent GLOW word contained in our Koofers.com data set as illustrated in Appendix C. Using *High* star reviews as the focus category the computation is as follows:

A = 454, "great" appears in 454 *High* star reviews.
B = 67, "great" appears in 67 *Low* star reviews.
C = 1270, "great" does not appear in 1270 *High* star reviews.
D = 828,"great" does not appear in 828 *Low* star reviews.

Therefore the GLOW score for "great" would be:

$$RSV_{great,High} = 292.9 = 454 * Log \frac{454 * 828}{67 * 1270}$$

The SMOKE RSV score for "great" simply inverts the focus category, and is represented as $RSV_{great,\ Low}$.

The Document and Relevance Correlation, as developed by Fan, W., et al. (2005), expands RSV by combining facets of RSV and cosine similarity. DRC is can be calculated as follows:

$$DRC = \frac{A^2}{\sqrt{A + B}}$$

The Correlation Coefficient C, as first proposed by Ng, H., et al (1997), is a modified version of the chi square statistic that removes a specific weakness of chi square when applied in an information retrieval context. In a sense, C is a "reduced" version of Chi square that has been shown to be an effective term relevancy metric (Ng, H., et al, 1997). Following Fan, W., et al. (2005), our categorical variables are the terms or tokens from the review and the associated relevance of the term to the document where C is calculated as follows:

$$C = \frac{\sqrt{N} * (A * D - C * B)}{\sqrt{(A + B) * (C + D)}}$$

where *N* is the number of reviews or documents. We calculated each of these relevance metrics for each term or token contained in both the TripAdvisor.com and the Koofers.com data sets. Since we have binary class labels, each term or token has a GLOW and SMOKE calculation associated with RSV, DRC and C. For each observation contained in the document-term vector, we computed aggregated RSV, DRC, and C scores by summing each individual RSV, DRC , and C score related to each GLOW and SMOKE words based on whether the term or token was contained in the review.

5.3 Feature selection

After we completed the text preprocessing steps, the document-term vectors for the TripAdvisor.com data set and the Koofers.com data set contained 34,211and 9,110 independent variables, respectively. Many researchers may consider these numbers of independent variables untenable (Bakus and Kamel, 2006). There are several reasons for reducing the number of independent variables. One argument for independent variable subset selection is that many of the created features may be redundant and carry little additional differentiating information. Higher levels of redundancy among independent variable can result in multicollinearity, which can invalidate any required independence assumptions and thus produce models with questionable results. Another important reason for feature selection is that most classifier algorithms perform better with smaller feature sets (Čehovin and Bosnić, 2010; Salton, 1989). This argument also applies to ensemble learning modeling since the computational costs are even higher.

There are two general approaches to feature selection for classification modeling. Exhibit 4.5, adapted from Ruiz (2012), illustrates the logical data flow for these two general methods. The filtering approach completes feature selection before constructing a classifier. The features are

ranked based on an evaluation metric such as the correlation or information gain associated with the class label. Once ranked, the features are selected using the top *k* percentage of the full set or an absolute count of the ranked features. The selected features are then used as input for a new document-term vector so this computationally expensive process does not have to be performed again. The filter methods are relatively fast and are independent of the classifiers and ensembles. One weakness associated with the filter selection method is that they are less effective at identifying redundant features (Kantardzic, M., 2011). The wrapper method of feature selection embeds a classifier within a loop and iterates over the features set until the best set of features are determined based on a specific evaluation criteria such as overall classification accuracy. The wrapper method can be optimized. However, the process comes with a high computational cost and as a result, is not as popular as other feature selection methods when larger feature sets are present.

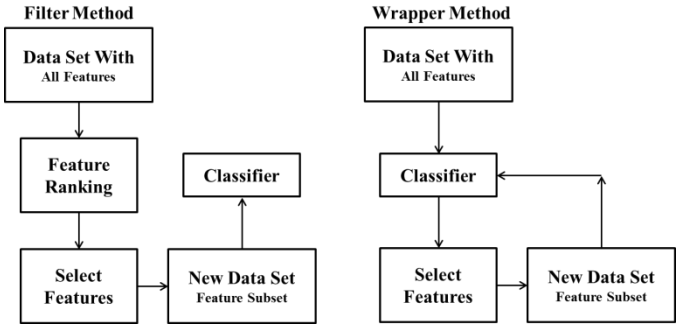


Exhibit 4.5. Approaches for Feature Selection

We chose to compare two feature selection operators from RapidMiner v5.3.015 each known for their computational efficiency: Weight by Correlation, a filter approach and Forward Selection, a wrapper approach. The Weight by Correlation operator ran quickly and produced a correlation coefficient for each feature-class vector pair. We used the Select by Weight operator to rank the feature set, to select the top one percent from the TripAdvisor.com data and the top ten percent set from the Koofers.com data set. We then saved these results as new data sets.

We then tested the Forward Selection operator, which starts with an empty feature and selects the single feature that provides the highest model performance increase. The operator adds each remaining feature, one at a time, to the existing feature set and compares the classifier performance. We used Naïve Bayes as the classifier due to its speed. Iteratively, the operator builds a feature set by choosing the single feature that produces the highest performance increase during each loop, until

a stopping condition occurs. Consequently, the first loop ran 34,211 passes. Once the best performing feature was selected, the loop ran 34,211 – 1 passes and so forth. Due to the extensive runtime required by the Forward Selection operator, we subsequently chose the Correlation by Weight operator for our analysis and retained 342 features (one percent) from the TripAdvisor.com data set and 911 features (ten percent) from the Koofers.com data set as our independent variable sets.

5.4 Supervised classifiers

We utilized three popular supervised classifiers, Naïve Bayes, logistic regression, and decision trees for this research project. All of the classifiers are well known for their text mining research streams as well as successful text mining applications (Bermejo, P., et al., 2011; Cao, Q., et al., 2011; Fuller, C. M., et al., 2011).

5.4.1 Naïve Bayes

The Naïve Bayes classifier is known for its simplicity, speed, and accuracy; making it an excellent choice for high dimensional data sets common in a text mining environment (Li, D., et al., 2013: Jiang, L., et al., 2013). The Naïve Bayes classifier is capable of integrating incremental updates, which is highly desirable from an ongoing operational viewpoint. Due the high degree of sparsity associated with most text mining data sets, it is important that the data mining platform provide the Laplace Correction option to prevent a zero probability calculation. The Naïve Bayes classifier is a probabilistic classification algorithm that assumes independence among the independent variables. In a text mining context, the independent variables are tokens derived from the text corpus. The naïve independence assumption states that tokens $T= \{t_1, t_2, t_3, \ldots t_n\}$ representing a document d_i that is classified as label C_j are all statistically independent. In our binary class environment, we calculated the predicted class C_j for review d_i the maximization of:

$$P(C_j|d_i) = argmax\left(\frac{P(C_j)\prod_{t=1}^{n} P(t_i|C_j)}{P(C_0)\prod_{t=1}^{n} P(t_i|C_0) + P(C_1)\prod_{t=1}^{n} P(t_i|C_1)}\right)$$

5.4.2 Logistic Regression

The generalized logistic regression model, as shown below, can be further simplified and applied in a binary classification context (Shmueli, G., et al., 2010). For our research, we transformed the

dependent variable Star Rating into a binary label by coding the label from each observation as High or Low. For example, following Wang (2005), let C_i represent the dependent variable having C classes. The probability of an observation with n independent variables, X_n being in class C_i, is represented as:

$$P(C_i) = \frac{e^{\left(\beta_0^{C_i} + \sum_{n=1}^{N} (\beta_n^{C_i} * X_n)\right)}}{1 + \sum_{c=1}^{C} e^{\left(\beta_0^{C_i} + \sum_{n=1}^{N} (\beta_n^{C_i} * X_n)\right)}}$$

The highest class probability calculation determines the class assigned to new observations (Shmueli, G., et al., 2010). There are several advantages of using logistic regression for text mining applications. This statistical model is based on a strong theoretical foundation, requires few actual modeling assumptions, and the results are transparent to the end user. Powel and Baker, (2007), discuss numerous text mining and sentiment analysis applications of logistic regression.

5.4.3 Decision Trees

Decision trees are one of most popular supervised classification techniques. Decision trees are intuitive and visual, which makes the classification output easy to interpret and to apply. These classifiers require minimal modeling assumptions, can handle missing data, require minimal parameter settings, and accept a multi-class dependent variable. One important advantage of using a decision tree is that the most discriminating features are automatically selected, while partitioning the training data set. This characteristic could be very useful when applying decision trees within a text mining application, because of the high dimensionality and sparsity common to most text data sets. In addition, once trained, the classifier is relatively fast classifying a new observation (Kantardzic, 2011).

Decision trees induce classification rules in a top down hierarchical manner by recursively partitioning a training data set into increasingly homogenous subsets. The homogeneity of each partition is measured by information gain, the Gini index, as well as other metrics (Shmueli, G., et al., 2010).

5.5 Ensemble learning methods

An ensemble meta-algorithm takes a set of classifiers and aggregates their individual results by an algebraic method. The basic motivation behind ensemble learning is that creating a committee of

experts is much easier than attempting to locate or derive a single genius. The accuracy or prediction from a committee of experts can be higher than any single committee member because an effective aggregation process has a tendency to remove uncorrelated errors and outlier results. Numerous empirical studies have shown that the combined prediction from a set of classifiers is more accurate than any individual classifier from the set (Lozano and Acuña, 2011; Major and Ragsdale, 2000). At a minimum, there are two ensemble design characteristics, diversity and accurate models, required to provide the potential for increased ensemble accuracy (Dietterich, 2000). Ensembles create diversity by allowing classifiers to make different errors on new or unseen observations. An accurate model is defined as a classifier that can attain an overall accuracy of greater than 50%, i.e. better than random guessing.

Over the past decade, ensemble learning research has become popular within the data mining community. However, due to the existence of numerous and at times conflicting research streams, the successful application of ensembles to specific problem contexts has been unstructured and seemingly random. Rokach (2009) proposed an ensemble taxonomy to provide a guide for future ensemble design and application. The taxonomy contains five dimensions: member dependency, diversity generation, combiners, ensemble size, and cross inducer. For our research, we selected four ensemble methods guided by Rokach's taxonomy, and supported by Bagheri, M. A., et al., (2013). Exhibit 4.6 illustrates four taxonomy dimensions along with our selected ensemble.

Ensemble Dimension	**Characteristic**	**Ensemble**
Ensemble size	Number of classifiers	Voting
Diversity generation	Training partitions	Bagging
Combiners	Results aggregation	Stacking
Cross Inducer	Classifier ensemble relation	Random Forest

Exhibit 4.6. Ensemble Taxonomy

5.5.1 Voting

The Voting meta-algorithm is the most intuitive ensemble learning method for aggregating classification results from a set of classifiers. Voting is normally used to combine the results from a set of different classification algorithms, although it sometimes used with a set classifiers constructed from the same base model, each using different parameter settings. Due to its speed and intuitive appeal, it is a common practice for researchers to use Voting as a base line measure when comparing the performance of multiple ensemble models (Sigletos, G., et al., 2005). For a binary

class problem, Voting gives the classification result based on a majority vote with respect to the class label.

5.5.2 Boot Strap Aggregation

Boot Strap Aggregation, as developed by Leo Breiman (1996), has become one of the most popular ensemble learning techniques due to its simplicity, performance and computational efficiency (Polikar, 2012). Boot Strap Aggregation, also known as Bagging, has a well established research stream that shows performance improvements for popular classifiers across numerous applications (Dietterich, T. G., 2000; Kim and Kang, 2010; Das, et al., 2009; Kim, 2009; Wang, G., et al., 2011). The Bagging meta-algorithm develops model diversity by constructing a set of new training partitions the same size as the original training partition by sampling with replacement from the original training partition. These new training partitions are used to train a set of classifiers where the subsequent results are aggregated by majority vote or averaging. Due to the resampling process, the ensemble members are trained on approximately 63% of the observations from the training partition, because the probability of any observation being selected is $1 - \left(1 - \frac{1}{N}\right)^N$. As N approaches infinity, the probability of any observation being selected converges to 0.6321. Therefore, in one sense, Bagging is a special case of classifier averaging similar to Voting.

5.5.3 Stacked Generalization

Stacked Generalization or Stacking is a hierarchical ensemble learning algorithm, that contains two logical tiers, where a second-level classifier aggregates the results from first-level classifiers (Wolpert, 1992). Typically, the first-level classifiers are each from different classification families, such as decision trees, artificial neural networks, Naïve Bayes, and so forth. The outputs from first-level classifiers, each trained from the same training partition, are used along with the original class labels as inputs for the second-level classifier. The second-level classifier is also known as a meta-learner. More formally, from training partition D_{Train} with N observations, use the leave one out validation method, and train L first-level classifiers on N-1 observations. Generate a classification from each first-level classifier for the test observation with the pattern $\{y_1, y_2, \ldots, y_L, t_N\}$ where t_N is the target for the test observation. Repeat the leave one out validation method until each observation has been treated as the test set, which yields a new training set D_{New} with N observations. Use D_{New} to train the second-level classifier (Dzeroski and Zenko, 2004).

5.5.4 Random Forests™

The Random Forests ™ ensemble, also known as decision tree forests, has become popular because the ensemble shares many of the key strengths of decision tree classifiers. Breiman and Cutler (2001) developed the Random Forest™ ensemble, subsequently trademarked the ensemble name and exclusively licensed it to Salford Systems. However, the algorithm code is publically available under the GNU General Public License (http://www.gnu.org/licenses/gpl.txt), a popular open source licensing program.

The Random Forest™ ensemble creates a set of small decision trees by combining the basic random sampling processes of bootstrapping, used in Bootstrap Aggregation with random feature selection similar to the process found in Random Subspace (Ho, 1998). The ensemble uses these two randomized sampling processes to generate the required ensemble diversity. A standard pruning process prevents over fitting by finding sub-trees that generalize beyond the current training data set and, once completed, the ensemble combines the model results by applying voting for a nominal class label or averaging for a numeric class.

The Random Forests™ ensemble is known for its ease of use, over fitting control, application versatility, and most importantly, its interpretability (Kocev, D., et al., 2013). Similar to decision trees, Random Forests™ can process noisy or missing data and then selects only the most important features during model generation. One of the most relevant strengths of the Random Forests™ ensemble, in a text mining context, is the ability to process a data set that contains an extremely large set of features, because sparcity and high dimensionality are common problems associated with text mining applications. When compared to Support Vector Machines or Artificial Neural Networks decision trees are usually less accurate. However, a Random Forests™ ensemble has accuracy on par with Support Vector Machines and Artificial Neural Networks, and a training phase an order of magnitude faster (Liaw and Wiener, 2002). The algorithm can be parallelized for more computational speed. Additionally, this ensemble has fewer model parameters and the model results are certainly more interpretable than Support Vector Machines and Artificial Neural Networks.

5.6 Evaluation metrics

We used two traditional classification evaluation metrics for model assessment and performance comparison: average accuracy and Kappa. We used the hold out data partitions, discussed in Section 5.8, from each model replication to calculate these metrics. We calculated average accuracy as follows:

$$Average\ Accuracy = \frac{\sum(T_i)}{N}$$

where T_i represents the True on-diagonal classifications from a confusion matrix and N is the total number of instances. We calculated Kappa as follows:

$$K = \frac{Proportion\ (Actual\ Agreement) - Proportion(Expected\ Agreement)}{1 - Proportion(Expected\ Agreement)}$$

The Kappa statistic adjusts overall accuracy downward by taking into account the correct predictions that may occur by chance. This calculation is the same as Cohen's Kappa Coefficient, except we use it in the context of classification evaluation. We used Cohen's traditional agreement scale for interpretation (Cohen, 1960). These two evaluation metrics will allow us to compare classifier performance across all base and ensemble model versions.

5.7 Experimental design

For both of our case studies, TripAdvisor.com and Koofers.com, we constructed nine different classification model configurations, utilizing RapidMiner v5.3.015 as our software platform. We created the three base classification models, as previously discussed, to establish baseline performance metrics. We then created one Voting ensemble containing all base models, three Bagging ensembles (one for each classifier), one Stacking ensemble containing all base classifiers, and one Random Forests™. Exhibit 4.7 offers a conceptual diagram of the process logic associated with our experimental procedure (Galar, M., et al., 2011).

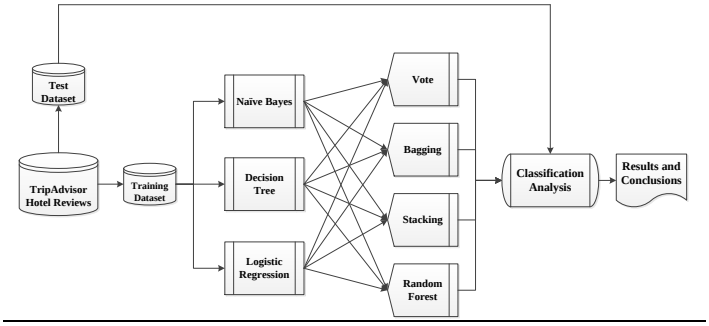


Exhibit 4.7. Conceptual Ensemble Experiment

To control our comparative environment, we followed a repeated measure experimental design, where the same *subject* is measured multiple times under different controlled *conditions*. In our experimental design, the subjects are the training and testing partitions created for a model replication and the experimental conditions are the nine different classifier model configurations. This experimental design allowed us to control the actual training and testing partition version that we subsequently used as input for each of the nine classifiers configurations during each of 30 replications (Sun, J., et al., 2011). More specifically, we used a stratified random sampling approach, based on the standard 67% / 33% ratio, which we replicated using 30 different random seeds, allowing us to create 30 different training and testing partitions. Exhibit 4.8 outlines the schema for our repeated measure experiment. It shows there will be a unique training and testing partition used as input during each replication, for each of the nine classifier model configurations. As an example, $m_{30,9}$ represents replication 30 of model configuration nine. This experimental design gave us the ability to perform the more conservative matched pair *t* tests to determine whether the differences in average accuracy between models are statistically significant (Ott and Longnecker, 2001).

	Base Model or Ensemble				
Training/Testing Partitions	1	2	3	…	9
1	$m_{1,1}$	$m_{1,2}$	$m_{1,3}$	…	$m_{1,9}$
2	$m_{2,1}$	$m_{2,2}$	$m_{2,3}$	…	$m_{2,9}$
3	$m_{3,1}$	$m_{3,2}$	$m_{3,3}$	…	$m_{3,9}$
⋮	…	…	…	…	…
30	$m_{30,1}$	$m_{30,2}$	$m_{30,3}$	…	$m_{30,9}$

Exhibit 4.8. Repeated Measure Experimental Design

We sampled the training partition a second time using a modified stratified sampling process. The RapidMiner Sampling operator balanced the training partition by applying a under sampling scheme, based on the weights we derived and illustrated in Exhibit 4.9. The weight values represent the percentage RapidMiner is to randomly sample from each class. With imbalanced data sets, traditional classifiers typically misclassify the minority class, usually the class of interest. This is even more problematic when the imbalanced data set is multinomial. Thus, it is common practice, that training partitions created from imbalanced data be balanced using one of several general methods, such as random under sampling and random over sampling. The testing or hold out sample

is not balanced, which allows for more realistic validation (Shmueli, G., et al., 2010; Yin, L., et al., 2013).

Class	Weight
Low	1.000
High	.295

Exhibit 4.9. Class Sample Weights

We selected the default parameter settings for most RapidMiner operators; however, we did make several exceptions. We selected the Laplace Correction parameter for the Naïve Bayes operator. We increased the ensemble size of the Bagging ensemble to 25 members (Kim, 2009; King, M.A., et al., 2014). We controlled our random seed values with the simple process of enumerating them 1 through 30, for partitioning the data set and any additional random process when required. The Decision Tree maximum depth was set to five, which significantly decreased computation time, as we are more interested here, in whether an ensemble can increase accuracy over the base classifier, not necessarily attempting to maximize absolute accuracy. The Random Forests™ number of trees parameter was set to 200 and the number of features randomly selected per level was set to the square root of the number features (Prinzie and Van den Poel, 2008).

6. Results and discussion

We performed our experiments on both case data sets using a personal computer with a 3.10 GHz Intel QuadCore CPU with 16 GB of RAM using the Windows 7 operating system. For each of our case studies we discuss GLOW and SMOKE word relevance, overall accuracy, and Kappa.

6.1 TripAdvisor.com results

One of the most interesting aspects from our research, which can be immediately beneficial to hotel management, was the discovery of the GLOW and SMOKE words associated with the TripAdvisor.com reviews while deriving our second order features. As a reminder, GLOW words are objective indicators of service excellence, while SMOKE words are objective indicators of service defects. We based our text analysis on the information science concept of word relevance feedback, which we measured using three metrics discussed in Section 5.2.3. We defined relevance as the degree or level of association a term or word has with our class label of High or Low. Our approach is a more sophisticated text mining approach when compared to a standard term count or

frequency analysis. Appendix C contains the most relevant terms for the TripAdvisor.com case. We observed several common themes emerging from our term relevance analysis. We observed that the top twenty GLOW words are all highly positive adjectives related to service excellence and indicative of customer satisfaction. We expected that the top twenty SMOKE words to follow a similar pattern. In contrast, the top twenty SMOKE words are split between negative adjectives related to service defects and verbs that indicate the existence of communication defects between the customer and hotel management.

We analyzed the TripAdvisor.com data set by applying both single and multiple criteria sorts. One of the most interesting observations was when we sorted by *Locality* and then by *GLOW RSV*. We could easily see the property leaders in each of the top ten markets based on the more objective *GLOW RSV* metric in contrast to the subjective Star Rating ranking. This sorting process would allow hotel managers to determine their relative standings in their market. If a hotel manager sorts the data set by *Locality* and then by *SMOKE RSV* they can easily detect service defects related to properties in a specific location. This sorting process implies a degree of defect severity, thus allowing hotel management to prioritize the service defects along with a possible solution.

As previously discussed in Section 2 a major goal of our research was the development of an automated service defect detection framework that researchers and practitioners could implement to assess and infer a more objective Star Rating from data obtained in large consumer review websites. In this research, we examined the classification efficacy of both base classifiers and ensemble classifiers with a difficult text mining context. Exhibit 4.10 documents the mean accuracy of the nine classifiers analyzed from the TripAdvisor.com case. The majority of the highest performing classifiers are ensemble configurations, with the exception of the Base Decision Tree classifier. This exception is rather interesting considering that the Base Decision Tree classifier outperformed the Random Forests™ ensemble, while the highest performing classifier was the Bagging Decision Trees ensemble. We observed that in all cases, except Naïve Bayes models, an ensemble configuration outperformed the base model configuration, i.e., the Bagging Logistic Regression outperformed Logistic Regression, and Bagging Decision Trees outperformed Decision Trees. Given the high degree of sparsity and dimensionality of the data set, we feel that these model accuracies support the case for using ensemble learning techniques for inferring whether a consumer

review rating should be High or Low.

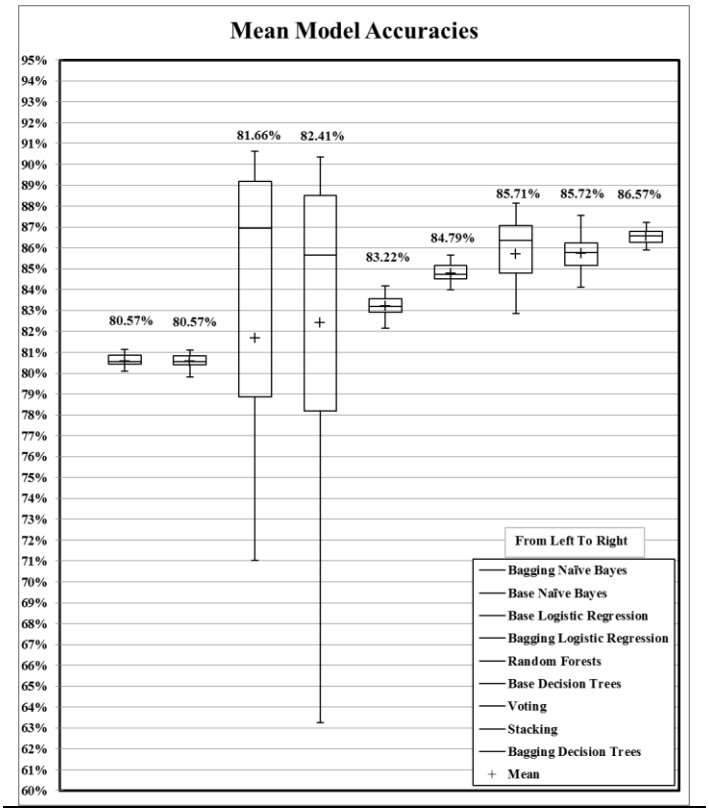


Exhibit 4.10. TripAdvisor.com Mean Model Accuracy

Exhibit 4.11 illustrates that the accuracies of the highest performing classifiers are statistically significantly better in terms of overall accuracy, than lower performing classifiers. We also noted how well the Voting ensemble performed. We did not expect this level of performance for the

	Bagging Naïve Bayes	Base Naïve Bayes	Base Logistic Regression	Bagging Logistic Regression	Random Forests	Base Decision Trees	Voting	Stacking	Bagging Decision Trees
Bagging Naïve Bayes									
Base Naïve Bayes	0.4856								
Base Logistic Regression	0.3130	0.3131							
Bagging Logistic Regression	0.1132	0.1133	0.2585						
Random Forests	0.0000 **	0.0000 **	0.2438	0.2970					
Base Decision Trees	0.0000 **	0.0000 **	0.0840	0.0632	0.0000 **				
Voting	0.0000 **	0.0000 **	0.0231	0.0053 *	0.0000 **	0.0076			
Stacking	0.0000 **	0.0000 **	0.0393	0.0185	0.0000 **	0.0000 **	0.4901		
Bagging Decision Trees	0.0000 **	0.0000 **	0.0171	0.0049 *	0.0000 **	0.0000 **	0.0113	0.0000 **	

Exhibit 4.11. Overall Accuracy: P-values < .05 * and .01 ** Level, Bonferroni Adjusted

Voting ensemble, as it is a rather simplistic ensemble that is typically used to ascertain baseline performance. As the accuracies of the Bagging Decision Trees and the Stacking ensembles are not statistically different from Voting, the case could be made based on simplicity, transparency, and speed to operationalize Voting. RapidMiner uses majority vote as the default aggregation method. In future work, we could change the aggregation method to plurality, or to confidence voting

methods such as a sum rule, or a product rule. We corrected for experiment wise error using the Bonferroni adjustment.

It is good practice to assess classifier performance using more than one performance evaluation method. We used Kappa as a second performance metric, which represents the correspondence between the actual label of High or Low and the predicted label of High, or Low. As illustrated in Exhibit 4.12, each model has a Kappa metric representing moderate agreement or higher.

Bagging Naïve Bayes	Base Naïve Bayes	Base Logistic Regression	Random Forests	Bagging Logistic Regression	Base Decision Trees	Voting	Stacking	Bagging Decision Trees
0.5082	0.5082	0.5871	0.5903	0.6058	0.6140	0.6338	0.6350	0.6556

Exhibit 4.12. Mean Model Kappa

6.2 Koofers.com results

Appendix C contains the list of the most relevant words derived from our Koofers.com data set. The GLOW words from our assessment are straightforward and intuitive. These top GLOW words represent positive reactions by students and in our opinion have a valence similar to "Differentiators" or "Exciters" contained in McGrath and MacMillan's Attribute Map. Most educators would appreciate being associated with any of the top GLOW words. As for the SMOKE words, a theme of "Enragers," "Terrifers," and "Disgusters" appear pointing to a clear pattern of service defects.

Exhibit 4.13 illustrates the mean accuracy for our Koofers.com case. We observed that the Voting ensemble performed well again, which to a degree informally supports the simplicity principle of Occam's Razor. Although the mean accuracies of the Voting and Bagging Logistic Regression are not statistically different, we noted the runtime for Bagging Logistic Regression was approximately ten times as long, which begs the question why run the Bagging Logistic Regression ensemble in future trials. We felt it is rather interesting that in both of our cases, the popular and highly hyped Random Forests™ ensemble performed worse than both the Base Decision Tree and the Bagging Decision Trees ensemble. We also observed that the Naïve Bayes and Bagging Naïve Bayes, performed poorly in both of our cases. This is somewhat counter intuitive, as the Naïve Bayes classifier is known for excellent performed in an applied text mining context. We noted both Base Logistic Regression and Bagging Logistic Regression performed well on the Koofers.com data set.

However, we noticed that these two classifiers ranked lower in mean accuracies from the TripAdvisor.com case. One possible reason for these results is the TripAdvisor.com data set displays a higher variance across the nine models. The variability of the TripAdvisor.com data set could be explained by the fact that it contains approximately five times the number of reviews; however, the differences in the data sets descriptive statistics shown Appendix B are not apparent.

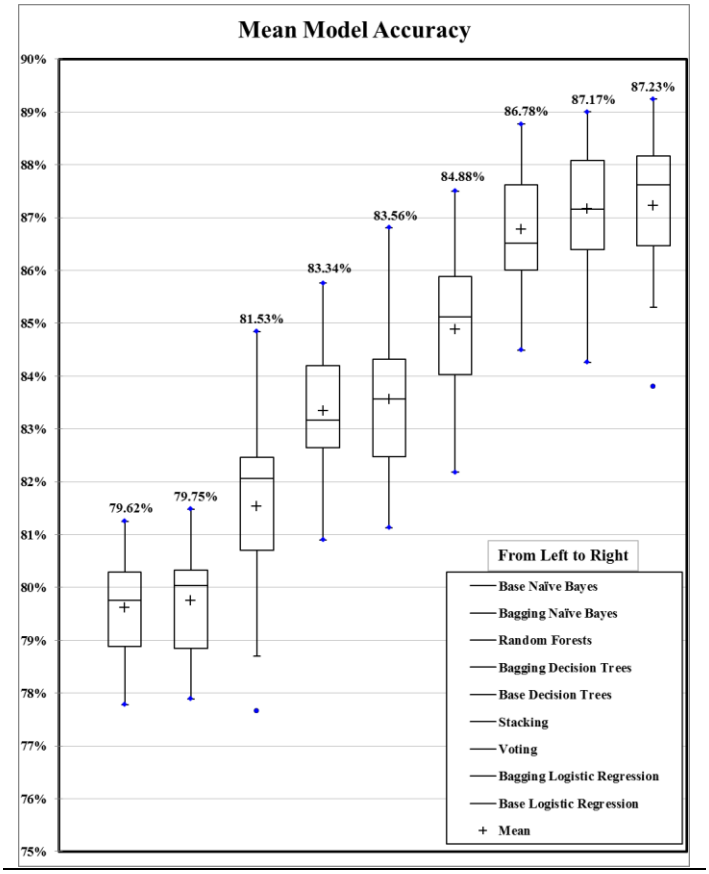


Exhibit 4.13. Koofers.com Mean Model Accuracy

Exhibit 4.14 illustrates that the accuracies of the majority of the classifiers are statistically

	Base Naïve Bayes	Bagging Naïve Bayes	Random Forests	Bagging Decision Trees	Base Decision Trees	Stacking	Voting	Bagging Logistic Regression	Base Logistic Regression
Base Naïve Bayes									
Bagging Naïve Bayes	0.0048 *								
Random Forests	0.0000 **	0.0000 **							
Bagging Decision Trees	0.0000 **	0.0000 **	0.0000 **						
Base Decision Trees	0.0000 **	0.0000 **	0.0000 **	0.2423					
Stacking	0.0000 **	0.0000 **	0.0000 **	0.0000 **	0.0000 **				
Voting	0.0000 **	0.0000 **	0.0000 **	0.0000 **	0.0000 **	0.0000 **			
Bagging Logistic Regression	0.0000 **	0.0000 **	0.0000 **	0.0000 **	0.0000 **	0.0000 **	0.0057		
Base Logistic Regression	0.0000 **	0.0000 **	0.0000 **	0.0000 **	0.0000 **	0.0000 **	0.0016 *	0.1543	

Exhibit 4.14. Overall Accuracy: P-values < .05 * and .01 Level, Bonferroni Adjusted**

different. Due to our multiple comparison context, we applied the Bonferroni correction.

We used Kappa as a second performance metric for our Koofers.com case. As illustrated in Exhibit 4.15, each model has a Kappa metric representing a moderate agreement or higher.

Base Naïve Bayes	Bagging Naïve Bayes	Random Forests	Bagging Decision Trees	Base Decision Trees	Stacking	Voting	Bagging Logistic Regression	Base Logistic Regression
0.5325	0.5340	0.6065	0.6398	0.6468	0.6743	0.7092	0.7178	0.7196

Exhibit 4.15. Mean Model Kappa

7. Conclusion, implications, and future directions

In this paper, we presented an integrated framework for service improvement identification and prioritization using text analytics. We demonstrated how our text analytic process could simplify the ongoing day-to-day management of the accelerating volume of online consumer reviews. We used online consumer reviews from two service industries as cases for the basis of our research. An objective approach to online review management can be challenging for service related industries due to the difficulties associated with the discovery of both the service provider's own service and the difficulty of discovery of aspirational service attributes in that service market that satisfy or, perhaps better still, excite consumers. The standard Star Rating system is subjective, arbitrary and inconsistent across consumer review sites. Consumer reviews are loaded with phases such as "the hotel was bad" or "the professor was strange." A Star Rating based on subjective or emotive phases provides minimal if any constructive information to guide service industry managers towards discovering a service defect or identifying a service opportunity. We transformed the difficult to interpret multinomial Star Rating label contained in both of our service related data sets into a less ambiguous and more actionable binary class label.

Applying user relevance feedback from the field of information science, we developed the constructs of GLOW and SMOKE words that allow service industry managers to infer an impartial and objective Star Rating, discover service defects and prioritize solutions in a systematic manner. The addition of these second order features to both case data sets significantly increased the overall accuracies of the base classifiers and the subsequent ensemble classifiers from a rather uninformative range in the low 50% to our superior accuracy results shown in Exhibit 4.10 and Exhibit 4.13. We argue from a service management perspective, that customer relationship management should include the active management of online reviews using GLOW and SMOKE words instead of an immediate reaction to reviews with low Star Rating. Although Star Rating systems in use on many

service related consumer review websites have known weaknesses, we suggest it cannot be ignored because of the subsequent effect a rating change can have on the linkage between a consumer experience and future purchasing decisions (Pathak, B., et al., 2010; Sparks and Browning, 2011; Yacouel and Fleischer, 2011; Anderson, 2012).

RSV scores are conveniently granular and summable, in contrast to categorical labels such as *High Star* and *Low Sta*r or Likert scale items like *1 Star* and *5 Star*, which are not summable. Thus, if we want to find the SMOKE score for a specific review, we can sum the RSV scores for all terms contained in that review. Similarly, if we want to find the SMOKE score for a hotel property, we can sum the RSV scores for all reviews that pertain to that hotel property. Likewise, if we want to find the SMOKE score for a city, we can sum the RSV scores for all reviews for all hotel properties in that. The summability of SMOKE scores makes them amenable to inclusion as a new column alongside each review, and advanced PivotTable analysis can then be performed on reviews, to determine the "smokiness" of different queries through the data set. For example, a hospitality manager could determine how "smokey" their hotel properties located in New York City are in 2014 and compare this metric to those same hotel properties in previous years. In addition, a hospitality manager could determine how "smokey" their New York City properties were last year compared to their competitors' properties in the same city. The granularity and summability of SMOKE and GLOW scores facilitates, for the first time, data pivoting on both structured data and text, and adds significant text analytic capability beyond what is possible with traditional categorical labels.

Our study is subject to several limitations. We developed our text analytic framework relying on the results from two online consumer review sources. We caution that our analysis may not generalize to other service related industries. Implementation and testing of our framework in additional service industries such as, automotive sales and service, consulting, food and beverage industry, or health care could provide additional evidence of generalizability. Our data sets could contain a structural bias, due to the self-selection nature of the review submission process. It may seem intuitive that reviews that contain complaints with an associated low rating would represent the majority, while in fact quite the opposite is true. TripAdvisor.com has publically acknowledged that approximately 75% of their posted reviews contain a Star Rating of four or five. Lastly, our Koofers.com data set is relatively small, which may also limit the generalizability of the analysis.

We feel that we have several future research opportunities. At a minimum, we would like to construct data sets from several additional popular online consumer review websites and compare the results with our results from this research. The main issue left to address is whether our framework generalizes to other service related consumer review websites. Another possible research idea is to create a new TripAdvisor.com data set and run the analysis using an updateable Naïve Bayes algorithm. We then would wait a period of time, then add new reviews from the TripAdvisor.com website and assess the performance of the classifier. Using an updateable version of the Naïve Bayes algorithm could greatly simplify the adoption and ongoing operation of our automated service defect detection framework.

References

Abrahams, A. S., J. Jiao, G. A. Wang and W. Fan (2012). "Vehicle defect discovery from social media." Decision Support Systems 54(1): 87-97.

Abrahams, A. S., J. Jiao, W. Fan, G. A. Wang and Z. Zhang (2013). "What's buzzing in the blizzard of buzz? Automotive component isolation in social media postings." Decision Support Systems 55(4): 871-882.

Adeva, J. J., García , J. M. Atxa, Pikatza, M. Carrillo, Ubeda and E. Zengotitabengoa, Ansuategi (2014). "Automatic Text Classification to Support Systematic Reviews in Medicine." Expert Systems with Applications 41(4, Part 1): 1498-1508.

Aggarwal, C. and C. Zhai (2012). A Survey of Text Classification Algorithms. Mining Text Data. C. C. Aggarwal and C. Zhai, Springer US**:** 163-222.

Alibeigi, M., S. Hashemi and A. Hamzeh (2012). "DBFS: An effective Density Based Feature Selection scheme for small sample size and high dimensional imbalanced data sets." Data & Knowledge Engineering 81–82(0): 67-103.

Anderson, C. (2012). "The Impact of Social Media on Lodging Performance." Cornell Hospitality Report 12(15).

Ayeh, J. K., N. Au and R. Law (2013). "“Do We Believe in TripAdvisor?” Examining Credibility Perceptions and Online Travelers’ Attitude toward Using User-Generated Content." Journal of Travel Research 52(4): 437-452.

Bagheri, M. A., G. Qigang and S. Escalera (2013). A Framework Towards the Unification of Ensemble Classification Methods. Machine Learning and Applications (ICMLA), 2013 12th International Conference on, IEEE.

Bakus, J. and M. Kamel (2006). "Higher Order Feature Selection for Text Classification." Knowledge and Information Systems 9(4): 468-491.

Bermejo, P., J. A. Gámez and J. M. Puerta (2011). "Improving the performance of Naive Bayes multinomial in e-mail foldering by introducing distribution-based balance of datasets." Expert Systems with Applications 38(3): 2072-2080.

Bellman, Richard Ernest (1961). Adaptive Control Processes: A Guided Tour, Princeton University Press.

Breiman, L. (1996). “Bagging Predictors.” Machine Learning. 24 (2): 123–140.

Breiman, L. (1996). “Bias, variance, and Arcing classifiers.” Technical Report 460. Berkeley, California: University of California, Department of Statistics.

Breiman L. (2001). “Random Forests.” Machine Learning. 45 (1), pp 5-3

Bronner, F. and R. de Hoog (2011). "Vacationers and eWOM: Who Posts, and Why, Where, and What?" Journal of Travel Research 50(1): 15-26.

Cao, Q., W. Duan and Q. Gan (2011). "Exploring determinants of voting for the “helpfulness” of online user reviews: A text mining approach." Decision Support Systems 50(2): 511-521.

Čehovin, L. and Z. Bosnić (2010). "Empirical Evaluation of Feature Selection Methods in Classification." Intelligent Data Analysis 14(3): 265-281.

Cheung, C. M. K. and M. K. O. Lee (2012). "What Drives Consumers to Spread Electronic Word of Mouth in Online Consumer-opinion Platforms." Decision Support Systems 53(1): 218-225.

Clerides, S., P. Nearchou and P. Pashardes (2005). Intermediaries as Bundlers, Traders and Quality Assessors: the Case of UK Tour Operators. Discussion paper no. 5038. Centre for Economic Policy Research, London.

Cochrane, K. (2011). Why TripAdvisor.com is Getting a Bad Review, Guardian News and Media Limited.

Cohen, J. (1960). "A Coefficient of Agreement for Nominal Scales." Educational and Psychological Measurement 20(1): 37-46.

Das, R., I. Turkoglu and A. Sengur (2009). "Effective Diagnosis of Heart Disease Through Neural Networks Ensembles." Expert Systems with Applications 36(4): 7675-7680.

Dellarocas, C. (2003). "The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms." Management Science 49(10): 1407-1424.

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. First International Workshop on Multiple Classifier System, Lecture Notes in Computer Science. J. Kittler and F. Roli. Cagliari, Italy, Springer-Verlag**:** 1-15.

Dietterich, T. G. (2000). "An Experimental Comparison of Three Methods for Constructing

Ensembles of Decision Trees: Bagging, Boosting, and Randomization." Machine Learning 40(2): 139-157.

Drummond, C. and R. C. Holte (2003). C4.5, Class Imbalance, and Cost Sensitive: Why Under-sampling Beats Over-sampling. ICML'2003 Workshop on Learning from Imbalanced Data Sets (II) Washington, DC, USA, ICML.

Duan, W., B. Gu and A. B. Whinston (2008). "Do online reviews matter? — An empirical investigation of panel data." Decision Support Systems 45(4): 1007-1016.

Dzeroski, S. and B. Zenko (2004). "Is Combing Classifiers with Stacking Better than Selecting the Best One?" Machine Learning 54: 255-273.

Fan, W., M. D. Gordon and P. Pathak (2005). "Effective profiling of consumer information retrieval needs: a unified framework and empirical comparison." Decision Support Systems 40(2): 213-233.

Feldman, R. (2013). "Techniques and Applications for Sentiment Analysis." Communications of the ACM 56(4): 82-89.

Freund, Y. and R. E. Schapire (1996). Experiments With a New Boosting Algoritm. International Conference of Machine Learning, San Francisco, CA, Morgan Kaufmann Publishers.

Fry, R. (2014). Young Adults, Student Debt and Economic Well-being. Social and Demographic Trends Project. Washington, DC, Pew Research Center.

Fuller, C. M., D. P. Biros and D. Delen (2011). "An investigation of data and text mining methods for real world deception detection." Expert Systems with Applications 38(7): 8392-8398.

Furnkranz, J. (2002). "Round Robin Classification." Journal of Machine Learning Research 2: 721-747.

Galar, M., A. Fernández, E. Barrenechea, H. Bustince and F. Herrera (2011). "An Overview of Ensemble Methods for Binary Classifiers in Multi-class Problems: Experimental Study on One-vs-One and One-vs-All Schemes." Pattern Recognition 44(8): 1761-1776.

Galar, M., A. Fernández, E. Barrenechea and F. Herrera (2014). "Empowering Difficult Classes with a Similarity-based Aggregation in Multi-class Classification Problems." Information Sciences 264(0): 135-157.

Ho, T. K. (1998). "The Random Subspace Method for Constructing Decision forests." IEEE

Transactions on Pattern Analysis and Machine Intelligence 20(8): 832-844.

http://www.bls.gov/opub/mlr/2012/01/art4full.pdf, last accessed July, 2014.

http://forbes.com/2010/04/28/tourism-new-york;lifestyle-travel-las-vegas-cities_slide.html, last accessed January, 2011.

http://www.gnu.org/licenses/gpl.txt, last accessed October, 2014.

http://www.forbes.com/sites/greatspeculations/2013/03/08/heres-why-we-believe-tripadvisors-user-base-will-continue-to-climb/, last accessed July 2014.

http://www.hoovers.com/company-information/cs/company-profile.TripAdvisor.com_Inc.52db0c0fef6f8c9b.html, last accessed July 2014.

http://www.thetimes.co.uk/tto/money/consumeraffairs/article3095761.ece, last accessed July 2014.

http://www.travelweekly.com/print.aspx?id=250964, last accessed July 2014.

http://www.tripadvisor.com, last accessed July 2014.

Hu, N., L. Liu and V. Sambamurthy (2011). "Fraud Detection in Online Consumer Reviews." Decision Support Systems 50(3): 614-626.

Hu, N., I. Bose, N. S. Koh and L. Liu (2012). "Manipulation of online reviews: An analysis of ratings, readability, and sentiments." Decision Support Systems 52(3): 674-684.

Isabella, J. and R. M. Suresh (2011). Opinion Mining Using Correlation Based Feature Selection. International Conference on Software Engineering and Applications, Singapore, Global Science and Technology Forum.

Jeacle, I. and C. Carter (2011). "In TripAdvisor We Trust: Rankings, Calculative Regimes and Abstract Systems." Accounting, Organizations and Society 36(4–5): 293-309.

Jiang, L., Z. Cai, H. Zhang and D. Wang (2013). "Naive Bayes text classifiers: a locally weighted learning approach." Journal of Experimental and Theoretical Artificial Intelligence 25(2): 273-286.

Joshi, S. and B. Nigam (2011). Categorizing the Document Using Multi Class Classification in Data Mining. Computational Intelligence and Communication Networks (CICN), 2011 International Conference on, Gwalior, India, IEEE.

Jun, S., S.-S. Park and D.-S. Jang (2014). "Document Clustering Method Using Dimension Reduction and Support Vector Clustering to Overcome Sparseness." Expert Systems with Applications 41(7): 3204-3212.

Kantardzic, M. (2011). Data Mining: Concepts, Models, Methods, and Algorithms. Hoboken, N.J., John Wiley: IEEE Press.

Kim, YongSeog. (2009). "Boosting and Measuring the Performance of Ensembles for a Successful Database Marketing." Expert Systems with Applications 36(2): 2161-2176.

Kim, M. J. and D. K. Kang (2010). "Ensemble with Neural Networks for Bankruptcy Prediction." Expert Systems with Applications 37(4): 3373-3379.

Kim, Y. (2009). "Boosting and Measuring the Performance of Ensembles for a Successful Database Marketing." Expert Systems with Applications 36(2, Part 1): 2161-2176.

King, M. A., A. S. Abrahams and C. T. Ragsdale (2014). "Ensemble Methods for Advanced Skier Days Prediction." Expert Systems with Applications 41(4, Part 1): 1176-1188.

Kocev, D., C. Vens, J. Struyf and S. Džeroski (2013). "Tree ensembles for predicting structured outputs." Pattern Recognition 46(3): 817-833.

Kuncheva, L. I. (2004). Combining Pattern Classifiers Methods and Algorithms, Wiley.

Lahlou, F. Z., A. Mountassir, H. Benbrahim and I. Kassou (2013). A Text Classification Based Method for Context Extraction From Online Reviews. Intelligent Systems: Theories and Applications (SITA), 2013 8th International Conference on.

Li, D., L. H. Liu and Z. X. Zhang (2013). Research of Text Categorization on WEKA. 2013 Third International Conference on Intelligent System Design and Engineering Applications, New York, IEEE.

Liaw, A. and Wiener, M. (2002). "Classification and Regression by Random Forest." R News Vol. 2/3, December.

Litvin, S. W., R. E. Goldsmith and B. Pan (2008). "Electronic Word-of-Mouth in Hospitality and Tourism Management." Tourism Management 29(3): 458-468.

Liu, B. and L. Zhang (2012). A Survey of Opinion Mining and Sentiment Analysis. Mining Text

Data. C. C. Aggarwal and C. Zhai, Springer US**:** 415-463.

Liu, H. and H. Motoda (2007). Computational Methods of Feature Selection. Boca Raton, Chapman and Hall/CRC.

Liu, Y. and M. Schumann (2005). "Data Mining Feature Selection for Credit Scoring Models." Journal of the Operational Research Society 56(9): 1099-1108.

Looker, A., Rockland, D., and E. Taylor (2007). "Media Myths and Realities: A Study of 2006 Media Usage in America." Public Relations Tactics, pp. 10, 21–22.

Lozano, E. and E. Acuña (2011). Comparing Classifiers and Metaclassifiers
Advances in Data Mining. Applications and Theoretical Aspects. P. Perner, Springer Berlin / Heidelberg. 6870: 56-65.

Major, R. L. and C. T. Ragsdale (2000). "An Aggregation Approach to the Classification Problem Using Multiple Prediction Experts." Information Processing and Management 36(4): 683-696.

McGrath, R. G. and I. C. MacMillan (2000). The Entrepreneurial Mindset: Strategies for Continuously Creating Opportunity in an Age of Uncertainty. Boston, Mass., Harvard Business School Press.

Mehta, B. and W. Nejdl (2009). "Unsupervised Strategies for Shilling Detection and Robust Collaborative Filtering." User Modeling & User-Adapted Interaction 19(1/2): 65-97.

Miner, G., D. Delen, J. F. Elder, A. Fast, T. Hill and R. A. Nisbet (2012). Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. Waltham, MA, Academic Press.

Moraes, R., J. F. Valiati and W. P. Gavião Neto (2013). "Document-level Sentiment Classification: An Empirical Comparison Between SVM and ANN." Expert Systems with Applications **40**(2): 621-633.

Ngo-Ye, T. L. and A. P. Sinha (2012). "Analyzing Online Review Helpfulness Using a Regressional ReliefF-Enhanced Text Mining Method." Association of Computing Machinery Transactions on Management Information Systems 3(2): 1-20.

Nielsen, F. Å. (2011). "A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs." from http://arxiv.org/abs/1103.2903.

O'Connor, P. (2010). "Managing a Hotel's Image on TripAdvisor.com." Journal of Hospitality Marketing & Management 19(7): 754-772.

O'Mahony, M. P. and B. Smyth (2010). "A Classification-based Review Recommender." Knowledge-Based Systems 23(4): 323-329.

Ott, R. L. and M. Longnecker (2001). An Introduction to Statistical Methods and Data Analysis. Pacific Grove, Calif., Duxbury - Thomson Learning.

Pang, B. and L. J. Lee (2008). Opinion Mining and Sentiment Analysis. Hanover, MA, Now Publishers.

Pathak, B., R. Garfinkel, R. D. Gopal, R. Venkatesan and F. Yin (2010). "Empirical Analysis of the Impact of Recommender Systems on Sales." Journal of Management Information Systems 27(2): 159-188.

Peterson, R. A. and Merino, M. C. (2003). "Consumer Information Search Behavior and the Internet." Psychology and Marketing, *20*(2), 99–121.

Polikar, R. (2012). Ensemble Learning. Ensemble Machine Learning. C. Zhang and Y. Ma. New York, Springer US**:** 1-34.

Porter, M.F. (1980). "An Algorithm for Suffix Stripping, " Program, 14(3) pp 130−137.

Powell, S. G. and K. R. Baker (2007). Management Science: the Art of Modeling With Spreadsheets. Hoboken, John Wiley & Sons.

Prinzie, A. and D. Van den Poel (2008). "Random Forests for Multiclass Classification: Random MultiNomial Logit." Expert Systems with Applications 34(3): 1721-1732.

Puri, A. (2007). "The Web of Insights: The Art and Practice of Webnography." International Journal of Market Research, 49, 387–408.

Robertson, S. (1986). "On Relevance Weighting Estimation and Query Expansion." Journal of Documentation 42: 182-188.

Rokach, L. (2009). "Taxonomy for Characterizing Ensemble Methods in Classification Tasks: A Review and Annotated Bibliography." Computational Statistics and Data Analysis 53(12): 4046-4072.

Ruiz, R., J. C. Riquelme, J. S. Aguilar-Ruiz and M. García-Torres (2012). "Fast feature selection aimed at high-dimensional data via hybrid-sequential-ranked searches." Expert Systems with Applications 39(12): 11094-11102.

Salton, G., A. Wong and C. S. Yang (1975). "A Vector Space Model for Automatic Indexing." Communications of the Association for Computing Machinery 18(11): 613-620.

Salton, G. (1989). Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Reading, PA, Addison-Wesley.

Sebastiani, F. (2002). "Machine Learning in Automated Text Categorization." ACM Computing Surveys 34(1): 1-47.

Shmueli, G., N. R. Patel and P. C. Bruce (2010). Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner. Hoboken, N.J., Wiley.

Sigletos, G., G. Paliouras, C. D. Spyropoulos and M. Hatzopoulos (2005). "Combining Information Extraction Systems Using Voting and Stacked Generalization." Journal of Machine Learning Research 6: 1751-1782.

Sparks, B. A. and V. Browning (2011). "The Impact of Online Reviews on Hotel Booking Intentions and Perception of Trust." Tourism Management 32(6): 1310-1323

Sun, J., M. y. Jia and H. Li (2011). "AdaBoost Ensemble for Financial Distress Prediction: An Empirical Comparison With Data From Chinese Listed Companies." Expert Systems with Applications 38(8): 9305-9312.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll and M. Stede (2011). "Lexicon-based Methods for Sentiment Analysis." Journal of Computational Linguistics 37(2): 267-307.

Tan, C. M., Y. F. Wang and C. D. Lee (2002). "The Use of Bigrams to Enhance Text Categorization." Information Processing & Management 38(4): 529-546.

Tu, Y. and Z. Yang (2013). "An Enhanced Customer Relationship Management Classification Framework with Partial Focus Feature Reduction." Expert Systems with Applications 40(6): 2137-2146.

Vermeulen, I. E. and D. Seegers (2009). "Tried and Tested: The Impact of Online Hotel Reviews on Consumer Consideration." Tourism Management 30(1): 123-127.

Vinodhini, G. and R. M. Chandrasekaran (2014). "Measuring the Quality of Hybrid Opinion Mining Model for E- Commerce Application." Measurement.

Wang, G., J. Hao, J. Ma and H. Jiang (2011). "A Comparative Assessment of Ensemble Learning for Credit Scoring." Expert Systems with Applications 38(1): 223-230.

Wang, G., J. Sun, J. Ma, K. Xu and J. Gu (2014). "Sentiment Classification: The Contribution of Ensemble Learning." Decision Support Systems 57(0): 77-93.

Wang, Y. (2005). "A Multinomial Logistic Regression Modeling Approach for Anomaly Intrusion Detection." Computers and Security 24(8): 662-674.

Wasikowski, M. and X.-w. Chen (2010). "Combating the Small Sample Class Imbalance Problem Using Feature Selection." IEEE Transactions on Knowledge & Data Engineering 22(10): 1388-1400.

Witten, I. H. (2005). Text Mining. Practical Handbook of Internet Computing. M.P. Singh, Chapman and Hall/CRC Press, Boca Raton, Florida: 14-1 -14-22.

Wolpert, D. H. (1992). "Stacked Generalization." Neural Networks 5(2): 241-259.

Xia, R., C. Zong and S. Li (2011). "Ensemble of Feature Sets and Classification Algorithms for Sentiment Classification." Information Sciences 181(6): 1138-1152.

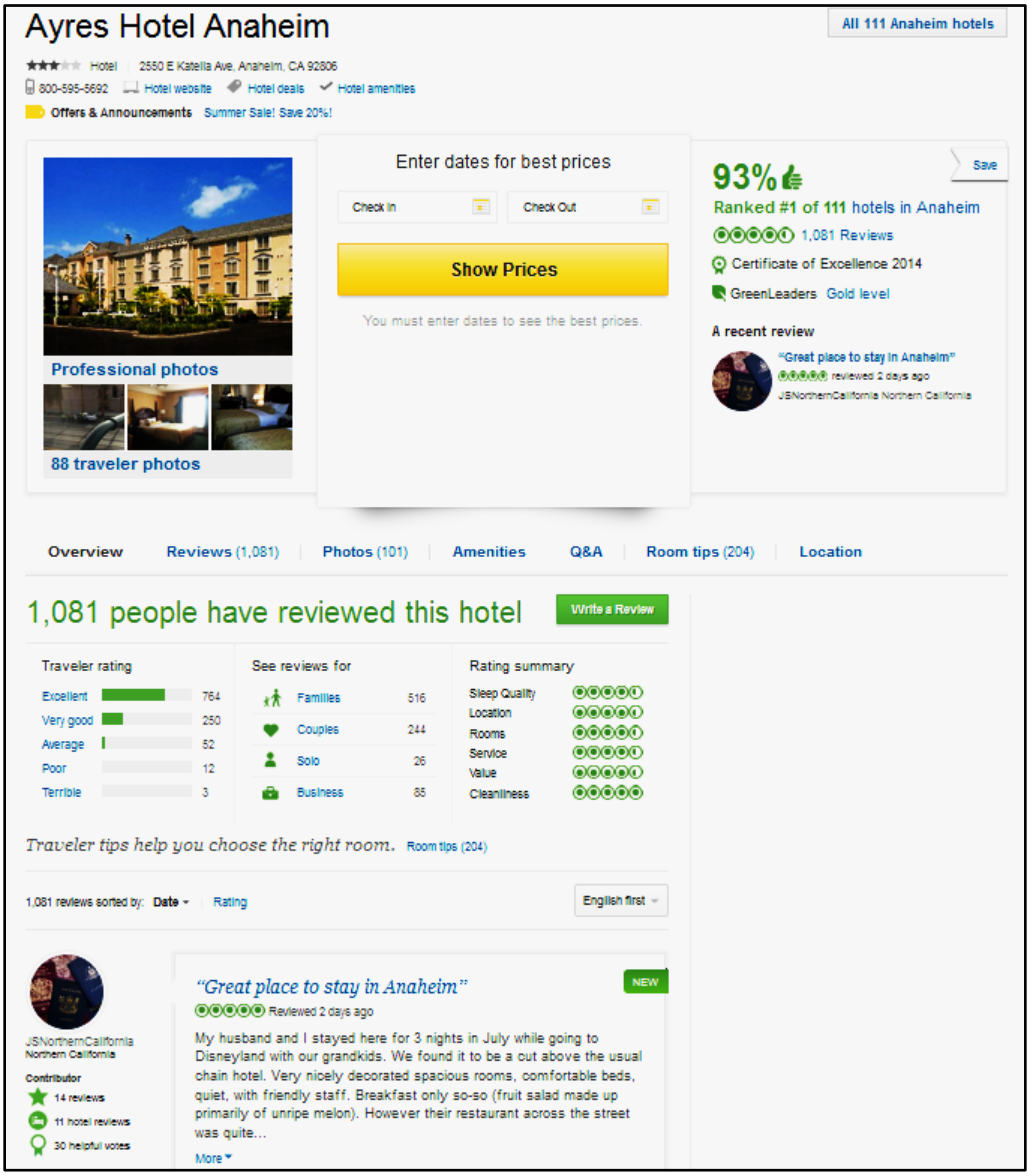
Xiang, Z., Z. Schwartz, J. H. Gerdes Jr and M. Uysal (2015). "What Can Big Data and Text Analytics Tell Us About Hotel Guest Experience and Satisfaction?" International Journal of Hospitality Management 44(0): 120-130.

Yacouel, N. and A. Fleischer (2011). "The Role of Cybermediaries in Reputation Building and Price Premiums in the Online Hotel Market." Journal of Travel Research 50: 1-8.

Yin, L., Y. Ge, K. Xiao, X. Wang and X. Quan (2013). "Feature selection for high-dimensional imbalanced data." Neurocomputing 105(0): 3-11.

Zhang, W., T. Yoshida and X. Tang (2011). "A Comparative Study of TF*IDF, LSI and Multi-words for Text Classification." Expert Systems with Applications 38(3): 2758-2765.

Appendix **A**
TripAdvisor.com Web Page



Koofers.com Web Page

Appendix B

TripAdvisor.com Data Set Descriptive Statistics

Cities	**# of Records**	
Las Vegas	89934	
Orlando	49050	
New York City	47928	
Chicago	35009	
San Diego	24496	
Anaheim	22949	
Philadelphia	13237	
Atlanta	12038	
Miami	11740	
Houston	8043	
Star Rating Distribution		
1	14878	4.32%
2	20012	6.65%
3	36379	11.57%
4	94973	30.21%
5	148182	47.13%
Total	314424	
Star Rating	**Mean Word Count**	**St. Dev.**
1	223.7290	185.6379
2	243.8655	197.9679
3	222.8985	172.6013
4	191.9462	164.8688
5	167.4173	151.0221

Koofers.com Data Set Descriptive Statistics

Star Rating Distribution		
1	267	10.19%
2	294	11.23%
3	334	12.75%
4	565	21.57%
5	1159	44.25%
Total	2619	
Star Rating	**Mean Word Count**	**St. Dev.**
1	116.8240	79.6056
2	112.9728	62.4318
3	113.6108	68.6381
4	102.3982	57.2048
5	102.4763	58.6039

Appendix C

TripAdvisor.com

GLOW Word	RSV	Most prevalent usage contexts
great	7840.9	great place to, great location great, great place stay, a great location, overall great stay
staff	2822.2	staff helpful friendly, staff extremely friendly, clean staff friendly, staff friendly courteous, staff friendly room
excellent	2818.2	great location excellent, hotel excellent location, excellent location great, hotel excellent service, hotel staff excellent
loved	2597.7	loved the hotel, kids loved pool, we loved the, loved washer dryer, year daughter loved
friendly	2389.9	staff helpful friendly, staff extremely friendly, friendly helpful room, extremely friendly helpful, clean staff friendly
comfortable	2249.8	rooms clean comfortable, beds extremely comfortable, clean beds comfortable, room spacious comfortable, beds comfortable staff
perfect	2231.2	hotel perfect location, perfect place stay, location hotel perfect, location perfect easy, location perfect block
helpful	2118.3	staff helpful friendly, friendly helpful room, extremely friendly helpful, helpful room clean, clean staff helpful
wonderful	1983.5	wonderful place stay, staff wonderful friendly, a wonderful time, staff absolutely wonderful, hotel wonderful room
location	1851.7	great location great, a great location, and great location, hotel perfect location, room great location
clean	1774.7	room spacious clean, rooms spacious clean, rooms clean comfortable, clean staff friendly, pool area clean
definitely	1610.9	definitely stay hotel, definitely stay recommend, would definitely stay, overall definitely stay, experience definitely stay
fantastic	1450.4	hotel location fantastic, fantastic pool area, rooms fantastic great, fantastic middle strip, hotel room fantastic
amazing	1381.5	room amazing views, floor view amazing, amazing beautiful hotel, amazing stay hotel, bedroom suite amazing
spacious	1370.6	room spacious clean, rooms spacious clean, clean spacious rooms, room spacious comfortable, clean spacious comfortable
highly	1278.2	highly recommend hotel, hotel highly recommend, great highly recommend, hotel highly recommended, highly recommend others
breakfast	1247.2	continental breakfast morning, continental breakfast good, complimentary breakfast morning, breakfast morning good, friendly helpful b
recommend	1137.7	highly recommend hotel, hotel highly recommend, definitely stay recommend, recommend hotel looking, recommend hotel others
enjoyed	1107.6	overall enjoyed stay, hotel enjoyed stay, kids enjoyed pool, enjoyed stay definitely, enjoyed night stay
beautiful	1040.6	beautiful hotel with, hotel beautiful clean, beautiful hotel stayed, floor beautiful view, beautiful pool area
quiet	1035.6	room floor quiet, hotel clean quiet, clean quiet comfortable, clean comfortable quiet, floor quiet room
restaurants	846.5	lots great restaurants, lots shops restaurants, good restaurants walking, good restaurants area, lots restaurants walking
view	833.0	view empire state, great view strip, view times square, fountain view room, great view city
awesome	818.5	pool area awesome, awesome pool area, area awesome definitely, pool awesome kids, show awesome highly
square	799.4	walk times square, blocks times square, distance times square, middle times square, view times square
easy	782.2	easy access strip, times square easy, location easy access, hotel easy walk, location good easy
pool	747.2	pool area clean, great pool area, pool water slide, kids loved pool, kids enjoyed pool
love	686.5	love this place, love hotel location, toby keith love, love hate relationship, to love this
best	658.8	of the best, the best hotel, best place to, is the best, the best hotels
lovely	654.1	room lovely view, lovely pool area, lovely hotel but, lovely view hotel, grounds lovely pool
value	648.5	great value money, great value great, great value and, location great value, value and location
subway	644.1	square subway station, great location subway, blocks nearest subway, subway station blocks, court subway starbucks
city	625.4	new york city, great view city, of the city, city room tips, center city philadelphia
fabulous	614.2	fabulous location excellent, north tower fabulous, fabulous view room, stay fabulous location, rooms fabulous views
shopping	597.4	shopping walking distance, miracle mile shopping, shopping michigan avenue, great location shopping, restaurants shopping areas
walk	593.0	walk times square, minute walk times, walk navy pier, hotel minute walk, minute walk front
walking	579.2	hotel walking distance, great location walking, minutes walking distance, walking distance great, walking distance attractions
plenty	553.3	plenty room move, breakfast good plenty, plenty restaurants walking, size plenty room, plenty dining options
distance	531.1	hotel walking distance, distance times square, minutes walking distance, walking distance great, walking distance attractions
huge	523.5	huge flat screen, comfortable bathroom huge, hotel room huge, room huge comfortable, huge separate shower

TripAdvisor.com

SMOKE Word	RSV	Most prevalent usage contexts
room	1308.3	told room ready, room called front, said room ready, somewhere else room, call room service
told	1139.0	front desk told, told front desk, told room ready, told rooms available, desk told room
not	1122.4	not worth the, do not stay, not up to, would not stay, not the best
desk	706.0	called front desk, went front desk, front desk told, front desk said, front desk asked
called	687.2	called front desk, room called front, called desk told, waited minutes called, room called housekeeping
dirty	675.0	took dirty towels, bathroom floor dirty, room room dirty, room dirty carpet, dirty towel floor
night	644.8	good night sleep, arrived friday night, night called front, night arrived hotel, paid night stayed
worst	629.5	worst hotel stayed, worst experience hotel, worst part stay, worst part hotel, worst thing hotel
said	596.9	front desk said, said room ready, said room nice, said take care, called housekeeping said
front	572.2	called front desk, went front desk, front desk told, front desk said, front desk asked
poor	565.8	poor customer service, poor quality food, poor front desk, poor service hotel, great location poor
finally	505.9	front desk finally, finally front desk, finally gave room, phone calls finally, check time finally
asked	503.0	front desk asked, asked speak manager, asked front desk, asked moved told, asked credit card
manager	447.7	asked speak manager, front desk manager, speak manager told, general manager hotel, desk manager told
rude	436.0	desk staff rude, rude front desk, front desk rude, hotel staff rude, staff rude unhelpful
call	433.1	call front desk, call room service, room call front, phone call room, received phone call
but	426.8	great location but, but not great, nice hotel but, but nothing special, not bad but
terrible	414.3	room terrible service, service pool terrible, time terrible service, noise terrible hotel, service restaurant terrible
horrible	412.3	horrible customer service, customer service horrible, horrible pool area, horrible night sleep, horrible front desk
carpet	388.9	room noticed carpet, room dirty carpet, room carpet stained, room clean carpet, burn holes carpet
check	373.4	front desk check, check front desk, people waiting check, went check room, check told room
door	346.0	disturb sign door, room opened door, front desk door, room bathroom door, bathroom door shut
work	340.0	room keys work, room work room, remote control work, work called front, internet connection work
ok	339.5	just ok stayed, it was ok, ok for the, ok for a, ok place to
phone	334.3	cell phone number, phone front desk, phone call room, received phone call, phone calls finally
went	316.5	went front desk, went take shower, went room floor, went check room, fire alarm went
people	307.1	thin hear people, front desk people, people waiting check, hear people room, people working front
average	306.0	rooms average size, overall hotel average, great location average, average star hotel, average hotel great
checked	304.1	checked credit card, checked front desk, checked told room, checked went room, checked night stay
paid	295.6	paid full price, paid night stayed, paid night room, paid night stay, hotel price paid
sleep	292.8	good night sleep, made impossible sleep, good nights sleep, place sleep night, decent nights sleep
walls	287.5	walls paper thin, walls thin hear, paper thin walls, thin walls hear, thin walls heard
think	282.2	think twice booking, think star hotel, star hotel think, think room service, think twice staying
know	277.9	front desk know, hotel staff know, know customer service, hotel know people, know ahead time
charged	272.8	credit card charged, charged credit card, charged valet parking, charged room service, hotel charged night
filthy	271.1	carpet filthy bathroom, room filthy dust, carpets filthy travertine, carpeting rooms filthy, screen bathtub filthy
loud	266.8	loud music playing, extremely loud room, music pool loud, makes loud noise, loud room tips
hotel	261.3	worst hotel stayed, hotel looks nice, good thing hotel, nice hotel but, at this hotel
money	259.3	worth the money, your money stayed, money stay somewhere, spend earned money, money somewhere else
hours	255.1	room couple hours, hours room ready, waited hours room, arrived hours check, room hours sleep

Koofers.com

GLOW word	RSV	Most prevalent usage contexts
and	301.7	and you will, good teacher and, great teacher and, great professor and, and you can
great	292.9	a great teacher, is a great, a great professor, teachingshows a great, great teacher and
a	201.5	a great teacher, is a great, a great professor, learn a lot, is a good
interesting	201.3	the class interesting, very interesting and, a very interesting, was interesting and, the most interesting
you	179.6	long as you, you will do, and you will, make sure you, if you study
easy	177.0	is an easy, easy to get, is very easy, easy to understand, an easy a
best	154.2	of the best, is the best, the best teacher, the best teachers, the best professor
awesome	139.7	was an awesome, awesome intro philosophy, awesome professor with, its an awesome, is awesome you
is	111.0	is a great, is the best, is a good, is a very, is an easy
loved	98.4	also loved him, i loved him, a i loved, i really loved, hated loved shell
fun	93.4	is fun and, a fun class, class is fun, for a fun, tests not fun
lot	87.1	learn a lot, a lot about, learned a lot, you a lot, a lot from
very	84.3	is a very, is very easy, a very good, is very helpful, and is very
good	83.4	is a good, take good notes, good teacher and, was a good, a very good
willing	76.2	willing to help, is willing to, and is willing, and willing to, very willing to
amazing	75.7	s amazing not, amazing i am, was amazing if, amazing prof gives, amazing professor very
helps	70.8	helps you understand, teacher that helps, helps you retain, studying helps alot, 40 which helps
really	63.3	i really enjoyed, is a really, you really have, a really good, really knows what
but	63.1	but it is, of work but, but you have, class but if, the semester but
funny	63.0	he is funny, to be funny, was entertaining funny, pretty funny i, very very funny
makes	61.6	makes the class, and makes it, makes this class, which makes the, makes sure you
gives	59.6	gives you the, he gives you, gives a lot, he gives a, and he gives
dr	57.2	i took dr, dr wildy because, dr scott was, although dr knaus, dr c sucks
well	55.8	to do well, well on the, well in this, you do well, not do well
wonderful	51.3	grader wonderful lecturer, sweet and wonderful, lyman briggs wonderful, is wonderful i, wonderful older man
entertaining	51.1	was entertaining funny, class entertaining tests, every class entertaining, were entertaining and, entertaining though i
professor	47.6	a great professor, great professor and, the best professor, very good professor, great professor he
love	47.2	a love of, love of teachingshows, anytimeshows a love, you either love, love with sexit
cares	46.6	cares about her, cares about the, she really cares, she cares about, she cares but
fair	45.3	fair grader and, is very fair, very fair with, was very fair, are fair and
classes	44.9	his classes are, for other classes, classes ive ever, in other classes, time classes just
definitely	44.7	i would definitely, definitely one of, it is definitely, and you definitely, definitely knows his
enjoyed	44.3	i really enjoyed, enjoyed this class, enjoyed the class, really enjoyed the, really enjoyed it
pretty	43.3	a pretty good, it was pretty, can be pretty, hes a pretty, they were pretty
overall	40.2	the class overall, overall this class, is overall a, but overall i, overall you will
sure	40.0	make sure you, so make sure, to make sure, sure you understand, make sure that
final	39.1	before the final, midterm and final, the final and, the final was, your final grade
humor	37.9	sense of humor, of humor overall, humor is odd, humor is awesome, of humor she
highly	37.7	i highly recommend, highly recommend him, his grading highly, the material highly, 280 i highly
understanding	37.4	very understanding and, nice and understanding, understanding the material, a better understanding, a good understanding

Koofers.com

SMOKE word	RSV	Most prevalent usage contexts
not	191.3	do not take, i would not, would not recommend, not take this, i do not
worst	129.7	the worst teacher, worst teacher i, is the worst, of the worst, the worst class
this	105.9	not take this, away from this, dont take this, i took this, to take this
teach	102.4	to teach yourself, does not teach, how to teach, he doesnt teach, had to teach
on	87.0	you on the, on the board, on and on, on your own, on the overhead
when	65.3	when you ask, when he is, and when you, when you need, when you have
avoid	64.7	you can avoid, avoid him if, can avoid it, avoid if possible, avoid this instructor
no	63.3	no idea what, had no idea, i had no, you have no, no idea how
have	62.9	unless you have, have nothing to, i have never, we have to, you have no
like	61.9	seems like a, she doesnt like, feel like she, it seems like, did not like
the	60.7	the worst teacher, is the worst, of the worst, the most boring, is the only
to	59.0	to teach yourself, know how to, expects you to, nothing to do, to do with
doesnt	58.0	he doesnt teach, she doesnt like, he doesnt care, he doesnt know, she doesnt care
for	57.5	for you if, is for you, be ready for, this class for, off points for
how	55.3	know how to, not know how, how to teach, learn how to, how to use
anything	54.1	to learn anything, learn anything about, dont learn anything, anything if you, anything in class
i	51.7	i would not, worst teacher i, i do not, i had no, i made a
are	51.4	tests are extremely, unless you are, the questions are, if they are, are really hard
because	50.7	class because i, is because he, because i was, the class because, to class because
his	50.1	to pass his, his tests were, any of his, his test are, not take his
that	48.3	that she has, top of that, that her class, that he does, recommend that you
terrible	46.0	fast has terrible, be funny terrible, two hours terrible, terrible handwriting that, class was terrible
then	45.8	if not then, and then he, then this is, then you get, it and then
or	44.3	is boring or, in class or, or assignmentsnotes lecture, tests or assignmentsnotes, or hard to
even	44.1	and even if, even though he, even if he, even though you, i dont even
know	43.8	know how to, not know how, does not know, dont know what, know what he
nothing	43.2	nothing to do, have nothing to, there is nothing, has nothing to, nothing you can
from	42.7	away from this, take it from, from someone else, from what i, from this class
dont	42.7	dont take this, dont know what, dont take it, dont take him, dont get it
never	40.3	i have never, you never know, never know what, i never went, he has never
unless	39.5	unless you have, unless you are, unless you want, this class unless, class unless you
horrible	39.0	a horrible teacher, horrible on the, are horrible he, is abosoulty horrible,
hard	37.9	very hard to, hard to understand, hard to follow, it hard to, are really hard
my	37.9	of my life, of my class, of my time, my entire life, of my exams
be	37.4	be ready for, to be in, supposed to be, to be teaching, to be the
all	36.3	at all costs, at all i, him at all, all over the, at all and
does	35.8	does not teach, does not know, he does not, she does not, does not care
only	35.6	is the only, the only one, the only way, are the only, there are only
wrong	32.2	get it wrong, wrong his example, wrong shoes in, wrong because all, wrong dont take
idea	31.5	no idea what, had no idea, no idea how, have no idea, idea what he

Chapter 5:

Conclusions

"One must neither tie a ship to a single anchor, nor life to a single hope."
Epictetus

1. Summary

Ensemble learning techniques are considered one of the most important developments in data mining, machine learning, and artificial intelligence, in over a decade (Nisbet, R., et al., 2009). However, the extensive available research streams can be reduced to the basic observation that constructing a very accurate classifier is difficult, time consuming, and expensive; while constructing a classifier with a lower relative accuracy is trivial. Thus, the eventual strategy is to derive a classifier with a relatively high accuracy from a set of "weaker" classifiers.

BellKor's Pragmatic Chaos, the winners of the 2009 Netflix $1,000,000 Prize, reflect this insight, by stating, "predictive accuracy is substantially improved when blending multiple predictors. Our experience is that most efforts should be concentrated in deriving substantially different approaches, rather than refining a single technique. Consequently, our solution is an ensemble of many methods."

Ensemble learning techniques are becoming mainstream business tools. Ensemble models are now being integrated into numerous industry applications such as, weather pattern modeling, econometric modeling, credit scoring, fraud detection, and recommendation systems. This research introduces three business problems, details their specific industry issues and contexts, and provides empirical evidence that ensembles can improve these current business problems, thereby illustrating the business value of integrating ensemble methods into daily operations.

The first business problem addressed in this research was the high level operational issues related to consumer demand forecasting and the subsequent daily capacity management faced by an alpine ski resort located in the state of Utah in the United States of America. This research presented a

framework that has the potential to help resort management better predict daily skier demand. A basic econometric demand model was developed and tested with the three predictive models, Multiple Linear Regression, Classification and Regression Trees, and Artificial Neural Networks. A variety of ensemble learnings configurations, using Bagging, Stacking, Random Subspace and Voting along with the three previously discussed standalone models, were then developed. The data set contained 908 complete observations. The dependent variable, skier days, represented the demand experienced by the ski resort. The 25 independent variables were all numeric. The top four ranked models, that achieved the highest percentage accuracy improvements, were ensemble configurations.

The second business problem discussed in this research was how managers of sponsored search marketing services could better predict which pay-per-click advertising campaigns would be profitable. With higher pay-per-click campaign classification accuracy, managers are capable of maximizing overall campaign portfolio profitability. Classification models were first developed from the Naïve Bayes, Logistic Regression, Decision Trees, and Support Vector Machines algorithms. The overall accuracies of these four base classification models were compared to several ensemble configurations generated from Voting, Boot Strap Aggregation, Stacked Generalization, and MetaCost. The data set contained 8,499 pay-per-click campaign marketing observations. The textual content of a sponsored ad is quite limited, thus 271 numeric features were derived from stylistic, sentiment, and semantic analysis. A categorical dependent variable of *Success* or *Fail* was derived from a purchase rate cut-off level. Several classification performance metrics were calculated and compared, however the total profit of the selected portfolio of campaigns was used as the primary performance criterion. The results of the research indicated that three ensemble learning configurations produced the highest campaign portfolio profits.

The final business problem addressed in this research was how service related industries could achieve service improvements with better online consumer review management. A service improvement framework was introduced that integrated traditional text mining techniques with ensemble learning methods. The processes of defect discovery and remediation along with accolade discovery and aspiration were introduced. Unstructured data sets from two popular online consumer review websites were used to discover SMOKE words related to service defects and GLOW words

associated with service accolades. This research demonstrated how ensemble learning techniques can improve several performance metrics given the challenges of an unstructured data context.

2. Research contributions

Specific research contributions associated with a business domain were covered in Chapter 2 through Chapter 4. The following discussion details two high-level research contributions synthesized from this dissertation.

Real world data sets were utilized for each research project contained in this dissertation. This approach is in contrast to the numerous ensemble research projects that use data sets acquired from well know repositories, such the UCI Machine Learning Repository. These research data sets are valuable and interesting from a benchmarking perspective; however, the data sets are sanitized unlike what frontline business managers would have at their disposal. The data sets for this research were designed in a manner that allows for the systematic assessment of the effectiveness of ensemble learning techniques based on a combination of dependent variable data type and feature set data type. The research landscape, as illustrated by Exhibit 1.4, shows the three dependent variable-feature set combinations investigated for this dissertation. Ensembles were found to be an effective means of increasing the classification accuracy when compared to baseline classifiers in each business application introduced in this dissertation. However, the ensemble models developed from the combination of a continuous data type for the dependent variable with a structured feature set, explored in Chapter 2, generated the greatest percentage improvements in overall accuracy. This observation motivated a new research question: are there modeling options that could counteract the sparse matrix problem associated with the numerical representation of textual data (a problem encountered in Chapters 3 and 4). The next section discusses a unique solution that helps reduce the effects of sparcity.

The most interesting research contributions, inspired by a multidisciplinary approach for this dissertation, were the constructs of GLOW and SMOKE words and the subsequent quantified GLOW and SMOKE scores. Motived by the topic of relevance feedback and how it is measured using the Robertson Selection Value metric, both topics from the information sciences domain, this research uses the RSV metric as a numeric proxy for the constructs of GLOW and SMOKE word

scores. As previously discussed, the RSV measures the relative prevalence of a term from a category of postings, such as High star reviews, versus a contrasting category of postings from Low star reviews. For this research, the GLOW and SMOKE words are the relevant tokens contained within a consumer review that are probabilistically associated with the class label. For each consumer textual review, the numeric score for each GLOW word were added together producing an aggregate GLOW score for the observation. The same process was used to calculate the SMOKE score for the observation. These calculations were computed for all observations in the data set. These aggregated scores were used in this research for two distinct purposes: as new features or independent variables used in the classification and ensemble analysis, and as individual metrics for comparison between service offerings, service entity, local and other attributes. The application of GLOW and SMOKE word scores in a services environment numerically quantifies subjectivity in consumer reviews, which makes the content more actionable for service managers.

3. Research questions

Chapter 1 presented four broad research questions for analysis. This section provides a discussion and concluding remarks.

3.1 What are the advantages and disadvantages of ensemble methods when compared to standard single classification model techniques?

Ensemble learning methods are general strategies for improving the overall accuracy of classification and prediction models. Numerous research streams support and corroborate the improved accuracy that ensembles can realize by learning and then combining a set of individual models. Many technical authors, consultants, and industry associations associated with data mining extol, in an almost unending litany of praise, the universal benefits of ensemble learning. Several industry leading data mining platforms now include many well know ensemble algorithms attesting to the popularity of ensembles. The ability to increase the overall classification or prediction accuracy is the primary advantage provided by ensemble modeling. Ensembles have the additional subtle advantages of being capable of reducing bias/variance errors and having the ability to search multiple hypothesis spaces. It is generally easier to follow an ensemble building strategy as opposed to developing and parameterizing a single classifier. The current computing power available on the desktop makes it as simple to run a Random Forests^{TM} containing 1000 decision trees as configuring

one highly parameterized decision tree. The chapters included in this research illustrate across three different business contexts, the power and success of following an ensemble building strategy.

Given all the advantages of ensemble learning methods, the one fundamental disadvantage, loss of model interpretability, seems to outweigh the advantages. Numerous business classification or prediction processes could benefit from ensemble implementations, but in many cases, government regulations or laws requiring transparency prevent their use.

This research revealed in Chapter 4, illustrated by Exhibit 4.13, that ensemble methods are not a universal solution. Individual classification models can produce equivalent accuracy to ensemble methods, if sufficiently powerful second-order features, such as SMOKE and GLOW words, are included as additional data set features.

3.2 How can researchers accurately estimate ensemble accuracy and compare the accuracy of several ensemble models?

Analogous to the *No Free Lunch Theorem* (Wolpert and Macready, 1999), there is no one evaluation metric that can be used in every classification or prediction problem. This research applied numerous evaluation metrics, such as overall accuracy, precision, recall, kappa, and root mean squared error, to assess the performance of the base classifiers, as well as the ensemble configurations, applied in each of the business contexts discussed in this research. Due to being a prediction problem, the comparative analysis in Chapter 2 relied heavily on root mean squared error. In contrast, the comparative analysis in Chapter 3 utilized the evaluation metric of portfolio profit which is specific to the Internet marketing industry. Chapter 3 also provided a compelling example of how relying on only overall accuracy could be lead to erroneous conclusions. Additionally, Chapter 3 illustrated how false positives and false negatives, as measured by precision and recall respectively, provided a more granular explanation as to why the Naïve Bayes model configurations performed so poorly. Chapter 2 and Chapter 3 argued the case for using a repeated measures experimental design when comparing numerous classification model configurations. This research highlighted an operator contained in RapidMiner that provides the functionality of a random seed generator, so the data set sampling could be replicated across all model configurations. This design provides the statistical basis for the application of a matched pair *t* test to determine the statistical

significance of model performance differences and control for subject-to-subject variability. The resulting *P*-values were subsequently adjusted, to control for experiment-wise error, by applying the conservative Bonferroni correction.

3.3 Are there base classifiers that are more applicable for ensemble learning methods?

Classification models that have high variance error are algorithms that produce different classifiers from slight perturbations in the data set. These generated classifiers all have different overall accuracies. It is generally accepted that decision trees and artificial neural networks are high variance error classifiers (Kantardzic, M., 2011; Provost and Fawcett, 2013). This attribute makes these classifiers more amenable to ensemble learning techniques because they easily supply the needed modeling diversity required for successful ensemble learning. The analysis in Chapter 2 supported this general observation. The artificial neural network configurations produced superior results. However, the accuracy results contained in Chapter 3 and 4 do not support this claim.

The decision tree configuration accuracies detailed in Chapter 3 are rather unremarkable and fall into the middle of the group. Chapter 4 also provided a mixed analysis of the decision tree results. Although the bagging decision tree ensembles ranked first in overall accuracy in the TripAdvisor.com case, the accuracies of the remaining decision tree configuration, in both cases, fell into the middle of the group. Even the accuracy results of both cases from the popular Random Forests™ were unremarkable. The Random Forests™ results are interesting and present possible future research projects.

3.4 What are some of the insights and cautions that researches or business managers should be cognizant of when employing ensemble methods to data sets from actual business problems?

The most notable caution that the analysis of this research consistently reinforces is that, while ensemble learning techniques can produce remarkable accuracy improvements, the *process is not a panacea*. The subtle and somewhat neglected modeling issues of imbalanced data, unequal classification costs, interpretability, and model regularization all have a tendency to seep into the modeling process unnoticed. Furthermore, as Chapter 4 demonstrated, finding or deriving suitably powerful features, such as SMOKE and GLOW words, is as important as model selection, reinforcing that intelligent feature creation remains essential to classification modeling success.

Academia as well as industry should be cognizant of these modeling complexities and place them in the forefront of future research or application.

4. Future research

Model loss of interpretability is a major disadvantage that prevents many business entities and governmental agencies from implementing ensemble learning methods. The concepts of model interpretability, intuitiveness or comprehensibility are vague and subjective and are certainly available for scholarly research, taxonomy development and policy discussions. One promising research stream that combines the advantages of ensembles with interpretability of forward selection regression is the random generalized linear model (RGLM). The RGLM combines the advantages of higher accuracy and transparent feature selection of the Random Forests™ ensemble approach with the interpretability of the forward selection general linear model. The algorithm is currently available in an R package called randomGLM (Song, L., et al., 2013).

Model parameter optimization is a progressive research stream that could both augment and compete against ensemble learning techniques. Traditional optimization metaheuristics such as grid search, greedy search, and evolutionary search have long been "constrained" by computing resources to relatively small sets of parameter combinations. Although almost a cliché, advanced computing power is now becoming available to researchers in the form of 64-Bit operating systems, distributed and cloud computing, and the abundance of inexpensive RAM. Data mining platforms such as RapidMiner and WEKA now provide several operators that perform parameter optimization.

The concepts of GLOW and SMOKE scores can provide several opportunities for future research projects. One idea that may be the most applicable to service management is to explore how to partition an overall GLOW or SMOKE score into a more granular metric reflecting the different aspects or dimensions of service processes, such as the check-in process, local amenities, and room cleanliness. These new GLOW and SMOKE scores could be easily included in a service related data set as second order features for classification analysis. Lastly, the GLOW and SMOKE score concept could be used to explore the service evaluation at higher-level dimensions such as property groups, city, region, and country.

References

Song, L., P. Langfelder and S. Horvath (2013). "Random Generalized Linear Model: a Highly Accurate and Interpretable Ensemble Predictor." BMC Bioinformatics 14(1): 1-22.

Kantardzic, M. (2011). Data Mining: Concepts, Models, Methods, and Algorithms. Hoboken, N.J., John Wiley: IEEE Press.

Nisbet, R., J. F. Elder and G. Miner (2009). Handbook of Statistical Analysis and Data Mining Applications. Amsterdam; Boston, Academic Press/Elsevier.

Provost, F. and T. Fawcett (2013). Data Science for Business. Sebastopol, O'Reilly.

Wolpert, D. H. and W. G. Macready (1997). "No Free Lunch Theorems for Optimization." Evolutionary Computation, IEEE Transactions on 1(1): 67-82.