Optimization of Quarry Operations and Maintenance Schedules

Brennan K. George

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

Master of Science
in
Mining Engineering

Bahareh Nojabaei, Chair
Erik C. Westman
Manuel J. Barros-Daza

May 4, 2023
Blacksburg, Virginia

Keywords: Mining Industry, Big Data, Optimization, Maintenance, Machine Learning

Optimization of Quarry Operations and Maintenance Schedules

Brennan K. George

ABSTRACT

New technologies such as the Internet of Things are providing newer insights into the health, performance, and utilization of mining equipment through the collection of real-time data with sensors. In this study, data is utilized from multiple quarries and a surface coal mine collected through the software CAT Productivity and CAT MineStar Edge to analyze the performance of loaders and haul trucks. This data consists of performance metrics such as truck and loader cycle time, payload per loader bucket, total truck payload, truck plan distance, and loader dipper count. This study uses data analysis and machine learning techniques to analyze the performance of loaders and haul trucks in the mining operations used in the scope of this study. Data analysis of cycle time and payload show promising results such that there is an optimum cycle time for multiple loaders between 30-40 seconds that show a high average production. Furthermore, the distribution of production variables is analyzed across each set of loaders to compare the performance. The Caterpillar 992K machine in the rock quarries data set seemed to be the highest-yielding machine while the two Caterpillar 993K machines performed similarly in the surface coal mine data set. The Neural Network algorithm created a model that predicted the loader from the performance metrics with 90.26% accuracy using the CAT Productivity data set, while the Random Forest algorithm achieved a 79.82% accuracy using the CAT MineStar Edge data set. Furthermore, the use of preventative maintenance is investigated in the process of replacing Ground Engaging Tools on loader buckets to determine if maintenance was effective. Additionally, data analysis

is applied to Ground Engagement Tools maintenance to identify key preventative maintenance schedules to minimize production impact from equipment downtime and unnecessary maintenance. Production efficiency is compared before and after maintenance on Ground Engaging Tools and concluded that there was no material change in the average production of the mine based on that analysis. The insights gained from this study can inform future research and decision-making and improve operational efficiency.

Optimization of Quarry Operations and Maintenance Schedules

Brennan K. George

GENERAL AUDIENCE ABSTRACT

New technologies are helping us better understand the performance of mining equipment. This is done by using special sensors to collect real-time data on information such as how long it takes for trucks and loaders to perform their job, how much weight in the material they can carry, and how far they have to travel. Through the use of data analysis techniques and machine learning models, the data are analyzed to investigate optimum performance metrics. An optimum time of around 30-40 seconds is discovered for the loaders to output their best performance. We also discovered that through a comparison of normal distributions, some machines in similar working conditions perform much better. In the case of this study, it was found that the Caterpillar 992K loader machine outperformed all the other machines. Using machine learning models, we could accurately predict the loader unit from its data with about 80-90% accuracy. Maintenance practices are analyzed on loader bucket parts that assist in digging to prevent unnecessary maintenance or loss of production. Through analysis of maintenance records and production, it was found that there were no big changes after maintenance was performed. This information can help fuel future research as well as show where improvements can be made.

Dedication

This thesis is dedicated to all of my family and friends. To my parents, I cherish everything you have taught me and I thank you for consistently pushing me to be the best I can be. To my brothers, I appreciate you both leading the way as Hokies, guiding me through every step of my journey, and always being by my side. To my friends, we will always have our Fellowship and thank you for always being welcoming and caring for me. And to my best friend, you've always been a source of encouragement and will always be considered my brother.

Acknowledgments

I would like to thank my advisor, Dr. Bahareh Nojabaei for her continuous support and guidance throughout my research. I would like to thank Dr. Erik Westman for introducing and getting me interested in mining engineering back in my freshman year at the Galileo Slush Rush event. I would like to thank my other committee member Dr. Manuel Barros-Daza "Manny" for being an amazing professor and also a great friend. I would like to thank my classmates and close friends for keeping me on track, pushing me to do my best, and making lifelong friendships even though we are spread across the country. I would like to thank Jason Threewitts for taking a chance on me as an intern and bringing me onto the team at Carter Machinery. I look forward to working with you and everyone else on the team. I want to express gratitude to everyone who has been with me during my college career. This includes my family, friends, professors, and peers. You all have pushed me to do my best and made these last 5 years memorable.

Contents

List of Figures ix

List of Tables xii

1 Introduction 1

1.1 Motivation 1

1.2 Outline 4

1.3 Summary of Contributions 4

2 Previous Work Related to Recovery Optimization 6

3 Recovery Optimization Methodology 10

3.1 Data Collection 10

3.1.1 CAT Productivity Dataset 11

3.1.2 CAT MineStar Edge Dataset 12

3.2 Data Preprocessing 15

3.3 Data Analysis Methods 16

3.4 Statistical Analysis Methods 17

3.5 Machine Learning Methods 18

3.5.1 K-Nearest-Neighbor . . . 19
3.5.2 Random Forest . . . 19
3.5.3 Decision Tree Classifier . . . 20
3.5.4 Logistic Regression . . . 20
3.5.5 Neural Network . . . 20

4 Recovery Optimization Results **22**

4.1 Data Analysis . . . 22
4.2 Statistical Analysis . . . 34
4.3 Machine Learning . . . 37

5 GET Maintenance Application **44**

5.1 Methodology . . . 44
5.2 GET Maintenance Results . . . 45

6 Conclusions and Future Work **49**

Bibliography **52**

List of Figures

4.1 Distribution of Scatter plot of Average Payload per Bucket and Cycle Time for 992K Loader 23

4.2 Distribution of Scatter plot of Average Payload per Bucket and Cycle Time for 990K Loader 23

4.3 Distribution of Scatter plot of Average Payload per Bucket and Cycle Time for 990K Loader 24

4.4 Distribution of Scatter plot of Average Payload per Bucket and Cycle Time for 988K Loader 24

4.5 Distribution of Scatter plot of Average Payload per Bucket and Cycle Time for 988K Loader 25

4.6 Distribution of Scatter plot of Average Payload per Bucket and Cycle Time for 988K Loader 25

4.7 Distribution of Scatter plot of Average Payload per Bucket and Cycle Time for 988K XE Loader 26

4.8 Histogram Showing the Frequency of each Cycle Time Occurring in the 992K Loader Data set 26

4.9 Histogram Showing the Frequency of each Cycle Time Occurring in the First 990K Loader Data set 27

4.10 Histogram Showing the Frequency of each Cycle Time Occurring in the Second 990K Loader Data set 27

4.11 Histogram Showing the Frequency of each Cycle Time Occurring in the First 988K Loader Data set 28

4.12 Histogram Showing the Frequency of each Cycle Time Occurring in the Second 988K Loader Data set 28

4.13 Histogram Showing the Frequency of each Cycle Time Occurring in the Third 988K Loader Data set 29

4.14 Histogram Showing the Frequency of each Cycle Time Occurring in the 988K XE Loader Data set 29

4.15 Distribution of Scatter plot of Average Payload per Bucket and Cycle Time for the First 993K Loader 31

4.16 Distribution of Scatter plot of Average Payload per Bucket and Cycle Time for the Second 993K Loader 32

4.17 Histogram Showing the Frequency of each Cycle Time Occurring in the First 993K Loader Data Set 32

4.18 Histogram Showing the Frequency of each Cycle Time Occurring in the Second 993K Loader Data Set 33

4.19 Normal Distribution of the Variable Cycle Time for the Seven Loaders in the Rock Quarries 34

4.20 Normal Distribution of the Variable Fill Factor for the Seven Loaders in the Rock Quarries 35

4.21 Normal Distribution of the Variable Bucket Payload for the Seven Loaders in the Rock Quarries 36

4.22 Normal Distribution of the Variable Truck Total Payload for each Loader in the Surface Coal Mine 37

4.23 Normal Distribution of the Variable Loader Cycle Time for each Loader in the Surface Coal Mine 38

4.24 Normal Distribution of the Variable Truck Cycle Time for each Loader in the Surface Coal Mine 39

4.25 Heat Map Displaying Correlation Between Each Variable in the CAT Productivity Data Set 41

4.26 Heat Map Displaying Correlation Between Each Variable in the CAT MineStar Edge Data Set 42

5.1 Box Plot of Average Production for Each Set of Days Leading up to and After Maintenance on the GET of a 988K Loader 46

5.2 Percent Change Comparing Production Value from Before and After each Day of Maintenance on the GET of a 988K Loader 47

List of Tables

3.1 Sample Cat Productivity Production Data for Loaders from Rock Quarries . 12

3.2 Sample CAT MineStar Edge Production Data for Loaders from Surface Coal Mine 13

3.3 Number of Samples in each Data Set Before and After Preprocessing 16

4.1 R-squared values Obtained from the Linear Regression and Polynomial Regression Machine Learning Methods 40

4.2 Accuracy scores of the five machine learning algorithms used on the CAT Productivity data set for predictions. Testing data was used to get model accuracies. 41

4.3 Accuracy scores of the five machine learning algorithms used on the CAT MineStar Edge data set for predictions. Testing data was used to get model accuracy. 43

5.1 Assumed values for calculation of production value for each set of production days leading up to and after maintenance 45

Chapter 1

Introduction

1.1 Motivation

The development of new technologies has opened a broader view into machines and sensors through the expanded collection of big data. The Internet of Things (IoT) has taken hold in the mining industry, allowing connections between equipment and software to produce real-time data on numerous parts of the operation; hazard analysis, fleet management, condition monitoring, alarm systems, and even process optimization [1, 2, 3]. The IoT is a system of objects, sensors connected to these objects, and computers that allow for signals to be sent as information or received to control devices as well as to monitor them [4, 5]. IoT is a system that has countless uses and is frequently being used in networks such as the medical field and even in home automation [6, 7, 8]. As technology continues to develop with more sensors being integrated into machines and network connectivity gets faster IoT will become more powerful and useful in numerous applications [9]. In the mining industry IoT is a newer system that allows for easier data monitoring, such as performance or machine health, from machines across the operation and for automation that fully requires machines to communicate with the network to perform activities [4, 5, 10]. Machine companies that design machines for mining purposes, such as Caterpillar, are now fitting newer generations of machines with hardware to allow big data collection and monitoring [11, 12]. The new hardware connects sensors across the machine to user interfaces and maintains records of all

metrics. The new sensor-user interface connection then uses the structured data to output analyses of performance metrics, efficiency, and overall machine health [13].

Although these technological improvements exist, they are not always being fully utilized by companies in the mining industry. Mining is a mature industry that is accustomed to how things have been done in the past because it is cheaper than adopting new technologies in an ever-growing industry [1, 14]. Technology can benefit the mining industry by providing predictive analyses so that they can be proactive and not reactive to hazards such as mine disasters or machines breaking down which delay operations. Being proactive can help predict poor working conditions to improve the health of the workers, predict the failure of machines due to parts breaking or fluid leaks causing structural damage, or even prevent deaths such as the predicted landslide at the Bingham Mine in 2013 [15, 16]. The new technology can also help by determining the optimal performance of machines, providing operators with a production goal allowing the operation to achieve its full potential to keep up with demand. Identifying the optimal performance is important because once this is done then companies can acknowledge areas of their operation that are lacking. Once these areas are found they can start to improve them which could be done with more operator training as well as setting goals for operators to hit. In my experience, mining companies have created reward systems, such as gift cards or food, based on goal performance metrics for their operators to help improve productivity and this is helpful over time to increase production.

The purpose of this exploratory study is to investigate the benefits of big data collection and to facilitate the identification of key performance metrics using two Caterpillar software tools, CAT Productivity and CAT MineStar Edge [17, 18]. The two pieces of software are newer to the industry and are slowly being utilized by companies to monitor their production fleet. A thorough analysis of the data output from the machines would be beneficial in

determining the optimal production metrics as well as visualizing maintenance benefits. In this study, the data will be used to examine loader-hauler interactions to identify optimal performance metrics with the ultimate goal to increase yearly production. A loader, also known as a front-wheel loader, is a variable-sized machine with a bucket attachment on the front that is utilized in mining to move solid material from the production area [19]. Hauler units, also known as haul trucks, are used in mining for the transport of material between areas of the mine such as from the mining face to the processing facility [20]. The loader-hauler interaction is the process in mining by which the loading units fill haul trucks with material from the working face [21]. The material from the working face is first freed in the exploitation stage of mining in which explosives are used to break apart the rock. Once this is complete then the loading and hauling stages are conducted with the loader and hauler units to move the material [22, 23]. Both loaders and haulers have sensors on them that with the help of the IoT transfer data to dashboards so that users are able to monitor the speed of production. The production rate is frequently dictated by the speed of the loader-hauler interaction because if these machines are not up to optimal speed then downstream the operation will not have material to put through the production mill [24].

One hypothesis tested was if there was an optimum point at lower cycle times, ideally in the 30 seconds to 45 seconds range, in which the average payload will have a higher yield thus setting a production goal for operators [25]. Another hypothesis was evaluating whether the maintenance activity on GET loader buckets is appropriate as operators currently determine the frequency of replacement. The last hypothesis states that the use of machine learning will yield a predictive model of loader type and production values associated with that loader and regression models can be used to find additional variable relationships not already highlighted.

1.2 Outline

This thesis analyzes machines from two different types of mines, a surface coal mine and rock quarries. Chapter 2 reviews background information that is related to this topic as well as previous studies conducted that provide support for the motivation of this work. Optimization studies without the use of IoT technology are discussed with literature that displays the methods and results of these studies. Studies conducted using technology about predictive maintenance and predictive models to increase productivity and performance are then discussed. Additionally, research using machine learning algorithms is detailed to provide an introduction to the use of big data in these algorithms. Chapter 3 presents the methods chosen to analyze the data. In this research data analysis, statistical analysis, and machine learning methods were utilized to gather results. Chapter 4 provides the results of the experiment as well as a detailed analysis of these results and the implications to the operations as a whole. Chapter 6 presents the conclusions of this research conducted and the possibility for future work.

1.3 Summary of Contributions

The goal of this thesis is to investigate the uses of big data collection in the mining industry and to identify key performance metrics using Caterpillar software tools, CAT Productivity and CAT MineStar Edge. To this effort, the following was achieved:

- Conducted an extensive literature review on the uses of data analysis, predictive maintenance tools, and machine learning applications in the mining industry.
- Utilized data analysis techniques to identify key performance metrics for nine loaders between the two software.

- Performed statistical analysis on the nine loaders in order to identify the distribution of data and compare the variability of data between loaders in similar production conditions.
- Produced linear and polynomial regression machine learning models to investigate the relationship between production metrics. As well as utilizing machine learning prediction models to determine which was best for predicting production metrics given the type of loader.
- Measured production changes before and after maintenance was conducted on the Ground Engaging Tools of a loader to examine if this maintenance was beneficial.

Chapter 2

Previous Work Related to Recovery Optimization

Previous studies explored the optimization of fleet interactions without the use of technology that highly relies on equations that provide a number based on variables gathered from visual guesses as well as manual measurements. Matsimbe (2020) performed data analysis on shovel-truck interactions in a quarry in Malawi, to determine if they could optimize fleet size using different numbers of haul trucks for one shovel [26]. Using a stopwatch for cycle time as well as numerous equations for payload size that rely on the user's judgment of how full the shovel bucket is, they were able to determine increased fleet size with the current size shovel resulting in the queuing time of the haul trucks to go up by 6.40 minutes [26]. Nday et al. performed similar work in a mine in the Democratic Republic of Congo, with hand calculations that considered conditions of haul roads and the equipment as well as operator experience to lower cycle times by about 8% [27]. Finally, Samatemba et. al recently employed equations to discover the utilization rate, production rate, equipment availability, efficiency, and performance rate for haul trucks, loaders, and drill rigs in a Chibuluma South Mine in Zambia [28]. With simple analysis, these researchers were able to determine that all their machines were working with less than a 50% effectiveness rate translating to a large loss of revenue [28]. In the future, technology access will become more cost-effective and readily available for developing countries thus yielding a potential evolution of adoption over

time. In neither of the previous analyses, the machines were not connected with sensors to software that could accurately measure the payload, cycle times of the loader and hauler units, and queuing time of the trucks. If the technology was available, without having to use equations that take user opinion, therefore leading to potential user error, they would be able to provide a more accurate representation of how their fleet interacts with different size shovels and different fleet sizes. This accurate representation could lead to higher machine efficiencies, leading to higher production, and allowing the company to increase its revenue over time.

The inclusion of sensors and data in machines allows predictive maintenance strategies to be used more frequently. In the past, without access to technology to give performance metrics and machine health, maintenance would occur when machines broke down or when operators used visual or auditory cues to determine that damage had already occurred [29]. This lack of technology allows for the incorporation of human error. The operator may not know the signs to look for that there is a problem with the machine which could lead to more extensive damage creating a longer downtime for maintenance of the machine as well as a costlier repair. The addition of sensors and algorithms to machines can help circumvent this reactive maintenance planning. Equipment is more reliable, meaning less downtime when parts break, and there is a noticeable cost reduction since less is being repaired [30]. Algorithms such as those in machine learning can be utilized to create predictive models from past data to forecast and schedule maintenance for machines. Basri et al. reviewed this and found that through the computer-based approach to predicting machine failures, companies achieved better performance and productivity over time compared to reactive maintenance practices [31]. Due to current supply chain disruptions, it is vital for operations to predict when machines need parts instead of waiting until failure and potentially losing a machine for multiple months while waiting on replacement parts.

Machine learning has been a popular topic in numerous research fields and recent developments in technology allow for easy access and adoption. Supervised learning is a machine learning method in which the model is given a portion of the full data that is completely detailed and is used as a training data set [32, 33]. This training data is used to teach the model so that when the rest of the data, also known as the test data, is input into it, without any labels, it will be able to learn from the training data and predict the classification of the test data. For example, a group of researchers, Khan et. al., used supervised learning to classify unstructured text documents from the internet into reports, emails, views, and news in order to extract their highlights [34]. Unsupervised learning on the other hand is machine learning models that intake data that has no structure or labels and analyze them to produce clusters based on the model's own prediction [35, 36]. For example, Lopez et. al., utilized an unsupervised machine learning method in which they input human genome data without any labels and the machine learning method was capable of clustering multiple sclerosis patients accurately without having any input parameters to learn from [37]. Researchers in the mining industry have focused on this trend and are producing ever-growing experiments to see the capabilities of these algorithms alongside the collection of big data. Nobahar et. al. used five algorithms: linear regression, decision tree, K-Nearest Neighbor, random forest, and gradient boosting to simulate operations to optimize fleet selection [38]. The research found the Gradient Boosting Regressor algorithm was accurately able to predict the best fleet selection given performance metrics, weather conditions, and haul road routes with an 85% accuracy rate. Baek et. al. utilized a deep neural network algorithm to predict the ore production of a mine in the Republic of Korea based on performance metrics of the fleet of haul trucks [39]. The results were promising as their mean absolute percentage error for morning productions was 11.40% and for afternoon productions was 8.87% [39]. A mean absolute percentage error of less than 10 is excellent while between 10 to 25 is low but acceptable [40]. Machine learning algorithms and results should continue to improve over

time, and with the addition of big data collection in the mining industry, will expectantly be used for predicting performance metrics and required maintenance.

Chapter 3

Recovery Optimization Methodology

3.1 Data Collection

This study used data from two different software packages CAT Productivity and CAT MineStar Edge. These software are proprietary to Caterpillar and require individual companies to pay a subscription to have their machines connected to the network to access the data collected by sensors on the machine. These machines are connected to the network using a 4G signal transferred through what is called the Product Link box. This Product Link box allows for the health, utilization, production insights, and hours/location to be transferred to the software dashboards for use by the customer or owner of the software [41, 42, 43]. In this study, all machines used in this study have a Product Link box of generation PLE641 which allows for the advanced production metrics to be gathered. CAT Productivity had multiple loaders, aside from the ones used in this study, that had PLE641 boxes but due to the fact that the customers had not subscribed to the software, these machines only presented utilization data as well as cycle time metrics. These two software are capable of presenting similar performance metrics, machine health, and utilization but there are a few key differences. CAT MineStar Edge allows for real-time data collection and playback recordings of work machines are doing and easily ties which hauler unit is being loaded by the front-end loader [44]. On the other hand, CAT Productivity does not provide real-time data collection and does not provide as in-depth insights as CAT MineStar Edge since it is a

cheaper subscription [45]. CAT Productivity did not have full data filled out for haul trucks at each of these sites. During the exploration of the data available, many of the haul trucks only displayed cycle times and utilization and did not correspond load events to specific loader units. Therefore, due to this it was not feasible to perform a study on machines with one or two variables available in their data set. This data is not publicly accessible and was provided by the company Carter Machinery Company Inc. for use in this study. All data collected in this exploratory study was scrubbed of any customer-identifying information as well as asset number to prevent insight into customer production information.

3.1.1 CAT Productivity Dataset

The first data collected from CAT Productivity consisted of variables for basic production information with a data set for each of the seven loaders. The first data group used eight months of data from April to December 2022. The loaders used in this study were in different quarries across Virginia with similar production conditions to each other. These rock quarries, across Virginia, are surface operations that produce limestone, sand, and gravel. The first group of data was collected using the software Cat Productivity. The Caterpillar loader machines were of slightly different sizes based on their generation. The seven Caterpillar loaders consisted of one 992K machine, two 990K machines, three 988K machines, and one 988K XE machine. These machines are variable in size but frequently are used for similar-size quarries based on company choice. The loader data was tied to each haul truck thus enabling the gathering of truck metrics as well. The data collected contain variables for Date and Time, Bucket Payload (tons), Truck Total Buckets, Truck Total Payload (tons), Cycle Time (seconds), and Truck Id. Bucket payload is just the tonnage of the material that is in the loader bucket before it is dumped into the haul truck. Truck Total Buckets is the total number of loader buckets full of material that it takes to fill up the truck

Date	Bucket Payload (tons)	Truck Total Buckets	Truck Total Payload (tons)	Cycle Time (Seconds)	Truck ID
06/18/2022	16.94	4	68.93	45	1
07/02/2022	14.56	5	87.64	73	1

Table 3.1: Sample Cat Productivity Production Data for Loaders from Rock Quarries

before it departs the loading area. Truck Total Payload is the total tonnage of the material that the truck departs the loading area with after being filled by the loader. The cycle time is in regards to the loader and it is measured in the total time of the following four phases: picking up material from the working face to fill the loader bucket, swinging toward the haul truck, dumping the material in the haul truck bed, and finally swinging back toward the working face [46]. Truck Id is simply a way for the companies to distinguish which truck is being loaded by the loader. Table 3.1 shows a sample of the data collected.

3.1.2 CAT MineStar Edge Dataset

The second data set consisting of the two loaders from Cat MineStar Edge had numerous variables but only a few can be utilized for performance metrics analysis. This is because the majority of the information in this data set is made up of sensitive information to the company such as load and dump location in X, Y, and Z as well as latitude and longitude coordinates, haul routes with distinguishing information, machine serial numbers. Alongside the sensitive information, there was miscellaneous that was omitted due to it not being related to any performance metrics. For example, the variable cycle type was omitted because every data point had the same value which was ”HAUL” as well as other variables like haul operator which was not filled out for any data point. The second group consisted of two loader machines of the generation Caterpillar 993K. This second data set also consisted of eight months of data from a large-scale surface coal mine located in West Virginia. These

Date	Load Duration (seconds)	Truck Cycle Time (minutes)	Plan Distance (m)	Loader Dipper Count	Payload (tons)
04/16/2022	253	16.94	2150	6	157.3
05/12/22	341	14.56	1896	7	143.2

Table 3.2: Sample CAT MineStar Edge Production Data for Loaders from Surface Coal Mine

Caterpillar 993K are large pieces of machinery that are capable of large production operations such as the surface coal mine this study was conducted on. The most useful variables in the data set were Date, Load Duration, Truck Cycle Time (seconds), Plan Distance Full (m), Loader Dipper Count, and Resolved Payload (tons). Load duration is the length of time it took for these loaders to fill the haul trucks with the material. Truck Cycle Time is the total time that the haul truck goes through the phases of the start of loading, hauling to the dump site, dumping, hauling back to the loading site, and then pulling up next to the loader again [47]. Plan distance full is the total horizontal distance that the haul truck travels in order to reach the dump site from the loading site. Loader Dipper Count is the number of loader buckets it takes to fill the bed of the haul truck before it departs for the dump site [48]. Resolved payload is another name for Total Truck Payload. Table 3.2 shows a sample of the data.

There were limitations in the data collection. There was no possibility with the data from CAT Productivity to identify which haul truck was associated with each loader cycle. The Truck ID variable mentioned in Section 3.1.1 was never filled out with the distinguishable truck ID numbers. This means that the companies did not set up the system with a specific identifier for each haul truck. Due to this haul truck performances were not able to be identified. Therefore, this variable was deleted from the data sets. In a search through the CAT Productivity software, there were some haul trucks that were tied to each load cycle but then it was discovered that these loaders did not have upgraded subscriptions. This

lack of subscription means the only production metric shown for these loaders was cycle time. Comparing loader cycle time to haul truck cycle time did not seem reasonable for the scope of this research so these machines were not included in the analysis. Many of the loaders used in this research did not have accurate Truck Total Bucket variables. The way this information is stored in the system is it will keep adding one to the count of the total number of buckets per truck until the operator clicks a button that stores the number of buckets and resets the count. This button is usually tied to the horn on the machine as the loader operators honk the horn in order to indicate to the haul truck operators that loading is complete but this button is often never configured by the companies. Therefore, the data shows the count going up to 50 and then resetting frequently. For the loaders that did have this button configured, there were multiple occurrences where the count would be too high for a reasonable amount of buckets which means there was a human error in which the operator forgot to press the button.

Ideally, the haul truck metrics would be available alongside the loader metrics to compare and see if there are connections between the variables. With the machines included in this study, it was not possible. CAT MineStar Edge did have some limitations as well. One instance is the lack of individual loader bucket payload values. The software only gives the total payload per truck in the data sets which does not allow for the cycle time of the loader to be compared to the average payload per bucket. This lack of data means the analysis for the CAT MineStar Edge data was slightly different than that of the CAT Productivity analysis.

3.2 Data Preprocessing

Each data set downloaded from every loader was normalized for data analysis to be conducted input into the models. This process included identifying the range of cycle times for each loader to remove any large outliers, deleting missing data referred to in Section 3.1.2, and filtering out unnecessary variables that would not indicate production metrics and were sensitive information to the company.

For the first data set from Cat Productivity, numerous data points needed to be deleted. Given the lack of filled-out data, some variables were deleted from the data sets because of inconsistencies in subsequent data collection across each loader. These variables included Bucket Payload Sequence, Truck Total Buckets, Truck Total Payload, and Hauling Unit. All the loaders had the Product Link 641 boxes which allowed the setup of identifying how many bucket cycles it took to load the haul truck, the overall tonnage of the haul truck when full, and the haul truck associated with that cycle.

For the Cat MineStar Edge data set, the variables that were deleted were associated with sensitive information to the company. Each loader had its cycle time analyzed in a distribution to determine the frequency of task achievement creating low and high bounds for a cutoff time. Loader operators may perform multiple tasks while waiting for haul trucks to position themselves next to the loader at the mining face. On the low end, they may be cleaning the face or floor by picking up and dumping material repeatedly in a quick fashion. On the high end, they could be sitting with a bucket full of material waiting for a haul truck to pull up. Additionally, the fill factors for each data point were calculated in the CAT Productivity data sets which were done by taking the bucket payload and dividing it by the max bucket payload for each machine. Table 3.3 displays the number of samples in each data set before and after preprocessing.

	Number of Samples Before Preprocessing	Number of Samples After Preprocessing
993K	5557	3900
993K	7521	4365
992K	36633	32695
990K	72837	61155
990K	38727	31987
988K	37573	31724
988K	25961	20763
988K	66912	65894
988K XE	60652	57944
988K XE	60652	57944

Table 3.3: Number of Samples in each Data Set Before and After Preprocessing

3.3 Data Analysis Methods

Once the data sets for the eight months were gathered from CAT Productivity and CAT MineStar Edge, data cleaning was the first step in this study. Microsoft Excel was used to clean and prepare the data sets to verify the accuracy and reliability of the data. The data cleaning process involved removing sensitive information from the companies and variables that consistently had a lack of information. Following the removal of data, the fill factor for the CAT Productivity set of loaders was calculated which was conducted by dividing each bucket payload by the maximum payload possible for that machine. Due to a lack of individual cycle bucket payloads in the CAT MineStar Edge data set, the fill factor was not able to be calculated for the two loaders in the surface coal mine. Following the data cleaning process, the PivotTable function within Microsoft Excel was used to calculate the average payload for each cycle time. The frequency that each cycle time that appeared in the data set are calculated to determine the most common cycle times for each loader. This analysis provided us with insights into the payload distribution and helped to identify any patterns or trends shown in the data.

Next, the average payload for each cycle time and frequency of the cycle times was imported into the software Google Colab. This software allowed us to use Python code for data visualization purposes. Google Colab was chosen because it allowed us to run Python code on a web browser that synced to Google Drive so there was access to the code from any computer [49]. The figures were created using a combination of the Matplotlib, Pandas, and Numpy libraries. Scatter plots are created to display the distribution of data for average payload vs. cycle time, and histograms to visualize the frequency of cycle time for each of the loaders. These figures allowed us to gain insight into the relationships between the variables and to identify any outliers in the data. Finally, each of the figures was saved from the code to the hard drive of the host computer.

3.4 Statistical Analysis Methods

Google Colab was used to perform statistical analysis, calculations, and data visualization. The first step in this process was the calculate the measures of central tendency for the major production metrics chosen for each set of loaders. These measures of central tendency included the mean, standard deviation, and first and third quartile. For the CAT Productivity loaders, the seven loader units in this set are compared based on the variables loader cycle time, fill factor, and bucket payload. For the CAT MineStar Edge loaders, the two loader units in this data set are compared based on the variables total truck payload, loader cycle time, and truck cycle time. By calculating these measures of central tendency, a visualization of the distribution of the data as well as any trends or patterns that may appear were obtained. These measures of central tendency calculations provided insights into the performance of each loader compared to each other as well as to show any areas for improvement.

In order to visualize the distribution of data, the probability density function (PDF) is calculated using the measures of central tendency. The PDF is a way to display the normal distribution of a variable in a given data set and provides a comparison of different data sets when analyzing them [50]. The PDF was then calculated for each of the variables chosen for the sets of loaders. Through visualization of the PDF, any significant differences between the performance of the loaders in each set is identified to gain a better understanding of the underlying data. Limitations in the data occur with the lack of variables to compare between the loaders. So a necessary judgment are made as to which variables to focus on for the comparison of performance metrics.

3.5 Machine Learning Methods

The machine learning setup and analysis were also conducted using Google Colab because of the access to the machine learning libraries within Python. For the machine learning analysis, all of the individual loader data sets were put together but were distinguished from each other with a loader ID of 1 through 7 for the CAT Productivity data set and loader ID of 1 and 2 for the CAT MineStar Edge data set. The result was two separate files, one with all the data from CAT Productivity and one with all the data from CAT MineStar Edge. In the first step, a linear regression model, as well as a polynomial regression model, was built to explore the relationship between the target variable, the loader units, and the other features. From these models, R-squared values were calculated for the evaluation of the performance of the models. The R-squared value tells how much of the variance of the target variable is explained by the predictor variables which in this case is all the production metrics aside from the loader units [51]. After the linear and polynomial regression models were created for both data sets, an additional visual was created to identify the correlation

between each of the variables in the data set. We used this to investigate to see if there were hidden correlations between variables. These were shown by using heat maps which display on a grid the correlation between each of the variables.

In the next step, various classification models were built to predict the target variables. A random 20% of the data was chosen to be used as testing sets with the loader identifier being used as a parameter. The other 80% was used as the training data sets. Next, the two data sets were run through five different predictive models, K-Nearest-Neighbor, Random Forest, Decision Tree, Logistic Regression, and Neural Network. Clear results were obtained through this. The choice of each of the predictive models is discussed in Sections sections 3.5.1 to 3.5.5. These models produced accuracy scores capable of evaluating the performance of the models. The best accuracy score determined the best-performing model.

3.5.1 K-Nearest-Neighbor

The K-Nearest-Neighbor algorithm was chosen as one of the predictive models because it is versatile in being able to classify data but also to be able to predict based on the information input into its model [52]. This model works by taking given unlabeled data, in this case, the data not identified by its loader ID, and it attempts to accurately classify it by similar characteristics to the labeled data [53]. This was a great model for this research because it could be used either to predict what type of loader it is or use the type of loader to predict the production metrics.

3.5.2 Random Forest

The Random Forest method was chosen because of the use of randomization of the features it selects to split by the nodes [54, 55]. This algorithm is capable of splitting up into multiple

different sample groups, in what is called bootstrapping, and then it runs the algorithm on each group individually and finally brings them together at the end for classification [56]. This randomization hopefully helps accurately predict the classifications in the data sets.

3.5.3 Decision Tree Classifier

Decision Tree Classifiers splits information up on a map in which there are branches that split based on the difference in value in the data so if you followed the tree all the way down then you would ideally find the classification that is correct for that data point [57, 58]. These classifiers are typically great for discrete data which means quantitative data that is capable of only taking specific values which is the case for cycle time in this study since for the rock quarries it can only be between 0 seconds to 100 seconds [59].

3.5.4 Logistic Regression

The logistic regression model was chosen due to the poor performance of the linear and polynomial regression models that were initially used in the machine learning analysis. The logistic model was also chosen because it allows for expert analysis in determining the relationship between multiple independent variables and a dependent variable [60]. This was important in our consideration it this model works especially well for prediction problems or analysis.

3.5.5 Neural Network

The neural network model was chosen because of its complex nature of the model. The model created in this study was a simplistic version but it was capable of producing the

desired results using the neural network method. Neural networks are ultimately great at taking training examples, such as a subset of data, and analyzing it to gather information and then ultimately perform a given task based on what it has learned [61]. Neural networks are interesting as they are modeled after the human brain with nodes, which are decision points, that communicate to the other nodes through connections and transfer what they learn through this [62].

Chapter 4

Recovery Optimization Results

4.1 Data Analysis

Through data analysis of each loader's data set, a distribution of the average payload for each cycle time was obtained. Each loader was analyzed individually to determine if the results produced an optimum low cycle time in which the average payload per bucket is high. This optimum point would ultimately yield higher production down the line so this would be a goal for operators to strive for. The frequency of each cycle time was also included to determine the spread of the data. This spread could be tied to the average production to see how often the operators are achieving the optimum cycle time to give an indication of the speed of production. The loaders were analyzed between cycle times of 0 seconds to 100 seconds. Low cycle times would typically indicate the loader operators are performing non-production activities such as cleaning up the floor around the mine face for loose rock. The CAT Productivity software may distinguish this as a load event because it registers the payload and action of the loader picking up and dumping rock. Any cycle time above 100 seconds was not counted because these were infrequent and indicated the loader waiting for a haul truck to pull into the loading area. Figures 4.1 to 4.7 show the scatter plots of the average bucket payload vs. cycle time for all seven of these loaders from the rock quarries.

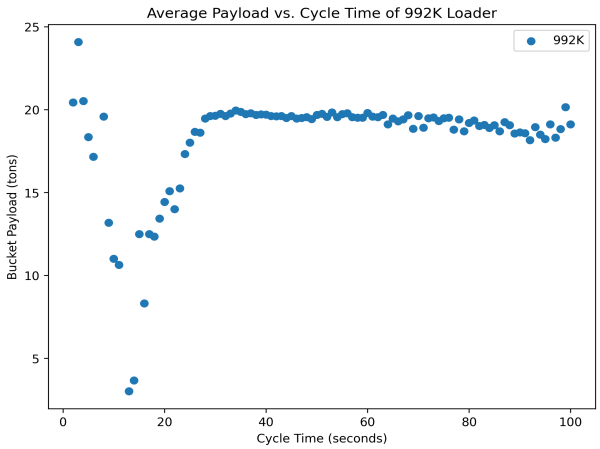

Figure 4.1: Distribution of Scatter plot of Average Payload per Bucket and Cycle Time for 992K Loader

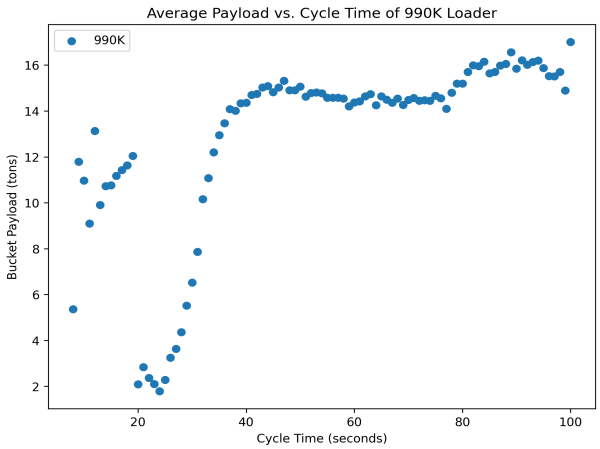

Figure 4.2: Distribution of Scatter plot of Average Payload per Bucket and Cycle Time for 990K Loader

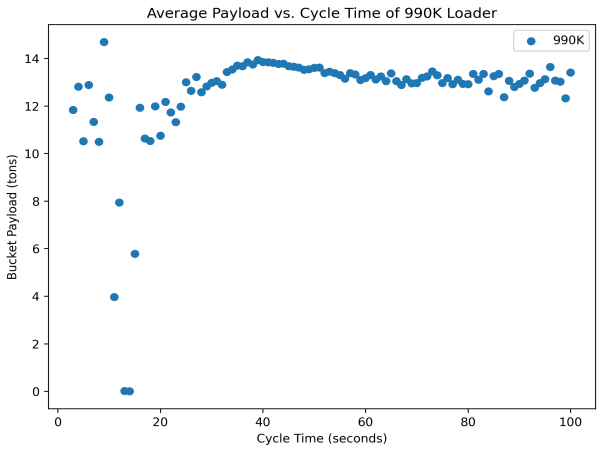

Figure 4.3: Distribution of Scatter plot of Average Payload per Bucket and Cycle Time for 990K Loader

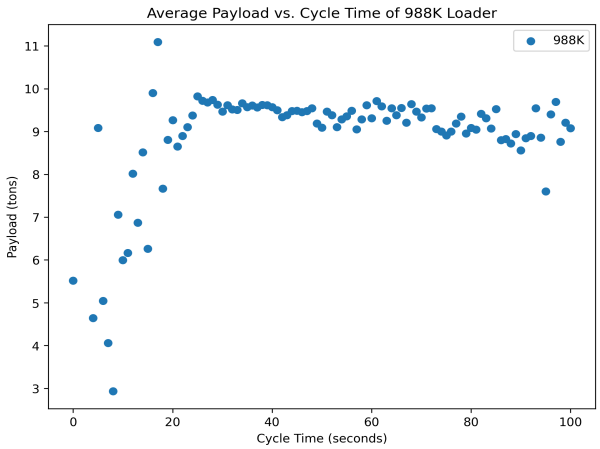

Figure 4.4: Distribution of Scatter plot of Average Payload per Bucket and Cycle Time for 988K Loader

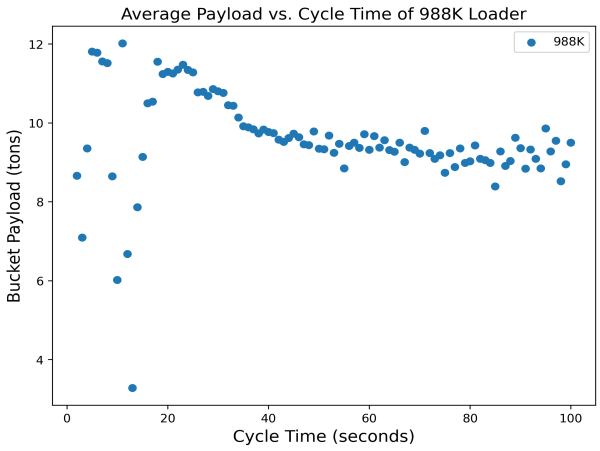

Figure 4.5: Distribution of Scatter plot of Average Payload per Bucket and Cycle Time for 988K Loader

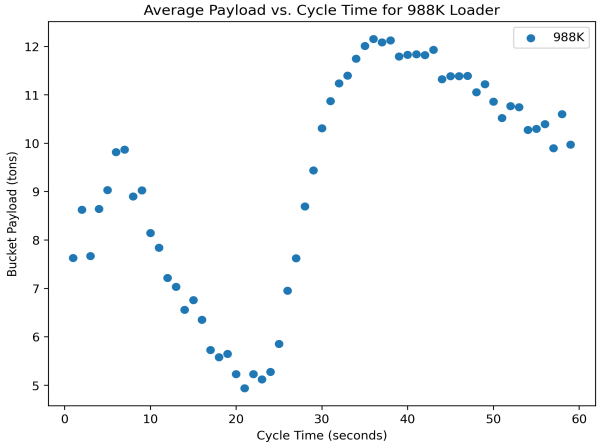

Figure 4.6: Distribution of Scatter plot of Average Payload per Bucket and Cycle Time for 988K Loader

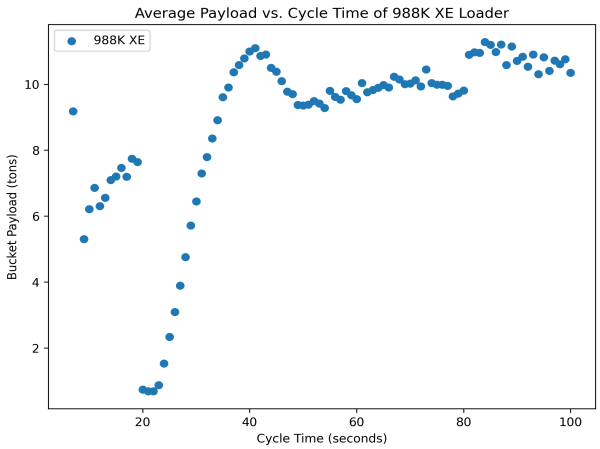

Figure 4.7: Distribution of Scatter plot of Average Payload per Bucket and Cycle Time for 988K XE Loader

Additionally, Figures 4.8 to 4.14 show histograms showing the distribution of the frequency of each cycle time.

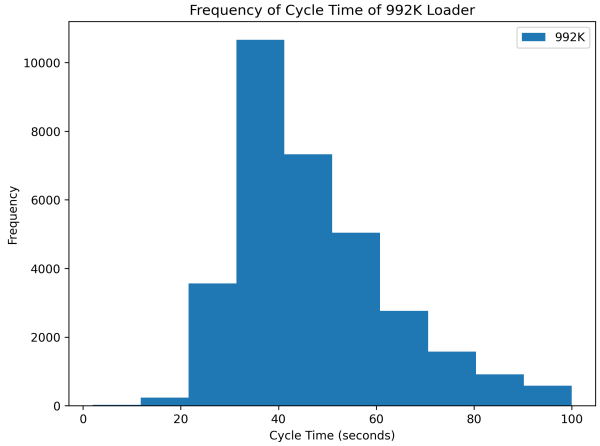

Figure 4.8: Histogram Showing the Frequency of each Cycle Time Occurring in the 992K Loader Data set

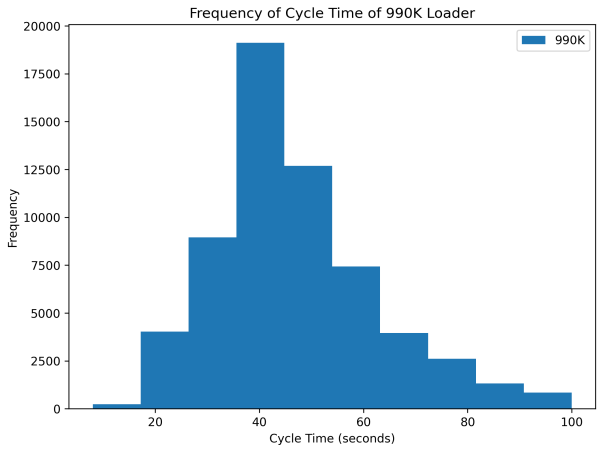

Figure 4.9: Histogram Showing the Frequency of each Cycle Time Occurring in the First 990K Loader Data set

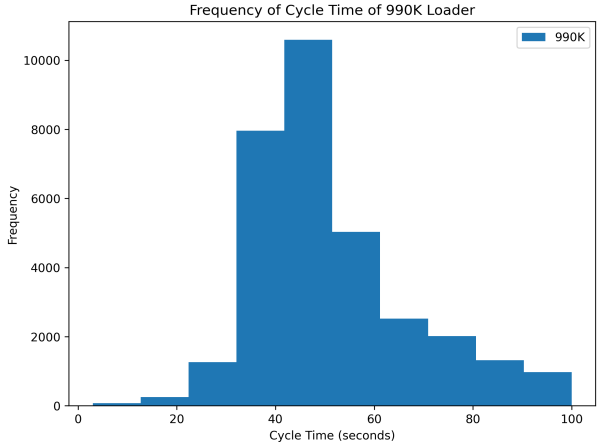

Figure 4.10: Histogram Showing the Frequency of each Cycle Time Occurring in the Second 990K Loader Data set

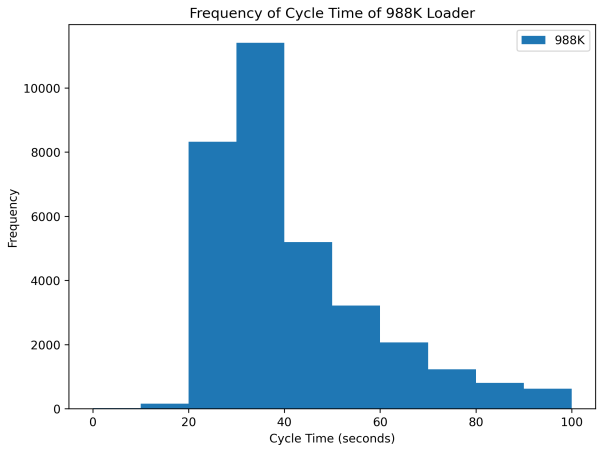

Figure 4.11: Histogram Showing the Frequency of each Cycle Time Occurring in the First 988K Loader Data set

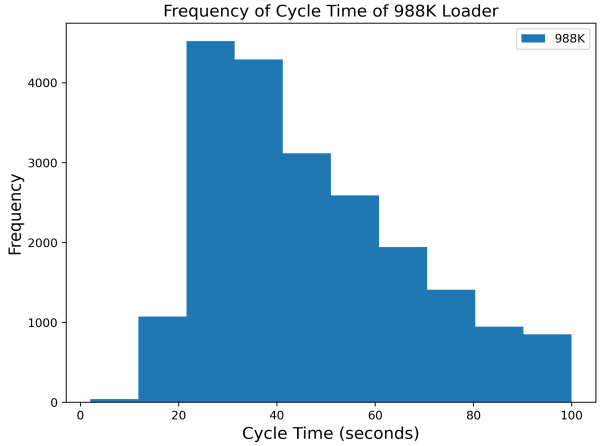

Figure 4.12: Histogram Showing the Frequency of each Cycle Time Occurring in the Second 988K Loader Data set

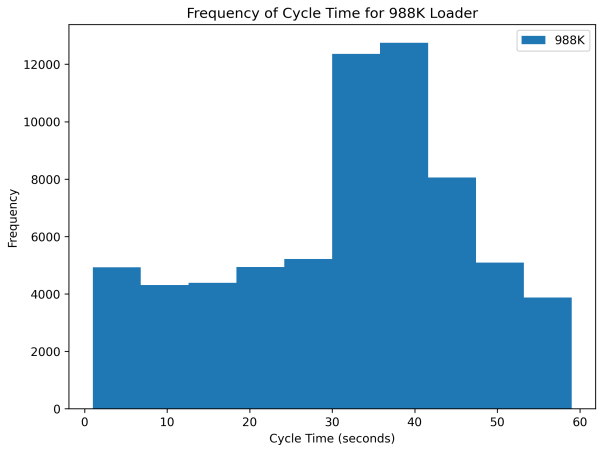

Figure 4.13: Histogram Showing the Frequency of each Cycle Time Occurring in the Third 988K Loader Data set

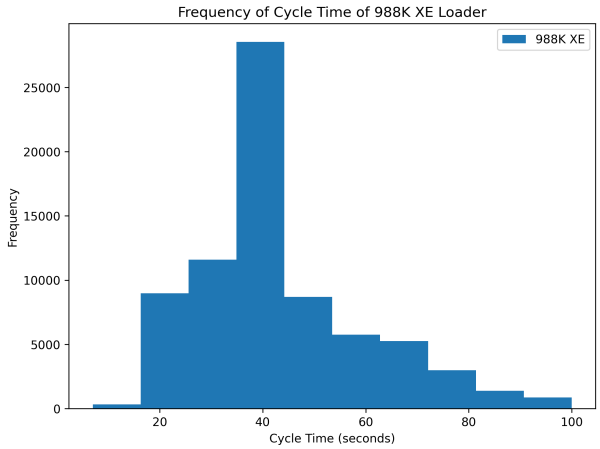

Figure 4.14: Histogram Showing the Frequency of each Cycle Time Occurring in the 988K XE Loader Data set

Of the data that was collected and displayed for the seven loaders in the rock quarries, three of them show promising results and the other four are inconclusive concerning cycle times and yields. The scatter plots of Figures 4.5 to 4.7 show the promising results. These figures indicate an optimum point around a cycle time of 32 seconds for Figure 4.5, 37 seconds for Figure 4.6, and 40 seconds for Figure 4.7 in which payload results in a higher yield. These results are promising because they fall in the range of an ideal cycle time which should be

between 30 seconds to 45 seconds. The lower range of this ideal cycle time, 30 seconds, is reasonable with an adequate operator but when the cycle times start pushing reaching 45 seconds then that is an area for improvement [25]. With machines of the size used in this study, it has been found that this range of cycle times is adequate for the operator to fill up the bucket a reasonable amount as well as to dump it into the haul truck. The scatter plots of Figures 4.1 to 4.4 show the inconclusive results. These plots are inconclusive because although there appears to be an optimum point for a cycle time that averages a high payload yield, it appears to level out after this point and only has slight changes to the average payload at each cycle time. The optimum point for these figures also appears to be around 35 to 40 seconds. The trends of a few scatter plots illustrate an increase in payload yield as they approach the higher bound of the cycle time range, 100 seconds. The increase in the payload is theorized that in a longer cycle time, operators repeatedly shovel into the blasted rock piles to gain higher fill factors for their buckets. The higher payloads in these instances are not significant enough to be useful and they sacrifice time that can be used to increase production over the shift or could be used for more cycles. Ideally, it would be better to stay in lower cycle times for production goal purposes.

The histograms of the frequency of each cycle time for each loader mainly show a normal distribution around the cycle times that represent the optimum point for a high yield in the payload. A few of the histograms, namely in Figures 4.8 to 4.12 and 4.14, are skewed toward higher cycle times which could suggest the operators are multitasking rather than simply loading haul trucks. More data near the higher bound of the cycle time could suggest operators are taking unnecessary lengths of time to load a truck or clean the floor of the pit which could affect downstream production. The normal distribution in these graphs centered around ideal cycle low cycle times suggests that the operators of these loader units are performing exceptionally. The higher cycle time skew means that there are numerous

more frequent high payload data points at these cycle times which could inflate the data. Figure 4.13 show no normal distribution and instead show a nearly uniform distribution and a skew toward lower cycle times respectively. The lower cycle time skew goes against reasonable cycle times as they would be too fast at these points and would not reasonably be able to maintain a high payload yield and dump into the haul truck in the same minimum length of time. The nearly uniform distribution seems inconsistent with the other loaders used in this study as it has fewer data points which could suggest it was used for a variety of activities and not just loading haul trucks with the material. Figure 4.13 suggest either abnormalities or mistakes in the data and would need further study with a site visit to analyze operations.

The two loader units from the surface coal mine were analyzed similarly to the seven loaders from the rock quarries. These loader units, however, did not have individual cycle times for each bucket payload in their data set. As a result, Figures 4.15 and 4.16 show the scatter plots of the average truck payload vs. cycle time for both of the 993K loaders from the surface coal mine.

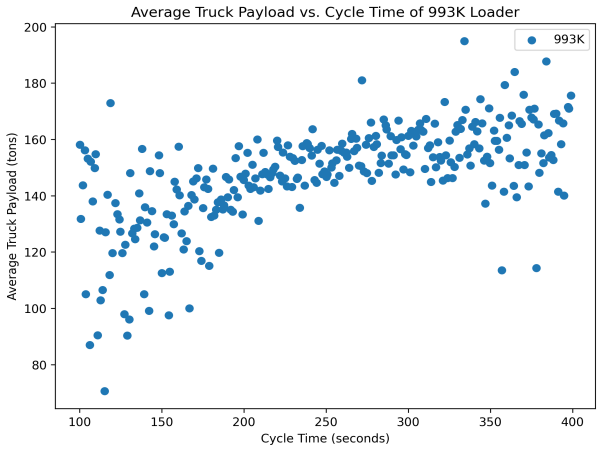

Figure 4.15: Distribution of Scatter plot of Average Payload per Bucket and Cycle Time for the First 993K Loader

Additionally, Figures 4.17 and 4.18 show histograms showing the distribution of the frequency

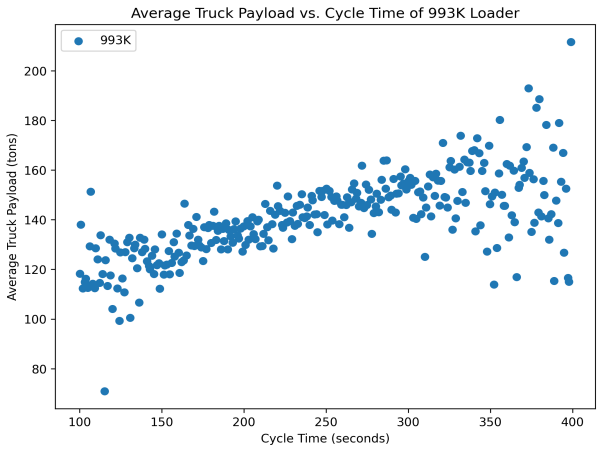

Figure 4.16: Distribution of Scatter plot of Average Payload per Bucket and Cycle Time for the Second 993K Loader

of each cycle time.

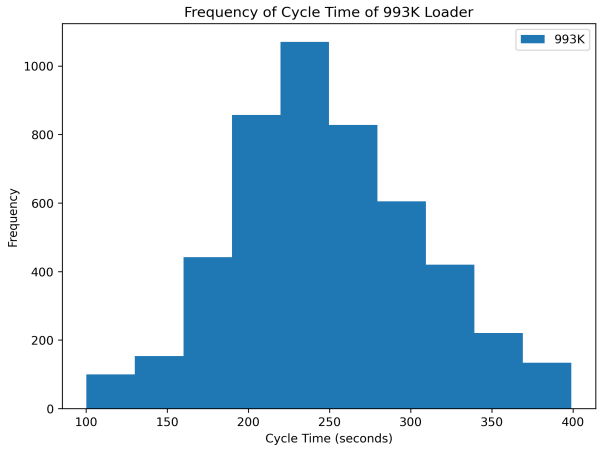

Figure 4.17: Histogram Showing the Frequency of each Cycle Time Occurring in the First 993K Loader Data Set

In Figures 4.15 and 4.16 the two 993K loaders do not concretely contain an optimum point around a cycle time that shows high payload yields. Although the scatter plots do not show an optimum point, they do have a positive trend with some outliers. The positive trend confirms the hypothesis that as you increase cycle time you may have higher payload yields due to repeatedly shoveling into the blasted rock piles to achieve a higher fill factor.

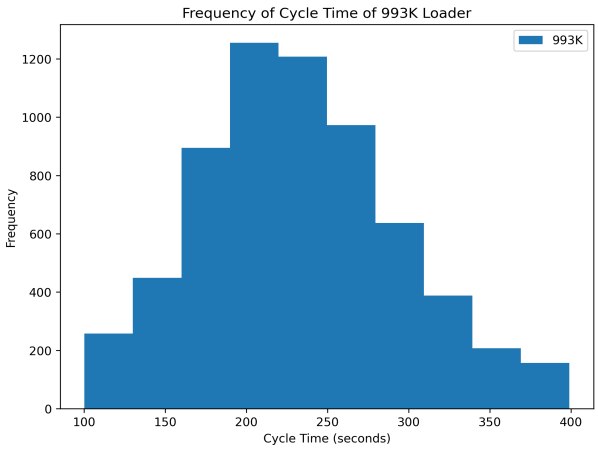

Figure 4.18: Histogram Showing the Frequency of each Cycle Time Occurring in the Second 993K Loader Data Set

Although this is true, the longer loading strategy must be tempered because there could then be an unintended effect where loading one haul truck for larger amounts of time could back up the entire operation. The low production for the shift could be due to longer cycle times or a backup due to long queue times at the loading site.

Figures 4.17 and 4.18 displaying the histograms of the two 993K loaders at the surface coal mines primarily show a normal distribution centered around the cycle time of 250 seconds. This centered point could suggest that for these larger operations such as these surface coal mines ideally operate around these times since they are more frequently hitting them. The normal distribution with a wide range also suggests that these loading units are not just used for loading haul trucks. They could be used for floor cleanup as well as moving material from stockpiles into crushers to continue production down the line.

4.2 Statistical Analysis

Statistical analysis was conducted on certain variables in the data to visualize their distribution for comparison. Cycle time, fill factor, payload, truck payload, and truck cycle time, were determined to be the best for determining the production performance of the loader units in their respective operation. The data was again split into two sets in this analysis based on quarry type. One data set included the rock quarries consisting of the seven loader units discussed in the data collection section. The other data set included the two loader units in the surface coal mines also discussed in the data collection section. The variables for the first set that were analyzed were the cycle time, fill factor, and payload. The variables presented in Figures 4.19 to 4.21 were chosen because all their data points were filled out across the seven loaders in the data sets.

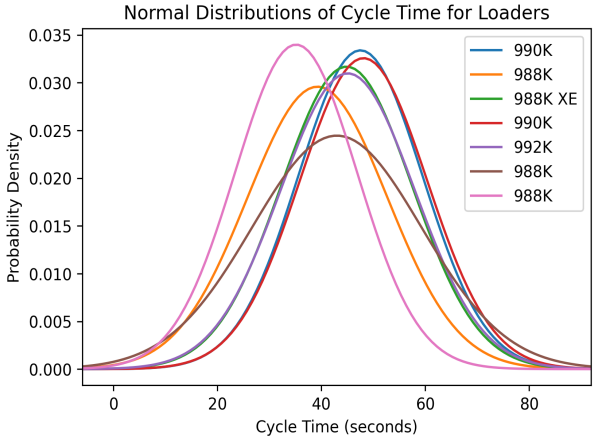

Figure 4.19: Normal Distribution of the Variable Cycle Time for the Seven Loaders in the Rock Quarries

In fig. 4.19 the normal distributions of the cycle times for the seven loaders were generally similar to each other. The distributions were centered around different values but multiple had similar size spreads. Multiple of the distributions was centered around a cycle time of approximately 50 seconds suggesting that this is in reality the average cycle time for rock

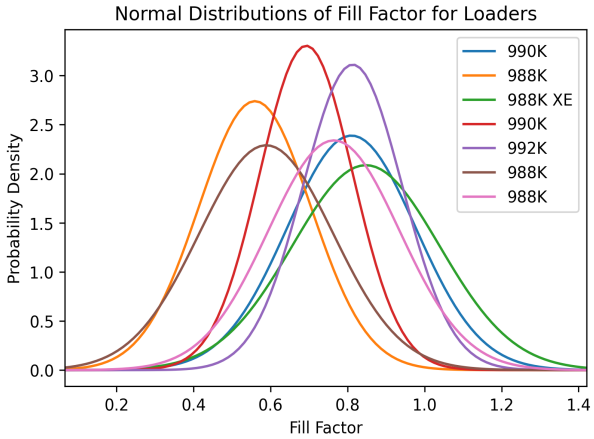

Figure 4.20: Normal Distribution of the Variable Fill Factor for the Seven Loaders in the Rock Quarries

quarries of this size. In Figure 4.20 the normal distributions of the fill factors for each of the loaders were considerably different. Multiple of them had low fill factors indicating the operators are not utilizing the size of their buckets to their fullest extent. One reason for the under utilization of buckets could be a lack of operator experience as well as loading units frequently being used for other purposes such as cleaning the face while registering these tasks as a loading cycle. Four loaders, 992K, one 990K, one 988K, and the 998K XE, showed optimal fill factors of about 80% or higher. The high fill factors are ideal because the operators are utilizing the machines to almost their full extent to fill the trucks resulting in fewer cycles per truck and lower truck cycle times. In Figure 4.21 the normal distributions for the bucket payload of each of the loading units vary considerably. Each generation of loading units is different sizes with different size buckets that allow for larger or smaller payload values. Although the distributions show their varying capacities, the quarries they operate in are nearly the same size with similar yearly production. Given the quarries' similar profiles and different machine options, the 992K loader would be the most beneficial for this type of operation with the correct operator. The 992K generally was one of the best-performing loading units in each of the variables shown in these distributions. The

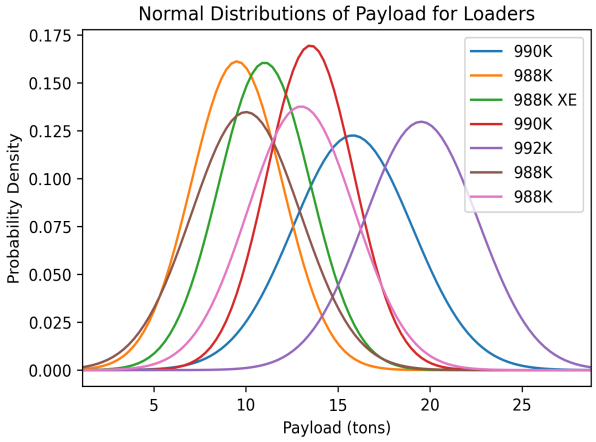

Figure 4.21: Normal Distribution of the Variable Bucket Payload for the Seven Loaders in the Rock Quarries

992K had a higher average cycle time than some of the other loaders but remained near the lower range of the ideal cycle time for loaders in this type of operation.

Total truck payload, loader cycle time, and truck cycle time were analyzed in the second data set of surface coal mines consisting of the two loader units. The surface coal mine data sets for the 993K truck had more data filled out allowing for the variables most related to production metrics to be analyzed. Figures 4.22 to 4.24 show the normal distributions of these two loaders. The variable bucket payload was not available for this data set because the software Cat MineStar Edge the total payload for each truck instead of each cycle's payload.

The two 993K loading units are the same type and size machine but one of the units outperformed the other in two of the three variable distributions. Examining the total truck payload distributions in Figure 4.22, the 993K machine, represented by the blue line, had a higher total truck payload of approximately 15 tons. The 993K, depicted as the orange line in Figures 4.23 and 4.24, had lower loader and truck cycle times by a small margin. In the future, the 993K loader could show a higher production even though it lagged slightly

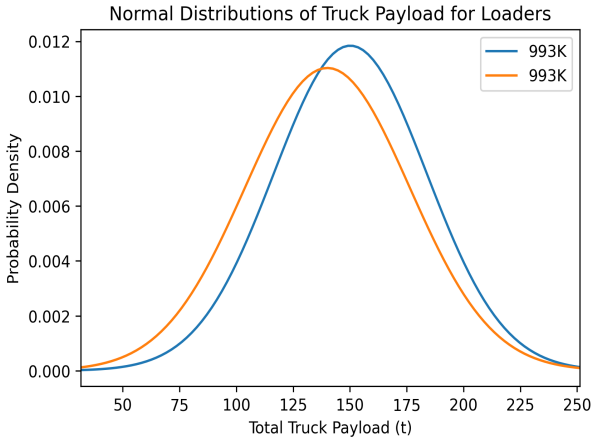

Figure 4.22: Normal Distribution of the Variable Truck Total Payload for each Loader in the Surface Coal Mine

behind in cycle time. Production variables can also be heavily influenced by operator experience, location of operation within the same mine, and the type of material the loading units are working on within the quarry. Different materials could be harder to dig into or fully fragmented from blasting allowing for easier digging and impacting production.

4.3 Machine Learning

Using Python, machine learning methods were used to analyze the data sets to initially determine correlations between the variables included. The first analysis was done on the data sets consisting of the seven loaders from the rock quarries. The first machine learning analysis used was to test linear regression as well as polynomial regression to see which fit better to the data. This regression method was used to primarily see the relationship between the variables in the data sets to predict the continuous target variables which were the loader units. Using the associated linear regression libraries within Python, the data was split into features and target variables. In the CAT Productivity model the features were Payload, Truck Payload, Cycle Time, and Fill Factor. In the CAT MineStar Edge model the

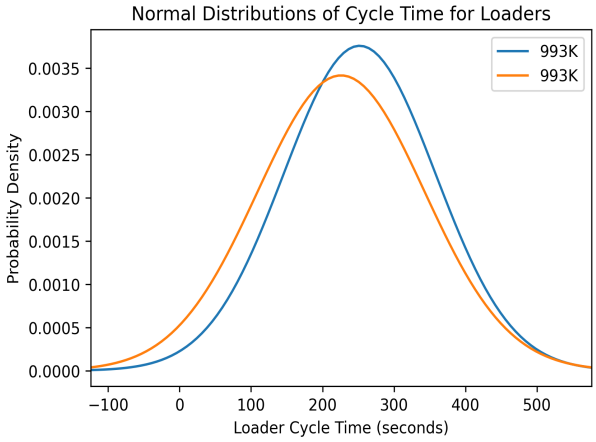

Figure 4.23: Normal Distribution of the Variable Loader Cycle Time for each Loader in the Surface Coal Mine

features were Load Duration, Total Truck Payload, Truck Cycle Time, Plan Distance and Loader Dipper Count. The target variable in this case was the loader. After the model was set up and trained and predictions were made on the training set, an R-squared value was obtained. For the linear regression model, this value was 0.098 which is a poor result. This means that 9.8% of the variance in the target variable, the loading unit, can be explained by the variables in the data sets. This essentially means that the linear regression model is poor at predicting the variation of loaders using the data. This is a surprising result as it was believed that there would be some form of linear relationship shown in the model.

The polynomial regression model showed better results but not great. The polynomial model was set up similarly to the linear regression model with just the addition of a higher degree to the function to account for the polynomial feature. Using the associated polynomial regression libraries within Python, the data was split into the features and the target variables which again were the loader units. After the model was set up and trained and predictions were made on the training set, an R-squared value of 0.366 was obtained. The polynomial regression R-squared value is better than the linear regression model but still is not a great result. The value of 0.366 means that only 36.6% of the variance in the target variable, the

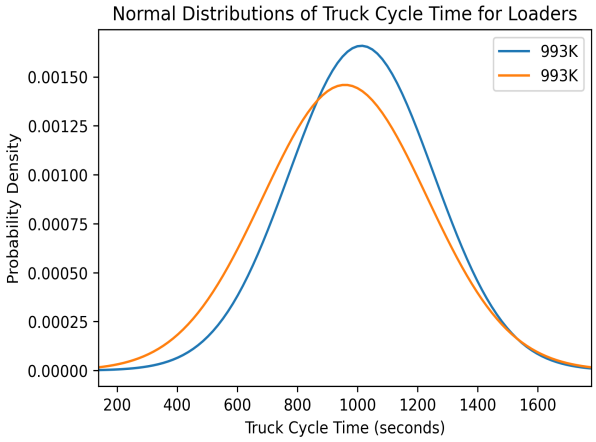

Figure 4.24: Normal Distribution of the Variable Truck Cycle Time for each Loader in the Surface Coal Mine

loading units, can be explained by the given variables in the data sets.

The data from the two loaders from the surface coal mine also were used in the linear and polynomial regression models. This data did not show as promising results as the rock quarries data. The linear regression method resulted in an R-squared value of 0.080 and the polynomial regression method resulted in an R-squared value of 0.152. Both of these R-squared values are considerably low and tell that the variables in the CAT Productivity and MineStar Edge data sets are not adequate to explain the variance in the target variable. More analysis can be done on this as there were only two loaders in this data set. Some specific features can affect the R-squared values in the regression models. Thus, it is recommended to get the R-squared using each feature separately to eliminate the features that could be adding noise to the regression models Table 4.1 shows the summary of the R-squared values across both data sets and both models.

Heat maps were obtained alongside the linear regression algorithm to display the correlation between each of the variables in the data sets. The heat map generated from the CAT Productivity data set is displayed in Figure 4.25. The heat map generated from the CAT

	Data Sets	
	CAT Productivity	**CAT MineStar Edge**
Linear Regression	0.098	0.080
Polynomial Regression	0.366	0.152

Table 4.1: R-squared values Obtained from the Linear Regression and Polynomial Regression Machine Learning Methods

MineStar Edge data set is displayed in Figure 4.26. In the CAT Productivity heat map, it is surprising to see almost no correlation between any of the variables. The variables fill factor and payload have a high correlation but that is given since payload is used to calculate the fill factor. The CAT MineStar Edge heat map displayed similar results as there are high correlations between the variables in this data set. The highest found was between the payload match score and the loader dipper count which was 0.54. These two variables do not correlate in any way to each other in definition so this calculated correlation was thrown out. The highest aside from this was the loader dipper count and the load duration with a 0.40 correlation. This is also surprising because the number of buckets to fill up the truck should be indicative of the duration of the load event. The fact that it is only 0.40 indicates there might be faulty data. Further research and more data would help to investigate this.

The next machine learning technique attempted was prediction models. The goal of this prediction was to utilize production metrics as the predictor variables to identify the type of loader. This prediction method was used to investigate how different the loader's production in similar mines was from each other. The benefit of these types of prediction models is that they can work backward. Future models should allow the user to input the type of loader and it should predict what the production metrics should be based on the mine they are working in. These future models however would require a larger spread of loader model types. This production prediction based on the loader model can be the basis for additional

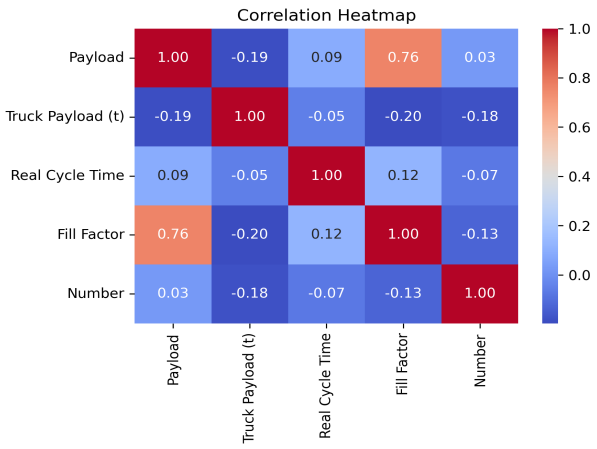

Figure 4.25: Heat Map Displaying Correlation Between Each Variable in the CAT Productivity Data Set

Machine Learning Method	**Accuracy of Prediction**
Neural Network	0.9026
K-Nearest Neighbor	0.8413
Random Forest	0.8104
Decision Tree	0.8062
Logistic Regression	0.7694

Table 4.2: Accuracy scores of the five machine learning algorithms used on the CAT Productivity data set for predictions. Testing data was used to get model accuracies.

future research. The models utilized in this study were K-Nearest Neighbors, Decision Trees, Random Forests, Neural Networks, and Logistic Regression. The full data set from CAT Productivity was run through each of these machine learning classifier algorithms and output an accuracy score. This accuracy score identifies how well the algorithm can predict values based on the training data set initially used in the algorithm. Table 4.2 shows the accuracy scores of the five models from the greatest to the least.

Of the five algorithms used, the Neural Network was the most accurate followed closely by the rest of the models. All five of the model has accuracy scores greater than 75% which are great results. Generally, any model with an accuracy score greater than 70% is considered

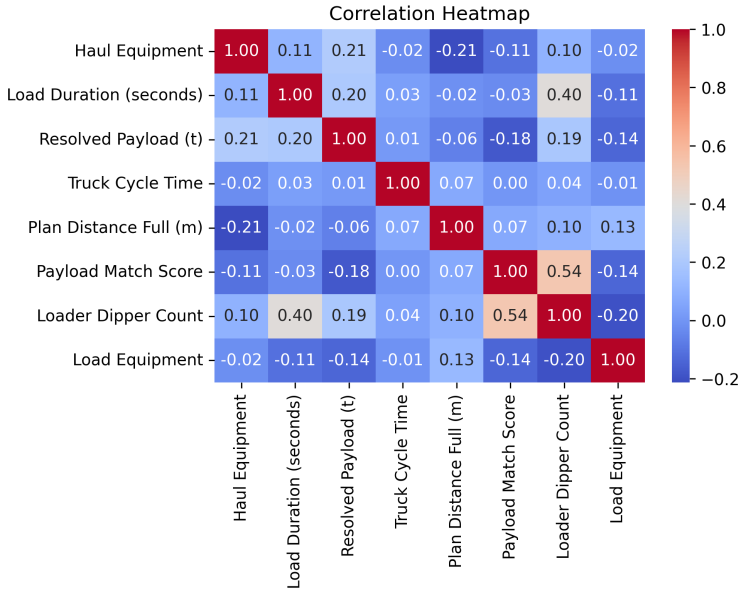

Figure 4.26: Heat Map Displaying Correlation Between Each Variable in the CAT MineStar Edge Data Set

to have great model performance. With a 90.26% accuracy, the Neural Network model was able to accurately estimate the target variable, the loading units. This also can go backward as the model could estimate the performance of any loader based on the full analysis of the performance metrics input. Future research and additional loaders' data collected would allow for predicting a loader's production metrics in an operation given the type of operation and the type of loader.

The same prediction models were used on the CAT MineStar Edge data set to predict the loading unit based on the data. The five models used in this analysis output accuracy scores and these can be seen in descending order in Table 4.3.

Of the five algorithms used for the analysis of the CAT MineStar Edge data set, the Random Forest model was the most accurate. The Random Forest model accurately estimated the target variable, the two loading units, with an accuracy score of 79.82%. The spread of accuracy scores for this data set was large. The neural network performed the worst for this

Machine Learning Method	**Accuracy of Prediction**
Random Forest	0.7982
Decision Tree	0.7485
K-Nearest Neighbors	0.7018
Logistic Regression	0.6220
Neural Network	0.5849

Table 4.3: Accuracy scores of the five machine learning algorithms used on the CAT MineStar Edge data set for predictions. Testing data was used to get model accuracy.

data set with a score of 58.49%. The models used for this data set are almost all lower than the models that the CAT Productivity data set was run through. A deeper analysis could be done on this as the MineStar Edge data set has more variables to use but also incorporates haul truck data while the Productivity data set only uses the loader data.

Chapter 5

GET Maintenance Application

5.1 Methodology

A thorough GET (Ground Engaging Tools) maintenance analysis is conducted using a combination of Microsoft Excel and Google Colab. Microsoft Excel is used to collect the average value of production and cost analysis, while Google Colab was primarily used for data visualization. To begin the analysis, the two data sets are combined which included the production metrics for the 988K loader that maintenance was recorded on and the data sets of the maintenance days GET replacement was conducted. This allowed us to compare the production average before and after maintenance, allowing us to better understand the impact of GET maintenance on production.

Using the PivotTable function in Excel, the average payload value for each day is first calculated. Then this data is matched with the days that maintenance was performed, creating a table of average production values before and after each day. For additional insight, several new variables are created by calculating the average value for payload for different periods of time before and after maintenance. This included five days, four days, three days, two days, one day, and the day of maintenance. Next, using the Matplotlib the distribution of these average payload values around maintenance is visualized. This allowed us to identify any patterns seen in the data based on how the GET maintenance impacted production.

Price of Product ($/ton)	5
Maintenance Time (hour)	1
Daily Production Time (hours)	7.5
Number of Buckets Per Truck	7
Cycle Time (seconds)	50
Price of Teeth ($)	2181.76

Table 5.1: Assumed values for calculation of production value for each set of production days leading up to and after maintenance

In order to conduct a cost analysis of the product of the production value during these periods, a few assumptions are made when calculating the cost. These assumptions can be found in Table 5.1. For example, a price of $5 per ton of product is assumed for a speculative analysis of limestone rock at the time. Through conducting this analysis, insights are gained into the economic impact of GET maintenance on production and were able to identify any areas for improvement.

Once the production values for each period before and after maintenance were calculated, a percent change was then calculated by comparing the before and after values of each maintenance day. The percent changes of production value and maintenance days were input into Google Colab and again with the matplotlib.pyplot library was plotted for data visualization.

5.2 GET Maintenance Results

Using the payload production values from Cat Productivity and the maintenance records of GET changes, an analysis was conducted on the change of production before and after maintenance. Only one maintenance record was obtained for a single 988K loading unit from one rock quarry. Unfortunately, many operations may not record when they perform maintenance or they may use paper records, which can easily be misplaced, to identify

when GET maintenance operations occurred. GET maintenance is quick, taking about 30 minutes to an hour, based on how many teeth need to be replaced so some operations may not consider this noteworthy to keep records of. Figure 5.1 shows the average production of each selection of days leading up to maintenance. So, -4 identifies the four days preceding maintenance, -2 identifies the two days preceding maintenance, 2 identifies the two days after maintenance, and so forth up to five days before and after maintenance. Figure 5.2 shows the percentage change in production value based on a cost analysis discussed in Section 5.1. The maintenance date is based on the date of each change in GET conducted on the 998K machine. The average payload used in the production cost analysis was determined by taking the set of days from the previous maintenance day up to the day in question and then also taking the set of days from the next maintenance day to the previous day in question. The percent change was calculated based on these days.

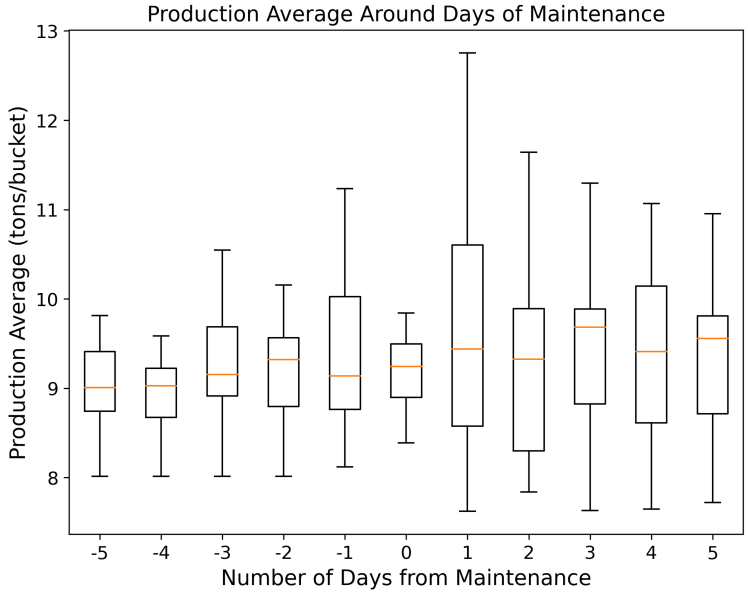

Figure 5.1: Box Plot of Average Production for Each Set of Days Leading up to and After Maintenance on the GET of a 988K Loader

The production average leading up to and after maintenance presented in Figure 5.1 suggests there is no distinguishable material change in production once the GET were replaced on

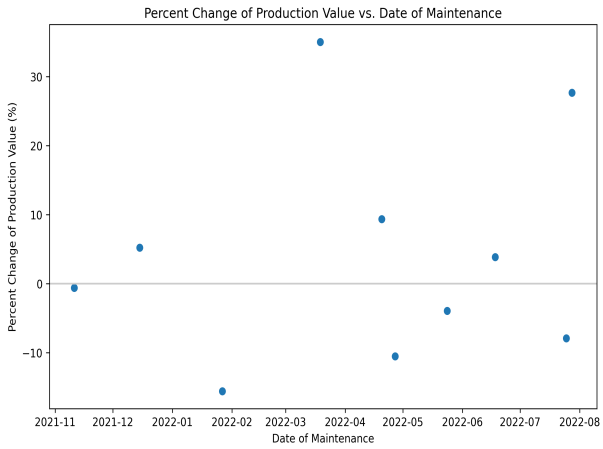

Figure 5.2: Percent Change Comparing Production Value from Before and After each Day of Maintenance on the GET of a 988K Loader

the loader bucket. Two sets of production days, -4 and -5, have low average payloads, but when comparing the sets of -3 to -1 to the sets of 2 to 5 there is relatively no change in the average production. The production is increasing before the maintenance and on the day of maintenance is high compared to the other values. The high value could be due to maintenance being performed at the start of the shift but there is no other indicator besides the date in the records of the time when maintenance occurred. The set of one day after maintenance has the highest production average but has the potential to be considered an outlier when tied with the sets of days after maintenance. The distribution of data one day after maintenance is high as the box and whisker shows a large spread so this could indicate the outlier due to one day of high production.

Figure 5.2 illustrates the percentage change of production value around the day of maintenance and suggests there is no distinguishable material change in production due to maintenance. The percentage change spans an even split of values in negative and positive change. It is worth noting that the positive percentage change has a higher range going up to approximately 32% while the negative percent change goes down to only about -15%. Initially, it was speculated that if GET were changed on the bucket, it would benefit the operation and

production would increase. Based on the data and the results calculated from this data, the results are inconclusive as to whether the current process of GET maintenance is beneficial or if operators are changing the GET too frequently. GET maintenance is generally determined by the operator's decision because they experience a harder time digging into the rock face or they visually notice too much wear. Therefore, for better results in the future, there should be a standard low measurement for the teeth that objectively indicates maintenance to be performed to change them out.

Chapter 6

Conclusions and Future Work

This thesis conducted an exploratory analysis of data for nine loading units using data analysis, statistical analysis, and machine learning techniques. Seven of the loading units' data came from the software CAT Productivity and two of the loading units' data came from the software CAT MineStar Edge. Multiple variables were used to identify key production parameters including bucket payload, loader cycle time per bucket, fill factor, truck cycle time, truck total payload, and loader cycle time per truck. In addition, an analysis of production values before and after GET maintenance was conducted to determine if the maintenance significantly impacted production values. Section 4.1 focused on data analysis to find the relationship between the average payload of each loader and their cycle times. This relationship was found to be inconclusive in determining an optimum point at which production can be increased down the line. The data across each loader varies between cycle time and the average payload. However, three loaders seem to indicate an optimum cycle time between 30 seconds to 40 seconds to yield a higher payload.

Statistical analysis was conducted in Section 4.2 to compare the distribution of production variables across the loaders performing in a similar work environment. The Caterpillar 992K and one of the 990K loading units were found to be the highest-yielding machines. In the analysis, one of the Caterpillar 993K loading units outperformed a similar 993K machine on all production variables for unknown reasons that will be the basis of future research. Influencing variables can include operator experience, location of operation within the same

mine, and type of material.

Machine learning was used in Section 4.3 to investigate correlations between variables in both data sets. Through a linear and polynomial regression model that each data set was put through, there were no significant correlations between any of the variables. Although there was a correlation between loader dipper count and the load duration in the CAT MineStar data set, these go hand in hand. Theoretically, the correlation should have been much higher. This indicates that future research should be done investigating why this correlation was not what it theoretically should be. For the CAT Productivity data set, the Neural Network algorithm created a model that was 90.26% accurate in estimating the loading unit based on the input variables. The K-Nearest Neighbor algorithm came in a close second with an accuracy of 84.13%. In future operations, this means that production variables can be predicted given the type of loader. For the CAT MineStar Edge data set, the Random Forest algorithm created a model with the highest accuracy of 79.82%. This is still a great accuracy score but surprising that the addition of more variables or the inclusion of haul truck data as well in the data sets may have caused this decrease. This can be an area for future investigation. The Decision Tree algorithm came in a close second with an accuracy of 74.85%.

After generating an average production comparison in Section 5.2 of before and after maintenance on the Ground Engaging Tools on loader buckets, there was no material change in the average production of the mine. This analysis still does not answer the question of whether they are replaced too frequently. The future analysis would attempt to create a prediction model for optimal maintenance intervals on the Ground Engaging Tools on loader buckets.

Data from CAT Productivity was found to be less detailed than data from CAT MineStar Edge and largely subject to human error. The task is repetitive because these cycles are short (less than one minute for the majority) and performed over an eight-hour shift. The

data reliability is low due to operator input. Future work could help improve these models described in the paper. Expanding to an additional year of research will provide more data to yield a more precise predictive model. The push for the digitization of machine information will also allow more machines to populate this software with data allowing for a wider variety of machine types to be considered and further refinement of predictive models. This would include haul truck units as well. Adding haul trucks and synchronizing the loader and haul trucks during the loading events would allow users to pinpoint where in the operation there is an opportunity for improvement and which units need improved training for their operators. As the popularity of remote monitoring increases, this will lead to more companies subscribing to these types of software and upgrading their machines to accommodate them. Site visits would help with future research to get an understanding of the operation and tasks required by loaders during their shifts. This insight could explain irregularities in the data but also allow for time studies to be conducted to figure out choke points in operation. Additional research should be done on ways to characterize a Ground Engaging Tool that needs to be replaced instead of relying on operator opinion, eliminating the potential for human error.

Bibliography

[1] Soofastaei, Ali. *Data Analytics Applied to the Mining Industry.* CRC Press, 2020.

[2] Fatemeh Molaei, Elham Rahimi, Hossein Siavoshi, Setareh Ghaychi Afrouz, and Victor Tenorio. A comprehensive review on internet of things (iot) and its implications in the mining industry. *American Journal of Engineering and Applied Sciences*, 13(3):499–515, 2020.

[3] Abdullah Aziz, Olov Schelén, and Ulf Bodin. A study on industrial iot for the mining industry: Synthesized architecture and open research directions. *IoT*, 1(2):529–550, 2020.

[4] Kiernan Brent George. *Analysis of Lightweight Cryptographic Primitives.* PhD thesis, Virginia Tech, 2021.

[5] Karen Rose, Scott Eldridge, and Lyman Chapin. The internet of things: An overview. *The internet society (ISOC)*, 80:1–50, 2015.

[6] S Soumya, Malini Chavali, Shuchi Gupta, and Niharika Rao. Internet of things based home automation system. In *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pages 848–850. IEEE, 2016.

[7] Farahnaz Sadoughi, Ali Behmanesh, and Nasrin Sayfouri. Internet of things in medicine: a systematic mapping study. *Journal of biomedical informatics*, 103:103383, 2020.

[8] Dongxin Lu and Tao Liu. The application of iot in medical system. In *2011 IEEE*

International Symposium on IT in Medicine and Education, volume 1, pages 272–275. IEEE, 2011.

[9] Shancang Li, Li Da Xu, and Shanshan Zhao. The internet of things: a survey. *Information systems frontiers*, 17:243–259, 2015.

[10] Rafael Laskier. Modernizing the mining industry with the internet of things. *Internet of Things and Data Analytics Handbook*, pages 521–543, 2017.

[11] Maria Holmlund, Yves Van Vaerenbergh, Robert Ciuchita, Annika Ravald, Panagiotis Sarantopoulos, Francisco Villarroel Ordenes, and Mohamed Zaki. Customer experience management in the age of big data analytics: A strategic framework. *Journal of Business Research*, 116:356–365, 2020.

[12] Chong-chong Qi. Big data management in the mining industry. *International Journal of Minerals, Metallurgy and Materials*, 27:131–139, 2020.

[13] Jonatan Adam Fekete. Big data in mining operations. *Copenhagen Business School*, 2015.

[14] Chaudhery Mustansar Hussain, Mosae Selvakumar Paulraj, and Samiha Nuzhat. *Source reduction and waste minimization*. Elsevier, 2021.

[15] Viktor Mayer-Schönberger and Kenneth Cukier. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.

[16] Kristine L Pankow, Jeffrey R Moore, J Mark Hale, Keith D Koper, Tex Kubacki, Katherine M Whidden, and Michael K McCarter. Massive landslide at utah copper mine generates wealth of geophysical data. *Gsa Today*, 24(1):4–9, 2014.

[17] Caterpillar Inc. Cat productivity. https://s7d2.scene7.com/is/content/Caterpillar/CM20190403-794f9-36f71, 2019.

[18] Caterpillar Inc. Operation and maintenance manual: Cat minestar edge production recording. https://s7d2.scene7.com/is/content/Caterpillar/CM20200115-7b6ae-db164, 2020.

[19] Scott Gamble, Craig S Coker, Frank Franciosi, and Robert Rynk. Composting operations and equipment. In *The Composting Handbook*, pages 341–408. Elsevier, 2022.

[20] Walter G Koellner, Gerald M Brown, José Rodríguez, Jorge Pontt, Patricio Cortés, and Hernán Miranda. Recent advances in mining haul trucks. *IEEE Transactions on Industrial Electronics*, 51(2):321–329, 2004.

[21] Ignacio Andrés Osses Aguayo, Micah Nehring, and GM Wali Ullah. Optimising productivity and safety of the open pit loading and haulage system with a surge loader. *Mining*, 1(2):167–179, 2021.

[22] Gabriel Santelices, Rodrigo Pascual, Armin Lüer-Villagra, Alejandro Mac Cawley, and Diego Galar. Integrating mining loading and hauling equipment selection and replacement decisions using stochastic linear programming. *International Journal of Mining, Reclamation and Environment*, 31(1):52–65, 2017.

[23] Elmira Tajvidi Asr, Reza Kakaie, Mohammad Ataei, and Mohammad Reza Tavakoli Mohammadi. A review of studies on sustainable development in mining life cycle. *Journal of Cleaner Production*, 229:213–231, 2019.

[24] Markus HA Piro and Ksenia Lipkina. Mining and milling. In *Advances in Nuclear Fuel Chemistry*, pages 315–329. Elsevier, 2020.

[25] Caterpillar. Get more wheel loader productivity. https://www.cat.com/en_US/articles/for-owners/how-to-calculate-cat-wheel-loader-productivity.html, 2023.

[26] J Matsimbe. Optimization of shovel-truck productivity in quarries. *Int. J. Res. Advent Technol*, 8(10):1–9, 2020.

[27] INM Nday and H Thomas. Optimization of the cycle time to increase productivity at ruashi mining. *Journal of the Southern African Institute of Mining and Metallurgy*, 119(7):631–638, 2019.

[28] Brighton Samatemba, Long Zhang, and Bunda Besa. Evaluating and optimizing the effectiveness of mining equipment; the case of chibuluma south underground mine. *Journal of Cleaner Production*, 252:119697, 2020.

[29] Marina Paolanti, Luca Romeo, Andrea Felicetti, Adriano Mancini, Emanuele Frontoni, and Jelena Loncarski. Machine learning approach for predictive maintenance in industry 4.0. In *2018 14th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications (MESA)*, pages 1–6. IEEE, 2018.

[30] Simon Robatto Simard, Michel Gamache, and Philippe Doyon-Poulin. Current practices for preventive maintenance and expectations for predictive maintenance in east-canadian mines. *Mining*, 3(1):26–53, 2023.

[31] Ernnie Illyani Basri, Izatul Hamimi Abdul Razak, Hasnida Ab-Samat, and Shahrul Kamaruddin. Preventive maintenance (pm) planning: a review. *Journal of Quality in Maintenance Engineering*, 23(2):114–143, 2017.

[32] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. Supervised learning. *Machine learning techniques for multimedia: case studies on organization and retrieval*, pages 21–49, 2008.

[33] Vladimir Nasteski. An overview of the supervised machine learning methods. *Horizons. b*, 4:51–62, 2017.

[34] Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, and Khairullah Khan. A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, 1(1):4–20, 2010.

[35] Peter Dayan, Maneesh Sahani, and Grégoire Deback. Unsupervised learning. *The MIT encyclopedia of the cognitive sciences*, pages 857–859, 1999.

[36] Muhammad Usama, Junaid Qadir, Aunn Raza, Hunain Arif, Kok-Lim Alvin Yau, Yehia Elkhatib, Amir Hussain, and Ala Al-Fuqaha. Unsupervised machine learning for networking: Techniques, applications and research challenges. *IEEE access*, 7:65579–65615, 2019.

[37] Christian Lopez, Scott Tucker, Tarik Salameh, and Conrad Tucker. An unsupervised machine learning method for discovering patient clusters based on genetic signatures. *Journal of biomedical informatics*, 85:30–39, 2018.

[38] Pouya Nobahar, Yashar Pourrahimian, and Fereidoun Mollaei Koshki. Optimum fleet selection using machine learning algorithms—case study: Zenouz kaolin mine. *Mining*, 2(3):528–541, 2022.

[39] Jieun Baek and Yosoon Choi. Deep neural network for predicting ore production by truck-haulage systems in open-pit mines. *Applied Sciences*, 10(5):1657, 2020.

[40] David A Swanson. On the relationship among values of the same summary measure of error when used across multiple characteristics at the same point in time: an examination of malpe and mape. *Review of Economics and Finance*, 5(1), 2015.

[41] Caterpillar Inc. *Cat Product Link Parts and Service Reference Guide*. Caterpillar Inc.

[42] Caterpillar Inc. *Connectivity Application Selection Guide - Radios*. Caterpillar Inc.

[43] Caterpillar Inc. *Cat Product Link Technology*. Caterpillar Inc.

[44] WesTrac. Minestar edge. https://www.westrac.com.au/technology/minestar/minestar-fleet/minestar-edge, 2023.

[45] WesTrac. Cat productivity. https://www.westrac.com.au/technology/cat-technology/cat-productivity, 2023.

[46] Metin Özdoğan and Hakkı Özdoğan. Cycle time segments and cycle time distribution curves of mining size wheel loaders-a case study. *Scientific Mining Journal*, 56(1):13–21, 2017.

[47] Siyuan Song, Eric Marks, and Nipesh Pradhananga. Impact variables of dump truck cycle time for heavy excavation construction projects. *Journal of Construction Engineering and Project Management*, 7(2):11–18, 2017.

[48] Caterpillar Inc. *CAT Operation and Maintenance Manual Cat MineStar Edge Production Recording*. Caterpillar Inc.

[49] Ekaba Bisong and Ekaba Bisong. Google colaboratory. *Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners*, pages 59–64, 2019.

[50] Sung-Hyuk Cha. Comprehensive survey on distance/similarity measures between probability density functions. *City*, 1(2):1, 2007.

[51] Jeremy Miles. R-squared, adjusted r-squared. *Encyclopedia of statistics in behavioral science*, 2005.

[52] Daniel T Larose and Chantal D Larose. k-nearest neighbor algorithm. 2014.

[53] Zhongheng Zhang. Introduction to machine learning: k-nearest neighbors. *Annals of translational medicine*, 4(11), 2016.

[54] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[55] Yanli Liu, Yourong Wang, and Jian Zhang. New machine learning algorithm: Random forest. In *Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3*, pages 246–252. Springer, 2012.

[56] Steven J Rigatti. Random forest. *Journal of Insurance Medicine*, 47(1):31–39, 2017.

[57] J. Ross Quinlan. Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, 28(1):71–72, 1996.

[58] Bahzad Taha Jijo and Adnan Mohsin Abdulazeez. Classification based on decision tree algorithm for machine learning. *evaluation*, 6(7), 2021.

[59] Madan Somvanshi, Pranjali Chavan, Shital Tambade, and SV Shinde. A review of machine learning techniques using decision tree and support vector machine. In *2016 international conference on computing communication control and automation (ICCUBEA)*, pages 1–7. IEEE, 2016.

[60] Fadi Thabtah, Neda Abdelhamid, and David Peebles. A machine learning autism classification based on logistic regression analysis. *Health information science and systems*, 7:1–11, 2019.

[61] Hsinchun Chen. Machine learning for information retrieval: Neural networks, symbolic learning, and genetic algorithms. *Journal of the American society for Information Science*, 46(3):194–216, 1995.

[62] Rene Y Choi, Aaron S Coyner, Jayashree Kalpathy-Cramer, Michael F Chiang, and J Peter Campbell. Introduction to machine learning, neural networks, and deep learning. *Translational Vision Science & Technology*, 9(2):14–14, 2020.