Marshall University
Marshall Digital Scholar

Theses, Dissertations and Capstones

2023

Identifying hazardous patterns in MSHA data using random forests

Olivia Milam
milam33@marshall.edu

Follow this and additional works at: https://mds.marshall.edu/etd

 Part of the Computer Engineering Commons, Risk Analysis Commons, and the Social and Behavioral Sciences Commons

Recommended Citation

Milam, Olivia, "Identifying hazardous patterns in MSHA data using random forests" (2023). *Theses, Dissertations and Capstones*. 1813.
https://mds.marshall.edu/etd/1813

This Thesis is brought to you for free and open access by Marshall Digital Scholar. It has been accepted for inclusion in Theses, Dissertations and Capstones by an authorized administrator of Marshall Digital Scholar. For more information, please contact beachgr@marshall.edu.

IDENTIFYING HAZARDOUS PATTERNS IN MSHA DATA USING RANDOM FORESTS

A thesis submitted to
Marshall University
in partial fulfillment of
the requirements for the degree of
Master of Science
in
Computer Science
by
Olivia Milam
Approved by
Dr. James D. McIntosh, Committee Chairperson
Dr. Tanvir Irfan Chowdhury
Dr. Ammar Alzarrad
Dr. Haroon Malik

Marshall University
August 2023

Approval of Thesis

We, the faculty supervising the work of Olivia Milam, affirm that the thesis, *Identifying Hazardous Patterns in MSHA Data Using Random Forests*, meets the high academic standards for original scholarship and creative work established by the Department of Computer Sciences and Electrical Engineering and the College of Engineering and Computer Sciences. This work also conforms to the editorial standards of our discipline and the Graduate College of Marshall University. With our signatures, we approve the manuscript for publication.

Professor James D. McIntosh, Department of Industrial and Systems Engineering	Committee Chairperson	3-29-23 Date
Dr. Tanvir Irfan Chowdhury, Department of Computer Sciences and Electrical Engineering	Committee Member	04-04-2023 Date
Dr. Ammar Alzarrad, Department of Civil Engineering	Committee Member	03/29/2023 Date
Dr. Haroon Malik, Department of Computer Sciences and Electrical Engineering	Committee Member	04-07-2023 Date

© 2023
Olivia Milam
ALL RIGHTS RESERVED

Acknowledgments

My thanks and gratitude to Dr. Haroon Malik.

Table of Contents

List of Tables

List of Figures

Abstract

Mining safety and health in the US can be better understood through the application of machine learning techniques to data collected by the Mine Safety and Health Administration (MSHA). By identifying hazardous conditions that could lead to accidents before they occur, valuable insights can be gained by MSHA, mining operators, and miners. In this study, we propose using a Random Forest machine learning model to predict whether a given mining violation will lead to an accident, and if so, whether it will be fatal or non-fatal. To achieve this, the model is trained on MSHA violation data and the sum of scheduled accident charges within 35 days of the violation. We experiment with different predictive models using varying data columns, training set sizes, prediction classes, and hyperparameters to achieve a reliable prediction. One of the challenges in generating these models is accurately predicting the sparse class of accidents, as opposed to the abundant class of no accidents. To address this, we propose utilizing sample minimizing to balance the false negative and false positive rate and create a more accurate predictive model. Our results demonstrate, with a high degree of confidence, the potential for machine learning to improve mine safety and health by identifying hazardous conditions and mitigating the risk of accidents.

Chapter 1: Introduction

1.1 Motivation

All employees have the right to a safe workplace. A safe workplace allows employees to satisfy their job requirements without compromising bodily safety and health. The goal of any workplace safety program is to eliminate injuries and fatalities by properly controlling or removing hazards. Effective workplace safety programs collect data on hazards, policy violations, and injuries on a periodic basis. The collected data not only facilitates gauging the safety aspect of a workplace, but it also provides valuable information that could be used to find patterns, identify new information, and forecast future issues. In the routinely collected safety records, there could be actionable information that may be used to prevent a dangerous situation from occurring.

Often, this type of historical data is not used to its full potential for prediction or forecasting. The data is often used as a measuring stick for whether a safety goal was met or unmet – an entry in a spreadsheet for book-keeping or record keeping purposes. If a metric becomes particularly high, perhaps that area will receive some remediation if the correct stakeholder notices. Moreover, these metrics are usually measuring a particular season of time at a particular place, not observing the full gamut of other locations, analogous scenarios, or prior knowledge.

Liberating predictive knowledge out of this historical data, hopefully, can be used to close a feedback loop to improve safety. Using the data as a measurement of safety is analogous to taking a daily temperature of the environment. It is an important prerequisite to deeper understanding of what causes unsafe environments. Predictive knowledge is analogous to forecasting the weather. However, the analogy stops short in that with safety, human action can

change what is written in the forecast. Enacting change, or completing the feedback loop, can allow workplaces to improve safety.

1.2 Past Mining Incidents

There are several mining disasters in living memory that underline the importance of preventing accidents and ensuring safety. One such incident that occurred in West Virginia in 2010 was the Upper Big Branch mine disaster. Twenty-nine miners tragically died and two were injured in an explosion. The investigative Executive Summary report of the disaster by MSHA (Mine Safety and Health Administration) states, "the physical conditions that led to the explosion were the result of a series of basic safety violations at [Upper Big Branch] and were entirely preventable" [1, p. 2].

In 2006, at Darby Mine No. 1 in Kentucky, five miners were fatally injured due to an explosion [2]. In the investigation report, MSHA states: "The accident occurred because the operator did not observe basic mine safety practices and because critical safety standards were violated" [3, p. 56]. Moreover, "the company was cited for six conditions and/or practices which contributed in some way to the accident" [3, p. 1]. This mine was sealed following the accident [3, p. 55]. Disregard for safety practices and violating safety standards directly contributed to this accident.

Aracoma Alma Mine #1 in West Virginia had an accident in 2006 that resulted in the deaths of two miners [2]. MSHA reports:

> "As a result of the investigation, MSHA issued 25 citations and orders for violations which contributed to the cause or severity of the accident. Of these, 21 were the result of reckless disregard on the part of the mine operator" [4, p. 2].

Two of the fatalities were found to be directly connected to several contributory violations [4, p. 2]. This report highlights the fact that violations or citations do identify hazardous conditions and can cause serious fatal accidents if not properly addressed.

The Sago Mine disaster in 2006 in West Virginia resulted in twelve miners receiving fatal injuries due to an explosion [2]. The final conclusion and root causes of the report do not indicate specific prior violations, but indicate lightning likely ignited methane gas and caused an explosion, which unsealed unused parts of the mine that had elevated carbon monoxide that then entered occupied areas [5, pp. 187-188]. However, "the operator was subjected to a higher level of enforcement pursuant to section 104(d)" [5, p. 8] due to prior MSHA inspection results.

Such disasters highlight how dangerous mining is and how steps must be taken to safeguard miner safety. One theme in most of the MSHA reports for these listed incidents is that if corrective safety actions were taken before the incidents, then, at the very least, the severity of the accidents would have been lessened.

1.3 Regulations

One of MSHA's tasks is to regulate the mining industry to ensure best practices, requirements, and policies are followed. These rules are established to keep miners safe. Inspections are a tool to ensure mining operators are complying with federal laws. MSHA inspectors are examining the environment of the mine looking for potential hazards and policy violations. Underground mines must be inspected at least four times a year by MSHA inspectors [6]. These inspections ensure operators and miners follow all required safety protocols and look for hazardous conditions. Through the course of their work, MSHA inspectors log their findings into a publicly available MSHA database. In addition, mine operators are required to send

MSHA reports when certain events occur, such as an accident. Operators are also required to send quarterly reports on production or other notable events.

Early identification of dangerous conditions in mines could be used as a tool to improve mine safety. MSHA currently has a Pattern of Violations (PoV) report where they notify mines operating in dangerous elevated conditions. This report is created by compiling violations and accidents to create statistics that are used as a heuristic to determine if the mine shows poor conditions. The compiled report indicates a historical trend of past violations. One major issue with the PoV report is that it requires a long trend, i.e., a chain of many violations, before action. However, studying such trends, i.e., the recorded pattern of violations, can facilitate in predicting not only the potential future violations in a timely manner but also prevent them.

1.4 Thesis Hypothesis

This thesis brings forward the following Hypothesis:

"It is possible to create a predictive model to determine hazardous mining conditions by mining the large volume of MSHA violation and accident data."

More simply, this thesis envisions that by leveraging a machine learning algorithm, predictive models can be created from the large volume of violation and accident data MSHA has recorded in their database. Using these models, it is possible to predict if an accident will occur within thirty-five days of a particular safety violation. This time span was selected because it is a long enough period to reflect potential issues and was based on a prior work [7]. Such an early prediction of potential violations can facilitate the mining operators to rectify the especially hazardous conditions —related to safety violations— before violations turn into an accident that can cost miners' lives.

1.4.1 Research Questions

Primary **Q**uestion (*PQ*)**: Can we utilize MSHA violation data to detect a potential future mining accident?**

***PQ* Explanation**: The aim of the thesis is to take MSHA mining safety records as an input to train a machine-learning model, then take a new unseen MSHA violation input to assess, and finally output how the new input ranks as a hazard. The output of this machine learning model will be a number of classes that reflects the likelihood of an accident based on the inputted safety violation. The thesis aims to produce a useful result to a mining safety subject matter expert that can then be acted on to improve safety.

There are many layers to solving this question, such as determining:

- What machine learning technique to use.
- What data to select as an input(s).
- What is the best measure of a violation.
- How tuning the machine learner can yield optimal results.

***PQ Findings*:** MSHA violation data does have predictive features that can be used to predict the likelihood of an accident. The Random Forest (RF) machine learning technique was selected because it produces explainable models. MSHA's violation and accident datasets contained the most important training features. Exploratory analytics was conducted on MSHA violation data to remove non-predictive and unrelated attributes. Related and important features were found and identified by reading MSHA data definitions and model experimentation. The SCHEDULE_CHARGE attribute in the MSHA accident's table was determined to be an important attribute to predict, with a thirty-five day timeframe. Many different iterations of models were used to help determine model quality and improve each model.

***RQ 1*:** What MSHA data is needed to build a robust prediction model?

***RQ* 1 Explanation:** MSHA has amassed a large body of accurate data on U.S. mines through inspections and operator reports. Machine learning techniques can be applied to this data to build and train an incident prediction model. Quality of the prediction model greatly depends upon the quality of the underlying data. In order to make machine learning work well on new tasks, it might be necessary to design and train better features. Therefore, feature engineering is applied to identify important features (i.e., attributes) essential for building a quality predictive model. Feature engineering works to identify features that are deemed to facilitate prediction (among the hundreds available in the MSHA data) and removes features that act as noise to the model. At a high-level, feature engineering removes all the features that do not aid in answering the question "if an accident is likely to occur". A feature that facilitates predicting the likelihood of accident to occur is retained. For example, the "*day of week a violation occurs*" may be important. A violation on a specific day of the week may act as a weight on the importance of the violation. This could indicate intuitive knowledge that a subject area expert might know, for example, that most mines are not generally inspected on a certain day of the week. Beyond feature engineering, it could be possible other datasets could aid in the predictive strength. An example related dataset might be the average price of coal for that region at the time the violation occurred. The RQ1 seeks what independent variables, or features, are most useful and relevant for constructing incident prediction model using MSHA data.

***RQ 1* Findings:** This thesis identified a predictive set of violation features through researching MSHA data definitions and through model experimentation. Removing datetime features improved model understandability and eliminated some possible areas of data leaks.

Additionally, several sample sizes were utilized to better balance the abundant sample class of no-accidents with the sparse sample class of accidents.

***RQ 2*: What value should be predicted that best indicates a potential accident?**

***RQ* 2 Explanation:** Selecting what value to predict potential accidents from is not apparent in this dataset. There are several candidate dependent variables in the MSHA dataset that measure a hazard or accident. The difficulty lies in selecting a value to predict that subject matter experts are familiar with, have intrinsic meaning to the set at large, and is a useful metric. Careful selection will help better ensure the machine learning model applies to the real world.

***RQ* 2 Findings:** An aggregated calculation of the accident's SCHEDULE_CHARGE over thirty-five days was selected as the feature to predict. This feature was selected because it is a standardized numerical variable that MSHA uses to indicate accident's severity. MSHA has a schedule of charges to assign to an accident. For example, a schedule charge of 6,000 indicates a fatal accident and a schedule charge of 300 indicates the loss of a thumb. Narrowing predictive classes into No Accident, Non-fatal Accident, and Fatal Accident also made it possible to better analyze the predictive quality of the predictive model through a confusion matrix.

1.5 Outline

Mining stakeholders need a robust predictive model that has explainable results. This thesis presents a literature review of traditional reporting mining operators and regulators use, an overview of statistical based approaches, the impact of Internet of Things (IoT), and finally machine learning approaches and how the apply to this problem.

Next, the methodology used to create a machine learning model to predict potential safety issues is explained. The selected machine learning algorithm was the Random Forest technique.

Python was selected as the programming language to build the model using scikit-learn's Random Forest implementation. Because MSHA has a large library of data sources, data had to be selected, pruned, and prepared for use in the model. Selecting a meaningful safety variable to predict was a key part of data selection.

Three Random Forest models were created with varying results. The first model used many data features and attempted to predict the SCHEDULE_CHARGE over thirty-five days of a given violation. The second model removed some data features in an attempt to limit data leaks. The final model limited features, changed from predicting the SCHEDULE_CHARGE over thirty-five days to the category of a given schedule charge (No Accident, Non-fatal Accident, and Fatal Accident), and improved sampling of accident observations during model training.

Chapter 2: Literature Review

Determining the safety of a mine and correcting potential issues before an accident occurs is a preoccupation of both MSHA and mining operators. A few approaches to this problem are through traditional recordkeeping and reporting, statistical analysis, utilization of Internet of Things (IoT) sensor network data, and machine learning. Machine learning is most relevance to the thesis and is covered in the most detail.

2.1 Safety Recordkeeping and Reporting

2.1.1 Traditional Recordkeeping

A traditional approach to identifying hazardous mining conditions is through the use of recordkeeping and time-bound reporting. This approach is familiar to both mining operators and regulatory bodies. Operators and regulators create reports on some specified timeframe attempting to identify if the mine is trending in an unsafe direction. Trends in the data such as increasing or decreasing safety are discovered through subject-matter experts reviewing reports and making judgements based on the report. MSHA has reporting requirements for all U.S. mines, [8] and the U.S Securities and Exchange Commission (SEC) also has mandatory reporting requirements for U.S. publicly traded mining-operator companies.

The purpose of traditional reporting is to identify and act on safety hazards. Operators ideally self-regulate and create their own internal reports to correct issues before they turn into catastrophe or an injury. “Large companies tend to have better safety records than smaller companies due to greater numbers of professional engineers and better management” [9, p. 1]. Self-identifying and solving safety hazards early on can keep the mine safe and operating efficiently. Managers and engineers can adjust procedures based on reports to solve issues.

One of the MSHA traditional reports is the *Mine Injury and Worktime* report that is issued both quarterly and annually [10]. Operators are required to submit the data for these reports to MSHA because of Part 50 of Title 30 of the Code of Federal Regulations [11] [12]. Some examples of what kind of information is present in the reports include total number of fatal accidents, the number of non-fatal accidents with workdays lost, accidents with no days lost and the incident rate of these occurrences [13, p. 2]. Much of this data is also aggregated into summary form to provide regional and state-by-state safety overviews.

Publicly traded coal mining operators also have to report their safety record in their form 8-K SEC filings. They are required to report, "specified health and safety violations, orders and citations, related assessments and legal actions, and mining-related fatalities" [14]. Operators are also required to file a Form 8-K within four business days outside of periodic reporting if they, "receive notice from MSHA of an imminent danger order under section 107(a) of the Mine Act; notice of a pattern of violations under section 104(e) of the Mine Act, or notice of the potential to have a pattern of such violations" [15]. This reporting provides investors with clear information and knowledge about the safety record of the company they are investing in.

Traditional recordkeeping and reporting provides an important foundation for more sophisticated safety management. These basic reports are used by management and engineers to facilitate safe mines. The data collected for these reports will be used in machine learning and other techniques.

2.2 Statistical Analysis Based Approaches

2.2.1 Citation-related Reliability Analysis

One way to approach detecting potential mine safety issues is through citation-related reliability analysis (RA), based on statistical techniques. Harisha Kinilakodi and R. Grayson

champion the RA approach for mine safety issues. Citation-related RA is simply, “the probability of not getting a citation on a given inspector day, [and] is considered an analogue to the maintenance reliability approach, which many mine operators understand and use” [16, p. 1015]. Harisha Kinilakodi and R. Grayson demonstrate this approach on 31 mines of various sizes. They emphasis that citation-related RA is independent of the size of the mines. They state that, “70% of the underground coal mines are small-size mines (less than 50 employees)” [16, p. 1017]. For those 31 mines, they calculated the probabilities for zero, one or fewer, and greater than three citations [16]. An advantage of citation-related RA is that it is familiar to mining operators and can be applied at smaller operations.

2.2.2 MSHA’s Pattern of Violations (POV)

MSHA uses the data it collects on U.S. mines to improve safety and discover issues before they occur. MSHA uses a statistical analysis approach in their quarterly mining safety reports [10]. These reports cover the state of U.S. mining safety. The statistics used in the reports could be used as a summary of the state of a mine; however, the reports are retrospective rather than predictive. Another way MSHA attempts to be more predictive is through its Pattern of Violation (POV) criteria.

A POV is used to determine if a mine is exhibiting escalating risk factors. The goal of the POV criteria is to, “identify mine operators who have demonstrated a recurring pattern of Significant and Substantial (S&S) violations of mandatory health and safety standards at their mines. An S&S violation is one that is reasonably likely to result in a serious injury or illness” [17] . S&S violations are also explicitly marked as such in the Violation table provided by MSHA.

Each mine is evaluated at least once a year to determine if it meets the Pattern of Violations criteria. Two sets of criteria are used, meeting either one will result in issuing a Notice of Pattern of Violations. When notice of a POV is issued, MSHA may order mining to stop and require remediation of the violations [17].

POV Criteria One:

1. 50 S&S violations within 12 months
2. AND 8 S&S violations / 100 inspection hours within 12 months
3. AND .5 elevated citations / 100 inspection hours within 12 months
4. AND An Injury Severity Measure (SM) greater than other similar mines

OR

POV Criteria Two:

1. 100 S&S violations within 12 months
2. AND 40 elevated citations within 12 months

The POV approach is a good heuristic for identifying unsafe mines and exhibiting a high likelihood of safety issues. However, this approach only identifies potential safety issues in mines with consistent and high-risk associated violations. It does not capture mines that may have many low-risk citations that then increase the risk of accidents.

2.3 Internet of Things (IoT) Approach

2.3.1 Internet of Things and the Future of Mining

Many related-works also highlight the potential benefits of the Internet of Things (IoT) and how it can make mining safer and more autonomous. The information that IoT sensors could provide mining operators can make mining safer through accurate and instantaneous information on the mine's environment and equipment status. Sensor nets are already common in mines to monitor air quality and other factors, but the cost-effectiveness of inexpensive IoT devices could amplify sampling [18].

A critical aspect of IoT devices is that they only provide data, not analysis. These IoT sensor nets must be combined with statistical and machine-learning models to be a genuinely effective safety tool. However, because data collection and data analysis are planned together when creating an IoT solution, an interesting synthesis or more novel solution that would be unique from machine learning alone can occur.

2.3.2 U.S. Industrial Internet of Things (IIoT)

The U.S. National Institute for Occupational Safety and Health (NIOSH) investigated existing mining sensor systems in U.S. underground coal mines to determine if they could be used to create Industrial Internet of Thing (IIoT) systems. They found that, "out of 40 percent of the installed post-accident systems require minimal or no modification to support IIoT applications" [19].

The authors discuss the potential benefits of IIoT use in underground coal mines. They discuss that such a system could be used to monitor individual employee's exposure to hazardous conditions, predictive maintenance, disaster forecasting, automation, ventilation on demand, remote diagnostics, post-accident coordination, and use in water systems [19, pp. 7-8]. Some

challenges they discuss are security and privacy, harsh physical environment, availability of networks, and creating a data analytics system specific to coal mining [19, pp. 9-10].

2.3.3 Forecasting and Prewarning of Coal Mining Safety Risks

Chong-mao et. al, discuss potential uses of IoT as applicable to coal mining [20]. They explored how an IoT network could be a beneficial pre-warning system in an underground coal mine.

Data on the physical environment can be collected and processed to alert to hazardous conditions before they occur. For example, the Chong-mao et. al state: “Before rock outbursts, there are changes and fluctuations of mine pressure, electromagnetic radiation, infrared radiation, temperatures and other data. Similarly, there are fluctuations in gas emission quantity, temperature, electromagnetic radiation, and other data before coal and gas outbursts. Similarly, there are relevant pre-warning indicators before water inrush accidents, roof accidents and fires” [20, p. 11581]. The authors also discuss how this data will need to be processed and classified to be used effectively in a real-time warning system, discuss big data, and the challenges of implementing [20].

The authors also mention that in China, “small and medium-sized coal mines have low productivity and frequent safety accidents” [20, p. 11579]. This aligns with other research on U.S. mines on operation size and appears to be a theme.

2.4 Machine Learning Approach

2.4.1 Creating Predictive Models

A machine learning approach forms a model that uses past data to make predictions when given new data that the model has never encountered before. For coal mining safety, this would

mean using historical data, such as past accidents, violations, or sensor readings, and using that data to form a model that predicts hazardous conditions in a coal mine.

There are many techniques and options when creating predictive models. For this use case, machine learning techniques that create transparent decisions are ideal. Users of the model need to understand how the model found a result using the given data.

2.4.2 Machine Learning Classification Models for More Effective Mine Safety Inspections

One of the most comprehensive uses of machine learning on MSHA data is Jeremy Gernand's use of a Random Forest model in his paper *Machine Learning Classification Models for More Effective Mine Safety Inspections* [21]. The guiding question in Gernand's paper is to find, "what types of inspection findings are most indicative of serious future incidents for specific types of mining operations" [21, p. 1]. He builds both a single regression tree to explore this question and a Random Forest model to predict the lost-time incident rate for a given year. His goal is, "predicting whether or not a fatal or serious disabling injury is more likely to occur in the following 12-month period" [21, p. 1]. The two most important factors he found for his model were number of worker-days and total penalties due.

To prepare the MSHA Part 50 data, Gernand limited the model training data to active underground coal mines and the years 2000 to 2014. The MSHA MINE_ID was used to cross-reference data. He incorporated data from the Mines, Accidents, and Violations MSHA data tables. His final dataset contained records from 310 mines. The data in the model includes mine background data, aggregated total recorded violations, and aggregated accident details. Values were aggregated by year by mine. Rate of total lost work time is the dependent variable for this experiment [21].

The dependent variable selected for the model was, “the rate of accumulated lost days from injuries to total level of effort at a particular mine in a particular year” [21, p. 7] or, he otherwise states, “the rate of total lost work time as measured in days per 100,000 worker-days of operational effort” [21, p. 2]. One important comment Gernand makes about the choice of what to predict is that there are alternate options for safety measures. He mentions fatalities and lost workdays or injuries per 200,000 worker hours as alternate safety metrics to predict. He determines these are not ideal because, the first metric does not capture enough insight into the mine and the second metric, fatalities, is too uncommon to be statistically represented properly [21].

To create the Random Forest model, Gernand used MATLAB’s treebagger function, which uses Leo Breiman’s algorithm, to create the model. The hyperparameters he selected for random tree creation include a leaf size of five and splitting to reach a purse state based on mean squared error (MSE). The hyperparameters for the random forest include a sample of 1,000 random trees, and only randomly selecting a third of possible training variables for each tree [21].

One result from Gernand’s work was detecting variable importance. The three most important variables in the model were worker-days per week, total penalties due, and number of employees [21]. He also discovers that the two most important variables, worker-days and total penalties due, negatively correlate with the dependent variable.

He discusses a few reasons why worker-days per week is such an important metric. First, a high number of worker-days means there is more chance for an accident to occur. Second, number of worker-days is a close approximation to size. He also states, “there also happens to be a well-established connection between the size of companies and their safety records. As organizations grow in size and total capital, their liability risk increases abreast giving these

larger organizations a greater incentive to put more effort in protecting the safety of their employees" [21, p. 5].

Total penalties due is also a key metric Gernand discusses. A finding he had was that incident rate decreased, as the average penalties increased in the prior year. The straightforward response Gernand has is that it could simply be operators responding to the penalties. They have monetary incentive to improve safety to avoid penalty. He also posits that penalties may tend to be for high severity issues, which can be directly identified by an inspector and corrected, instead of many harder to correct problems. Also, of great importance, he discusses, "it may also be possible that many penalties are often assessed after the fact for injuries that have already happened" [21, p. 5].

Because Gernand's work is so instrumental in this thesis, the findings of the paper were reproduced.

2.4.3 Reproduction of Machine Learning Classification Models for More Effective Mine Safety Inspections

Reproducing Gernand's work with similar parameters resulted in variable importance findings that align with the original work. However, total 'penalties due' did not rank as highly in this reproduction.

To prepare the data for the reproduction, the same MSHA Part 50 data sets were downloaded as Gernand used. The Mines data set was reduced to only include mines that are in active status, in the underground category, and produce coal. A difference from Gernand's work is that this set of mines are mines that are active, underground, coal mines as of 2020. Using exactly the same set of mines that Gernand used in 2014 is not possible because the MSHA data field only has the current mine status, not status changes or historical status. For example, a mine in 2014 may

have had the status “Active”, but in 2020 had the status of “Non-Producing”. That historical change in status is not captured in the MSHA database. That means it is not possible from the data to determine the mines in Gernand’s sample from the data directly. After selecting mines based on those attributes, 143 mines meet the criteria for modeling. Gernand’s set of eligible mines was 310. In addition to selecting certain mines to model, new model training fields were calculated based on the Accidents and Violations data set, as done in Gernand’s paper.

Worker days per week was calculated from the MSHA Mine field, number of employees, which was then multiplied by the field, days per week, for each record. An aggregation was also calculated for Mine days lost, job experience, total experience, and mine experience. Days lost is a summation of all days lost for a mine, as specified in the Accidents data set, in a given calendar year. The other aggregation fields were averages of the Accidents reported for the year. Other calculated training fields, based on the violations data, include summation of the dollar amount due for violations in a calendar year and the average amount due per violation. See

Table 2: Calculated Fields Used in Gernand’s Work for more information on the necessary calculations.

Independent Variables used in Reproduction of Gernand's Work					
Table	**Feature or Field**	**Minimum Value**	**Maximum Value**	**Mean Value**	**Feature Importance**
Mines	MINE_ID	-	-	-	0.1457
Mines	STATE_CODE	1.00	69.00	26.61	0.0567
Mines	NO_EMPLOYEES	0.00	3,663.00	47.06	0.0961
Mines	HOURS_PER_SHIFT	0.00	24.00	6.27	0.0201
Mines	AVG_MINE_HEIGHT	0.00	9,998.00	16.12	0.1068
Mines	MILES_FROM_OFFICE	0.00	600.00	92.53	0.0894
Mines Calculated	WORKER_DAYS_PER_WEEK	0.00	25,641.00	294.06	0.1322
Multiple	CAL_YR	0.00	2,020.00	1,739.77	0.0453
Accidents Calculated	AVG_TOTAL_EXPERIENCE	0.00	65.00	8.69	0.0776
Accidents Calculated	AVG_JOB_EXPERIENCE	0.00	65.00	6.41	0.0612
Accidents Calculated	AVG_MINE_EXPERIENCE	0.00	57.00	5.70	0.0645
Violations Calculated	SUM_AMOUNT_DUE	0.00	12,100,513.00	14,746.55	0.0649
Violations Calculated	AVG_AMOUNT_DUE_PER_ VIOLATION	0.00	57,853.07	266.86	0.0395

Table 1: Features in Gernand's Work

Calculated Fields in Reproduction of Gernand's Work	
Field	**Calculation**
WORKER_DAYS_PER_WEEK	NO_EMPLOYEES * DAYS_PER_WEEK
SUM_OF_DAYS_LOST	Summation of DAYS_LOST in Accidents Table by MINE_ID and CAL_YR
AVG_JOB_EXPERIENCE	Mean of JOB_EXPER in Accidents Table by MINE_ID and CAL_YR
AVG_TOTAL_EXPERIENCE	Mean of TOT_EXPER in Accidents Table by MINE_ID and CAL_YR
AVG_MINE_EXPERIENCE	Mean of MINE_EXPER in Accidents Table by MINE_ID and CAL_YR
SUM_AMOUNT_DUE	Summation of AMOUNT_DUE in Violations Table by MINE_ID and CAL_YR
AVG_AMOUNT_DUE_PER_VIOLATION	Mean of AMOUNT_DUE in Violations Table by MINE_ID and CAL_YR
DAYS_LOST_PER_100,000_WORKER-DAYS	SUM_OF_DAYS_LOST/ ((WORKER_DAYS_PER_WEEK/7) * 100,000)

Table 2: Calculated Fields Used in Gernand's Work

The dependent variable is also a calculated field in Gernand's work. Days lost per 100,000 worker days is the sum of days lost, divided by worker days. Worker days is the worker days per week divided by seven and then multiplied by 100,000. The random forest implementation for this reproduction is scikit-learn's Random Forest Regressor. The selected hyperparameters for the forest include 1,000 estimators or trees, use of mean squared error (MSE) criterion, a minimum of five leaf samples, bootstrapping on, and out of bag score on. Figure *1*. The resulting R2 was .9502 and the out of bag score was .9320. See Table 3: Results of Reproducing Gernand's Work for more details. The most important features found in the reproduction model were MSHA mine ID, worker days per week, average mine height, number of employees, and miles from office, as shown in Figure 1: Reproduction of Gernand's RF Feature Importance. Each of these variables, except mine ID, appear in the top eleven variables Gernand discusses. A Random Forest Classifier model was also constructed as a comparison point.

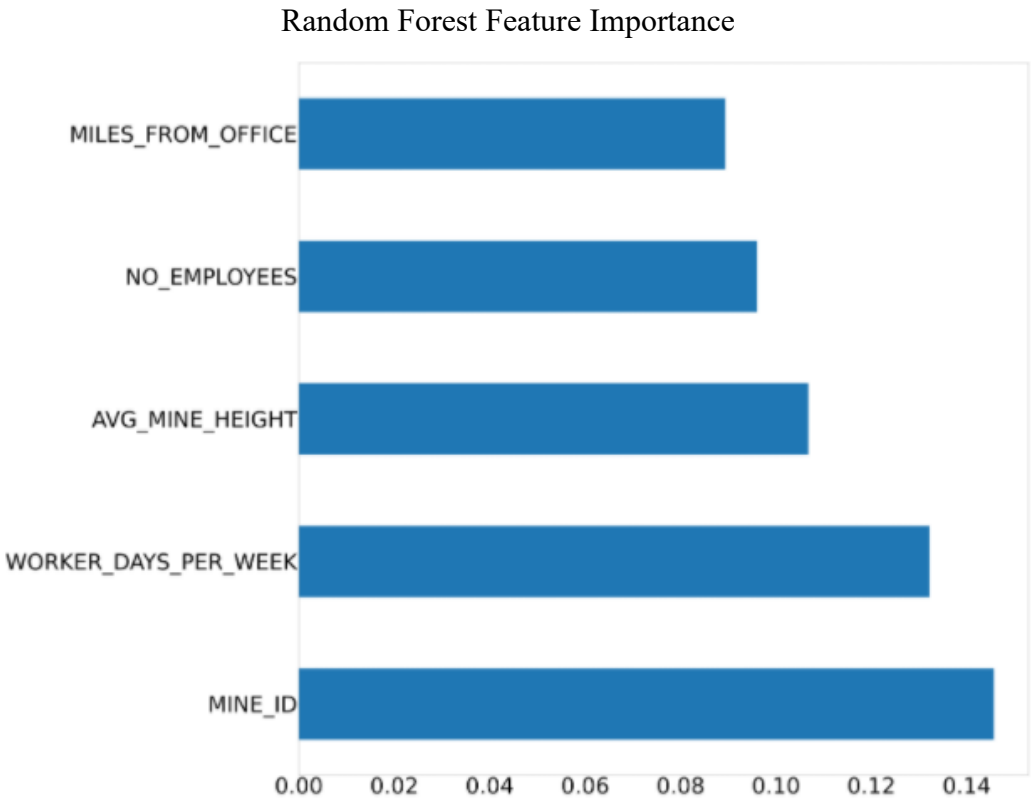

Figure 1: Reproduction of Gernand's RF Feature Importance

Results of Reproducing Gernand's Work					
Random Forest Type	**Record Count**	**R^2**	**Out of Bag Score**	**Weighted Recall**	**Weighted Precision**
Random Forest Regressor	1,008,368	0.9502	0.9321	NA	NA
Random Forest Classifier	1,008,368	0.9988	0.9972	0.9988	0.9988

Table 3: Results of Reproducing Gernand's Work

Chapter 3: Methodology

3.1 Overview

The machine learning technique selected for this thesis was the Random Forest (RF) algorithm. RF was selected because it produces robust and explainable results on structured data. The scikit-learn implementation of Random Forests was used to generate the RF model.

Interactive Jupyter Notebooks were the primary tool to create the predictive model. Jupyter Notebooks are an open-source web application that one can use to create and share documents that contain live code, equations, visualizations, and text. We chose to use Jupyter Notebook because it has become ubiquitous among data scientists. Moreover, the notebooks can be easily shared with research communities to reproduce or replicate the thesis findings.

Several Python libraries were used, including Pandas and scikit-learn. Pandas is used for data cleaning analysis. It is the best tool for handling real-world messy data; the MSHA data requires a lot of cleaning. Scikit-learn is the most commonly used library for machine learning in Python. It provides efficient tools for predictive data analysis. The scikit-learn library was used to implement the predictive models for the thesis.

The data was directly downloaded from MSHA databases and loaded into a Juptyer Notebook. Next, the data was processed using Pandas to make it suitable for the machine learner to build robust predictive models. Then a machine learner was created using a scikit-learn Random Forest Classifier.

Three different scikit-learn Random Forest Classifier models were created to attempt to predict future mining accidents from MSHA violation data. Different models were created to explore the impact of different features, the impact of predicting a category versus an exact number, and the different trade-offs between each model and improve real-world usability. The

primary goal is to create a model mining stakeholders can use, which means finding the best model with the greatest reduction in false negatives and false positives. Table 4: Overview of Models Presented in Thesis lists an overview of the models and how they differ from each other.

- The first model, **Model A**, uses almost the entirety of the accident record and attempts to predict the exact schedule charge aggregation for thirty-five days from the violation.
- The second model, **Model B**, removes some of the datetime features from the accident column and attempts to predict the exact schedule charge for thirty-five days from the accident. The thirty-five day charge is the aggregation of the MSHA value SCHEDULE_CHARGE as found on the Accidents table. This was chosen as the value to predict because it indicates accident severity. The goal is to use a given violation to attempt to predict what the thirty-five day aggregated SCHEDULE_CHARGE will be for that violation.
- **Model C** uses the features without datetimes and simplifies the schedule charge to be different classes or categories of charges., i.e., No Accident (SCHEDULE_CHARGE of zero), Non-fatal Accident (SCHEDULE_CHARGE of less than 6,000), and Fatal Accident (SCHEDULE_CHARGE of greater than 6,000). This categorization led to more interpretable results. Model C also has three parts – using unchanged training samples (C.1), increasing the sparse class of accidents in the training sample (C.2), and sample minimizing with weighed samples (C.3). *The unchanged training sample version* uses the original training set as-is in the other models. *The version that increases the sparse class* uses the training set along with two subsets. **Subset F** is all fatal samples from the training set. **Subset A** is all non-fatal accident samples from the training set. These sets are then combined as the **original set** plus **Subset F** twenty times plus **Subset A** ten

times as a modified training set. *The sample minimizing with weights version* uses the above modified training set, but only uses 5% of the samples. The random sampling selects based on weight, with significant and substantial violations as the higher weight or higher priority selecting criteria.

Model	Features	Predicting	Training Set	Is Boosted	Hyperparameters
A	Violations	Exact 35-day schedule charge aggregation from Accidents	Full Set	No	Estimators: 5 Leaf Samples 3 Max Features: .5
B	Violations without datetimes	Exact 35-day schedule charge aggregation from Accidents	Full Set	No	
C.1 Unchanged Features	Violations without datetimes	Category of 35-day schedule charge from Accidents	Full Set	No	
C.2 Increasing Sparse Classes	Violations without datetimes	Category of 35-day schedule charge from Accidents	Full Set 20x Fatal 10x Accident	Yes	
C.3 Sample Minimizing with Weights	Violations without datetimes	Category of 35-day schedule charge from Accidents	5% of Set 40x Fatal 10x Accident	Yes	

Table 4: Overview of Models Presented in Thesis

3.2 Technique

The machine learning technique selected for this thesis is the Random Forest algorithm. It was selected because the MSHA dataset is structured data, many safety outcomes are known within the MSHA data, RF models are robust, and RF models produces explainable results.

The MSHA dataset is structured data because it is tabular data that was stored in a relational database. It has a predefined data model, and each piece of information is categorized. For example, a mine's record states if it is an UNDERGROUND or ABOVE GROUND mine. An example of unstructured data would be text or an image. Individual parts on text or an image are not categorized. A RF technique may be applied to structured data.

Additionally, the MSHA dataset is a good candidate for a supervised learning algorithm, such as RF, because and the outcomes of safety accidents are known. This means the algorithm can use known results when constructing the model.

RF is also an ensemble algorithm, so it aggregates multiple estimators to form the best prediction. The individual estimators or individual decision trees used here are Classification and Regression Trees (CART) trees. Each CART tree produces a prediction. The resultant prediction of each of the CART trees is then aggregated according to the RF algorithm into a final result. Ensemble algorithms tend to produce robust results by having dissimilar estimators that average out predictive shortcomings and produce a stable overall model. For example, intuitively, an individual CART tree may be ideally suited to predicting UNDERGROUND accidents, another may be best suited to ABOVE GROUND accidents, and a third tree that has acceptable accuracy at predicting both types of accidents. By aggregating the results of all three trees, a more accurate result should be achieved when predicting both types of accidents. In practice, the trees are not divided this way, but the example illustrates how the results are more stable overall.

The most important reason the RF algorithm was selected is that the model produces explainable results, which is necessary for safety-critical decisions. Each decision tree in a RF shows where splits occur and how a prediction arrived at a given conclusion. The aggregating or final result selection process is also clear. This is useful for subject matter experts to use when evaluating a prediction. This algorithm provides transparent results that decision-makers can leverage. Each node of the corresponding tree shows where the tree split and how the predicted value was achieved. This is beneficial to stakeholders using the model because they can see the reasoning and identify potential flaws. Non-transparent algorithms are not suitable for safety decision making. See 3.2.2 Interpreting a Random Forest for more information.

3.2.1 Random Forest

A Random Forest model is a machine learning algorithm that uses multiple decision trees to make a prediction. Random Forests were introduced by Leo Breiman in his paper *Random Forests*. Breiman describes creating multiple decision trees that use different features to create each tree and then each tree votes for a specific prediction [22, pp. 5-6]. Ensemble techniques as a whole combine multiple estimators into one prediction. The idea behind an ensemble technique is that any given estimator may have a weakness, but when the different individual estimators are combined, then that that weakness is reduced because each estimator works differently. Estimators should also be created to be dissimilar from one another.

The scikit-learn implementation of Random Forests, which is used in this thesis, randomizes the sample from the training set and how many features or columns are used to create each individual tree [23]. It uses, “an optimized version of the CART algorithm” [24], for the individual decision trees that make up the Random Forest. The documentation for Random Forests goes on to say:

> "The purpose of these two sources of randomness is to decrease the variance of the forest estimator. Indeed, individual decision trees typically exhibit high variance and tend to overfit. The injected randomness in forests yield decision trees with somewhat decoupled prediction errors. By taking an average of those predictions, some errors can cancel out. Random forests achieve a reduced variance by combining diverse trees, sometimes at the cost of a slight increase in bias. In practice the variance reduction is often significant hence yielding an overall better model" [23].

These sources of randomness generate more dissimilar individual tree estimators, which create a more robust overall model. A notable difference in the scikit-learn implementation of the Random Forest algorithm and the original algorithm by Leo Breiman [22] that the documentation mentions is that, "the scikit-learn implementation combines classifiers by averaging their probabilistic prediction, instead of letting each classifier vote for a single class" [23].

3.2.2 Interpreting a Random Forest

This section will show how to create an interpret a basic Random Forest using an example. First, the environment must be setup, as shown later in 3.3 Environment Tools and Libraries, and the data must be processed, as shown later in section 3.4 Data. However, the purpose of this section is to show the basic process of Random Forest generation and interpretation.

For this example, a simplified Random Forest Classifier was created in scikit-learn with the hyperparameters of three estimators, a max tree depth of two, and only using half of the features when deciding a split. The three individual trees that make up this Random Forest Classifier are shown in Figure 2: Example Random Forest Trees. In scikit-learn, each tree is of the type DecisionTreeClassifier [25]. In this example, and as is the default in scikit-learn, each individual

CART tree was created by using Gini impurity as the deciding factor in creating splits [25]. Notice that there are three trees as the hyperparameters specified three estimators and are shallow with only two splits. These trees would not predict well in-practice because they have very limited decision splits, but they are easier to interpret for an example. Table 5 has an example violation record. The goal is to predict if this violation record will lead to No Accident, an Accident, or a Non-fatal Accident. The encoded value refers to how the model categorizes the value and how it is displayed in the trees produced by scikit-learn in Figure 2. All values must be numeric in the model. Using Table 5 as an example, it would proceed as follows through the trees in Figure 2:

CART Tree 1:

- At node SIG_SUB <= 1.5:
    - SIG_SUB in this example is yes, which is encoded as 2. 2<=1.5 is false, so will proceed to node VIOLATOR_VIOLATION_CNT <= 284.5.
- At node VIOLATOR_VIOLATION_CNT <= 284.5
    - VIOLATOR_VIOLATION_CNT in this example is 128. 128 <=283.5 is true, so proceed to left leaf node.
- Tree 1 predicts the record's result as **No Accident**.

CART Tree 2:

- At node SIG_SUB <= 1.5:
    - SIG_SUB in this example is yes, which is encoded as 2. 2<=1.5 is false, so proceed to node VIOLATOR_INSPECTION_DAY_CNT <= 399.5.
- At node VIOLATOR_INSPECTION_DAY_CNT <= 399.5:

- VIOLATOR_INSPECTION_DAY_CNT in this example is 170. 170 <= 399.5 is true, so proceed to left leaf node.

- Tree 2 predicts the record's result as **No Accident**.

CART Tree 3:

- At the node MINE_TYPE <= 2.5:
  - MINE_TYPE in this example is Underground, which is encoded as 3. 3 <=2.5 is false, so proceed to node VIOLATOR_INSPECTION_DAY_CNT <= 444.5.
- At the node VIOLATOR_INSPECTION_DAY_CNT <= 444.5:
  - VIOLATOR_INSPECTION_DAY_CNT in this example is 170. 170 <=444.5 is true, so proceed to the left leaf node.
- Tree 3 predicts **Fatal Accident**.

Mine Type	Likelihood	Significant and Substantial	Violator Violation Count	Violator Inspection Day Count	…
Underground (Encoded as 3)	Reasonably (Encoded as 4)	Y (Encoded as 2)	128	170	

Table 5: Example Violation Record

Tree 1 and Tree 2 predicted **No Accident** and Tree 3 predicted a **Fatal Accident**. The Random Forest would predict **No Accident** because it has the highest prediction probability when averaging each tree's prediction. Also shown in Figure 2 is the prediction probabilities that scikit-learn calculated, which is represented in the value column. This array corresponds to the categories No Accident, Non-fatal Accident, and Fatal Accident. According to the documentation, "the predicted class probability is the fraction of samples of the same class in a leaf" [25].

Table *7* shows each tree's prediction probabilities and the average the Random Forest would use for predicting.

	Tree 1	**Tree 2**	**Tree 3**	**Random Forest**
No Accident	69.64 %	70.23 %	34.03 %	57.97 %
Accident	13.92 %	13.64 %	30.34 %	19.30 %
Fatal Accident	16.42 %	16.11 %	35.62 %	22.72 %

Table 6: Example Scikit-learn Calculated Prediction Probabilities

Example Random Forest

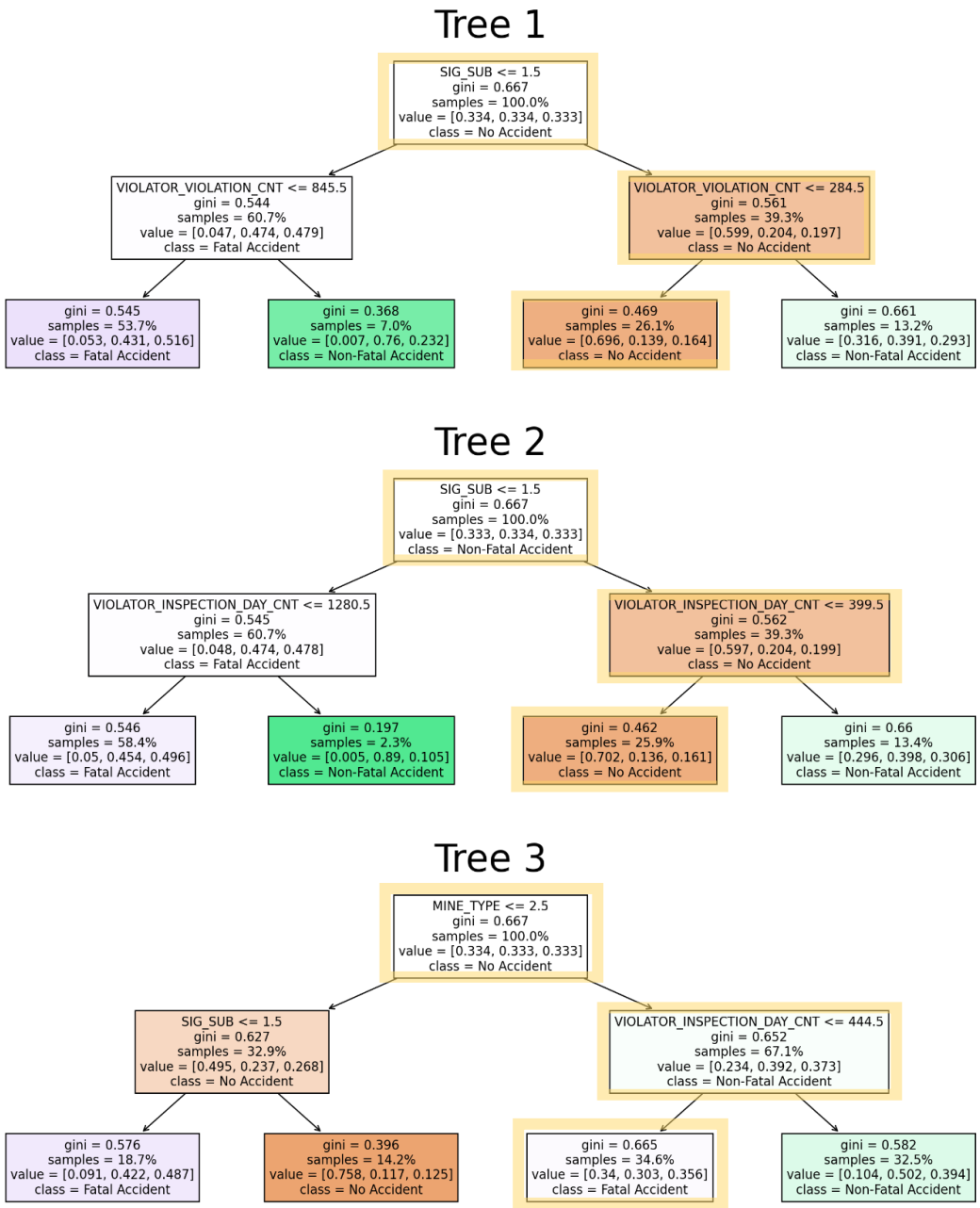

Figure 2: Example Random Forest Trees

This Random Forest was generated using scikit-learn. The RandomForestClassifier's hyperparameters were set to create three estimators, a depth of two, and only consider half of the features for each split. The number of accident samples was increased in the training sample. The yellow boxes highlight the path of the example violation in Table 5.

3.3 Environment Tools and Libraries

The environment used to create the machine learning model includes:

- Interactive Python (IPython) [26]
- Visual Studio Code [27]
- Anaconda Python Environment Management (conda) [28]
- Numerical Python (numpy) [29]
- Python Data Analysis Library (pandas) [30]
- Numerical Python (numpy) [29]
- Scikit Learn (scikit-learn) [31]
- Matplot Library (matplotlib) [32]
- FastAI Library v. 0.7 [33]

Interactive Python (IPython)

The primary software environment for creating the model was a Python interactive notebook. We selected Python due to its focus on simplicity, readability, and large volumes of readily available analytics and machine learning libraries, much needed for building our prediction models. An interactive environment was selected due to the ease of seeing the results of a given line of code immediately. An interactive Python environment has cells for each segment of related code. Selecting the run button will run a given cell's code. After running a cell, what was executed remains in the memory of the notebook if the notebook is live and the notebook server is running. For example, for this project, imports were the first cell of a notebook with no other code. After running this cell, the imports are now available in all subsequently ran cells. This code cell segment format allows for a more iterative approach through ease of access to variables and flexibility.

Visual Studio Code

Visual Studio Code (VS Code) was the primary Integrated Development Environment (IDE) used for this project. VS Code can create and run Interactive Python notebook files and can be used to control the Python environment [27].

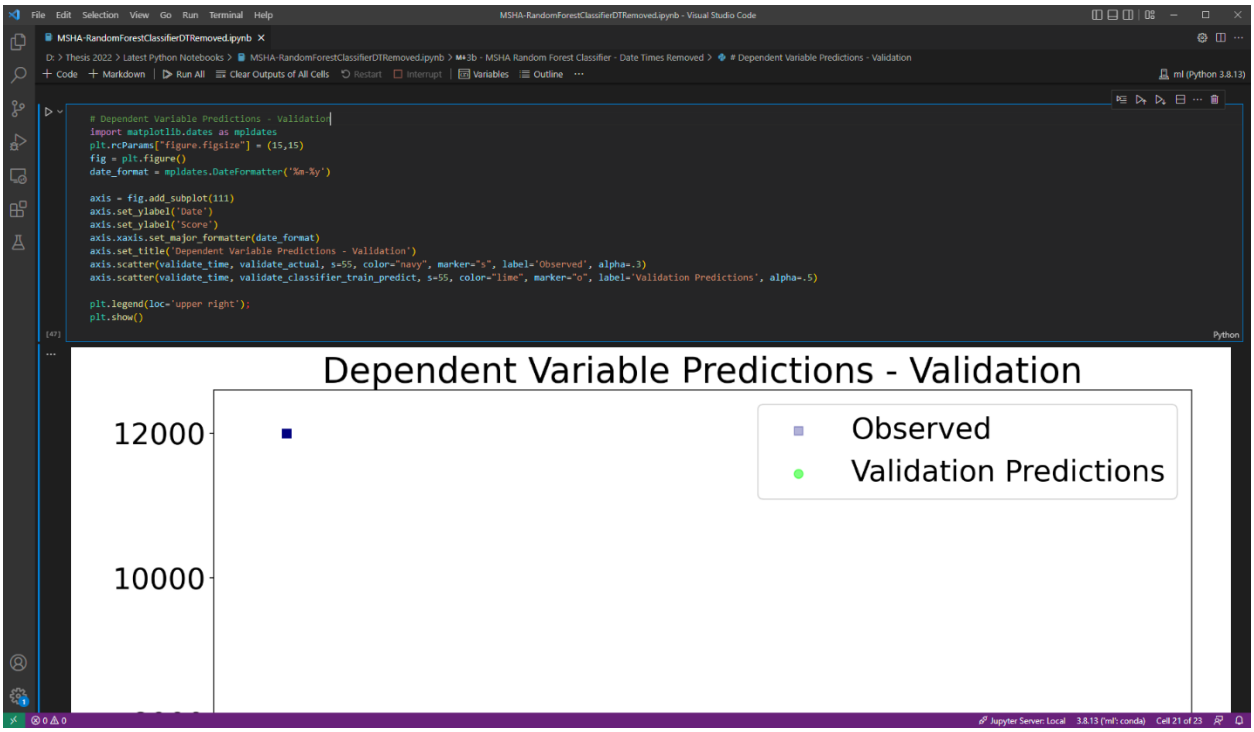

Figure 3: Example Cell Output in Visual Studio Interactive Python Notebook

Figure 3 shows a sample cell in an Interactive Python notebook running in Visual Studio code. The Python code is in a cell and is executed by activating the play button. In this example, the code is setting up a mathplotlib scatter plot. The code for the scatterplot is contained in one cell for ease of reuse and experimenting. After activating the play button, the code and the output will appear below, in this example, the output is the scatterplot below the cell. Other items to notice are the conda environment, in this case, ml (Python 3.8.13), and that it is a local Jupyter Server. Remote servers can also be used to execute the code. Variables, such as validate actual,

were specified earlier in the notebook, once a cell runs, then that cell's context is available to the other cells in the file.

Anaconda Python Environment Management (Conda)

Conda is a Python version and library management environment. Different Python environments are needed for different projects. Conda provides a toolset to organize library dependencies and control Python versions. Organizing library dependencies properly is important because it prevents different library requirements from interfering with each other. For example, if one project requires a certain version of the pandas library and another a different version, then a proper conda environment will prevent dependency conflicts. It also allows management of different versions of Python so that the same system can have one environment setup for Python 3.5 and another for Python 3.6, for example [28].

Numerical Python (numpy)

Numerical Python or NumPy is a Python library that optimizes arrays in Python and is a dependency of the pandas library and can be used with the Matplotlib package. It is important for arrays to be highly efficient due to the quantity of data and many operations required for this project [29].

Python Data Analysis Library (Pandas)

One Python package instrumental to creating this model is the Python Data Analysis Library (Pandas). Pandas is a powerful tool for processing and organizing data in Python. Dataframes are the central pandas data structure used in this project. A pandas dataframe, as stated in the documentation, is a, "two-dimensional, size-mutable, potentially heterogeneous tabular data ... [that] can be thought of as a dict-like container for Series objects" [34] A dataframe has columns and rows of data. Columns can be an incrementing integer index or use specified textual names.

Pandas dataframes have many features that makes data processing easier. Manipulating a column as a series of data can be accomplished with user-friendly array-like accessing. For example, the following syntax, dataframe['Column to Access'] = dataframe['Column to Access'] * 2, takes the column named 'Column to Access' and sets that column value to 'Column to Access' times two, row-wise. Dataframes also have functionality to sort, remove data, set datatypes, merge, or join two dataframes, and have built-in aggregation functions. It is also straightforward to initialize a dataframe from Comma Separated Values (CSV) files or Excel files, through the read_csv function. This is a powerful tool for cleaning up and reorganizing data.

Scikit-Learn

Scikit-learn is another central Python library used in generating the model. This machine learning package contains implementations of many machine learning algorithms, graphing functions, and data cleanup routines. The Random Forest classes are utilized to build the model. This will be discussed further when describing the model building process.

Matplot Library (matplotlib)

Matplotlib is the graphing utility for this project. Both pandas and scikit-learn leverage this library to create detailed graphs, charts, and diagrams. It is also used in its freestanding form when more advanced graphing features are needed. Notably for creation of a confusion matrix and scatterplots.

FastAI Library

FastAI is a key Python utility used in this project. This utility provides additional tools to prepare data for model training and enhances scikit-learn. For example, it has a built-in function to prepare categorical data and format it into numeric, which is a prerequisite for model building. An earlier version, 0.7, of fastai was used.

3.4 Data

MSHA is tasked with preventing mining accidents through educating miners, mine inspections, and establishing criteria for safe mining conditions. MSHA collects a wealth of data on U.S. mines through mine permits and inspection reports. “Every underground coal mine in the U.S. is inspected by an MSHA inspector on a quarterly basis to check whether a mine is complying with the mandatory health and safety standards. During an inspection, the MSHA inspector(s) issue citation(s) for mandatory health and safety standards that are violated and determine the degree of seriousness based on tabled criteria and their judgement” [16, p. 1016]. This data is then published by MSHA as reports, web tools, and raw collections.

The data used for this project was the raw flat files that MSHA exports and publishes from a relational database [35]. In this MSHA data collection, there are twenty sets of flat files in total [35].

MSHA Datasets

MSHA Dataset Name	**Description**
Accident Injuries	Accidents reported on MSHA form 7000-1
Area Samples	Physical samples
Coal Dust Samples	Operator and inspector dust samples
Conferences	Mediation on violations issued
Civil Penalty Dockets and Decisions	Petitions for changing violation penalties
Contractor Employment Production Quarterly	Coal production as reported by contractors
Contractor Employment Production Yearly	Coal production as reported by contractors

Controller Operator History	Shows historical changes in Controllers at a given mine
Employment Production Yearly	Coal production as reported by operators
Employment Production Quarterly	Coal production as reported by operators
Inspections	MSHA inspection listing
Mine Address of Record	Legal address of mine
Mines	Listing of all Mines since 1970
Noise Samples	Physical noise samples
Personal Health Samples	Physical samples
Quartz Samples	Quartz samples
Violations	Violations issued by inspectors on MSHA from 7000-3
Contested Violations	Violations protested by the operator
Assessed Violations	Violations that have been assessed penalties
107(a) Orders	Inspector ordered immediate withdrawal from mine

Table 7: MSHA Publicly Available Data Sets

The three datasets used in the machine learning model and of focus in this thesis are the Mines, Violations, and Accidents datasets. These datasets were selected because they directly relate to safety or contain detailed information on the mine. Factors contributing to the selection of these datasets over the multitude of choices are the relative importance of the information, and clear relation to the primary goal of predicting potential hazards.

Other datasets provided by MSHA do have valuable information but are more tangential to this goal. They would be valuable for our future work since they contain rich environmental samples, such as coal dust, coal production and other economic factors— these could also be of interest because they might provide insight into how operations change when prices are high or low, and total coal production— this may also provide insight into the size or capacity of the mine.

The **Mines** dataset is the first important data set used in the thesis for constructing the machine learning model. This MSHA dataset has general information on U.S. mines; it contains information such as MSHA ID, mine type, address, status of mine, number of employees, shifts, mine height, and mining methods used. This file contains biographical information on the mine [36]. The MSHA ID uniquely identifies a mine and is the primary key in the MSHA database; it is utilized by other datasets when they reference a mine [36]. Establishing a list of all U.S. mines is a critical first step upon which the rest of the datasets build upon.

Another important dataset chosen for training the model is the **Violations** file because it contains information on a mine's policy violations or potential dangers. A violation is one of the first recorded indications that something could be amiss from a safety standpoint within the mine. The Violations dataset is amassed from MSHA Form 7000-3 [35]. This form is for MSHA inspectors to issue citations to mines. The most notable fields in this dataset include what part of the mining section was violated, whether the violation was significant and substantial, the likelihood of occurrence, seriousness of injury, violation fee, and number of miners impacted [37].

One significant aspect of MSHA Form 7000-3 is the inspector's evaluation of the gravity of the specific violation. MSHA, in a citation handbook, defines gravity as,

> "… an evaluation of the seriousness of the violation. It is determined by the three factors listed in § 100.3 (e) (Determination of penalty amount; regular assessment). Section 100.3 states: 'Gravity is determined by the likelihood of the occurrence of the event against which a standard is directed; the severity of the illness or injury if the event has occurred or was to occur; and the number of persons potentially affected if the event has occurred or were to occur'" [38, p. 10].

The section the handbook refers to is The Federal Mine Safety and Health Act of 1977 (Mine Act) [38, p. 1]. The MSHA inspector's assessment of incident gravity is a qualitative determination of the violation that occurred. This qualitative determination is then recorded into a categorical quantitative metric. The gravity of an incident is beneficial when training the model because it naturally signals the importance of a violation.

The likelihood of a violation resulting in an accident is also judged by the inspector before being included in the violation dataset. The set of values for this field, i.e.,, Violation Likelihood, are No Likelihood, Unlikely, Reasonably Likely, Highly Likely, and Occurred. The Occurred designation is significant because it, "… can only be checked when an injury or illness has actually occurred" [38, p. 11]. The handbook provides special guidance for Occurred injuries. It provides an example that if an accident is worse than expected, the evaluation should be evaluated on what happened. The handbook provides a further example, where if an incident is better than typically expected, then the instance should be evaluated on the typical, worst case scenario [38, p. 12]. The gist of the handbook is that a conservative worst-case evaluation should always be selected in the case of a violation that resulted in an injury directly. Similar to gravity measurements, likelihood also can act as a signal how severe a violation was.

The third important data set is **Accidents and Injuries** dataset. The data set is composed from the information collected from MSHA Form 7000-1 [35]. When an accident occurs, the operator is required to fill out this form explaining what occurred to cause the accident or injury and how severe the incident was [39, pp. 28-35]. MSHA, for this form, defines an accident as meeting one of twelve criteria [39, pp. 3-4]. The criteria include deaths, near-fatal injuries, entrapments greater than thirty minutes, unexpected gas or liquid entry, gas or dust explosions, unplanned fires greater than thirty minutes, unplanned explosion, unplanned roof fall, rock outburst that disrupts work greater than an hour, unstable refuse pile, hoisting equipment damage disrupting work greater than thirty minutes, and deaths or injuries that occur due to the mine [39, pp. 3-4]. A few important fields include, degree of injury, mining method, experience, days lost, and scheduled charge [40]. "The MSHA inspectors also perform field checks on accuracy and for possible underreporting of accidents" [16, p. 1016].

One central field in the Accidents dataset is scheduled charges, which was selected as the dependent variable or value predicted in the machine learning model. This is the field that the model will attempt to predict. Scheduled charges, as stated in PC7014 Report on 30 CFR Part 50, "… are based on an estimate of the future loss of productive time brought about by an employee's permanent loss of a body member or permanent impairment of function. This measurement highlights the more serious injuries occurring in the mining industry" [39, p. 2]. "The scheduled charge for fatalities and permanent total disabilities is 6,000 days. Charges are assigned to determine the relative severity of certain injuries regardless of the actual days lost" [39, p. 17]. In a consistent and detailed manner, this field captures the severity of an accident. The table below from MSHA PC-7014 details how various injuries are categorized. For example,

an amputation of the hand is assigned a scheduled charge of 3,000 and an amputation of the foot is assigned a schedule charge of 2,400. See Table 9 for full list of scheduled charges by MSHA.

MSHA inspectors assign a schedule charge based on the MSHA PC-7014 table. This is a measure of severity. Table 7 shows a mean calculation using Pandas on the MSHA field "INJURY_SOURCE" for the Accidents MSHA table. This schedule charge is the mean schedule charge for all Accidents in the specified category. Of note is that the methane category injury source carries a mean charge of 6,000, which is a fatal schedule charge. Table 8 shows a mean calculation using Pandas on the MSHA field "NATURE_INJURY" for the Accidents MSHA table. Both tables were abbreviated to only include the mean schedule charges greater than one hundred. The mean schedule charge for each category of injury source and nature of injury helps illustrate how the scores scale depending on the Accident type. For example, an accident in a critical injury category, such as a heart attack, have a high mean schedule charge of 3,965.73, whereas more chronic conditions, such as black lung have a low mean schedule charge of 128.47. This shows that individual Accidents are weighed proportionally to the severity of the type and nature of injury. During model training, this scaling of severity of Accidents is an important criterion to ensure proper weighing.

MSHA Categorized Injury Source	Mean Schedule Charge
METHANE GAS-IN MNE & PROC	6000.00
OXYGEN DEFICIENT ATMOSPHR	3000.00
VEHICLES,NEC	2000.00
LANDSLIDE (SURF ONLY)	1500.00
WATER	1163.93
MISCELLANEOUS,NEC	921.30
TRANSFORMERS,CONVERTERS	807.69
NOXIOUS MINE GASES,NEC	418.60
NARO G RAIL CR,MTR-UG EQP	381.59
STD G RAIL CR,MTR-SURF EQ	362.26
FLAME,FIRE,SMOKE,NEC	310.71
KILNS,MELT FURNACE,RETORT	305.00
UNDERGROUND,NEC	260.87
ELEVATORS,CAGES,SKIPS,ETC	244.22
SAND,GRAVEL,SHELL	228.81
PASS CARS,PICKUP TRUCKS	224.79
PLANTS,TREES,VEGETATION	221.67
CRANES,DERRICKS	204.34
CONVEYORS,NEC	161.22
LONGWALL CONVEYOR	160.00
STREET,ROAD	155.84
HGHWY ORE CARIER,LRGE TRK	152.47
STEAM	150.00
GENERATORS	148.21
CAVING ROCK,COAL,ORE,WSTE	133.15
MINE JEEP,KERSEY,JITNEY	126.88
BELT CONVEYORS	121.20
RDIATNG SUBST OF EQIP,NEC	120.00
EXPLOSIVE-DIR REL TO INJR	119.77
UNDERGRD MINING MACHINES	118.37
ELECTRICAL APPARATUS,NEC	117.39
SURFACE MINING MACHINES	117.38
HOISTING APPARATUS,NEC	109.07
MACHINE-MILL,CLEANING PLT	105.63
LONGWALL SUPT,JKS & CHOCK	102.95
STORAGE TANKS AND BINS	101.92

Table 7: Mean Schedule Charge by Source

MSHA Categorized Nature of Injury	Mean Schedule Charge
HEART ATTACK	3965.73
SUFFOC,SMOK INHILAT,DROWN	3450.00
CEREBRAL HEMORAGE-NT CCUS	2238.46
ELECT SHOCK,ELECTROCUTION	1063.55
MULTIPLE INJURIES	963.44
CRUSHING	816.17
ASBESTOSIS	600.00
AMPUTATION OR ENUCLEATION	498.02
OTHER PNEUMOCONIOSIS,NEC	230.77
OTHER INJURY,NEC	173.07
PNEUMOCONIOSIS,BLACK LUNG	128.47
POISONING,SYSTEMIC	123.38

Table 8: Mean Schedule Charge by Nature of Injury

The schedule charge of an Accident will be aggregated to form the variable the model will predict. Schedule charge was selected over the other candidate variable of severity because the charge levels are predetermined, whereas the operator or inspector judge severity qualitatively with fields such as LIKILYHOOD. An aggregation of the ‘schedule charge’ over thirty-five days of accidents was selected because one accident could have multiple heralding violations and was used in a prior work [7]. The larger timespan also gave some lead time between violation and accident for events to occur.

Table of Scheduled Charges in Days

A. For Loss of Member – Traumatic or Surgical

Amputation involving all or part of bone		Thumb	Fingers			
			Index	Middle	Ring	Little
Distal phalange	---	300	100	75	60	50
Middle phalange	---	---	200	150	120	100
Proximal phalange	---	600	400	300	240	200
Metacarpal	---	900	600	500	450	400
Hand at Wrist	3,000	---	---	---	---	---

Toe, foot, and ankle			
Amputation involving all or parts of bone		Great toe	Each of other toes
Distal phalange	---	150	35
Middle phalange	---	---	75
Proximal phalange	---	300	150
Metatarsal	---	600	350
Foot at ankle	2,400	---	---

Arm	
Any point above elbow, including shoulder joint	4,500
Any point above wrist and at or below elbow	3,600

Leg	
Any point above knee	4,500
Any point above ankle and at or below knee	3,000

B. Impairment of Function

One eye (loss of sight), whether or not there is sight in the other eye	1,800
Both eyes (loss of sight), in one accident	6,000
One ear (complete industrial loss of hearing), whether or not there is hearing in the other ear	600
Both ears (complete industrial loss of hearing), in one accident	3,000
Unrepaired hernia (For repaired hernia, use actual days)	50

Table 9: Table of Scheduled Charges in MSHA PC-7014 Appendix C *[39, p. 23]*

3.5 Data Preparation

The MSHA provided Mines, Violations, and Accidents datasets were first loaded into an interactive python notebook by using the pandas library read-csv function. Read-csv converts the raw CSV into a pandas DataFrame. Required parameters for these datasets are to use latin encoding and pipe(|) delimination. The pipe delimters facilitates to distinguish between the different attributes value in CSV file. The date fields also need to be specified directly for read-csv to parse them correctly too, MSHA labeled these columns as <Column_Name>_DT. Figure 4 shows the MSHA provided Mines table data before processing. Figure *5*: Notebook with Mines Loaded into a Pandas DataFrame shows how the data will appear in Visual Studio after processing. Notice that the live variables may be viewed in the bottom panel. The Mines data used had 88,685 rows and 59 columns. Once in the Pandas DataFrame, the data may be manipulated and cleaned and exported out as another .csv file or loaded as a feather format [41] to easily transport between Notebooks.

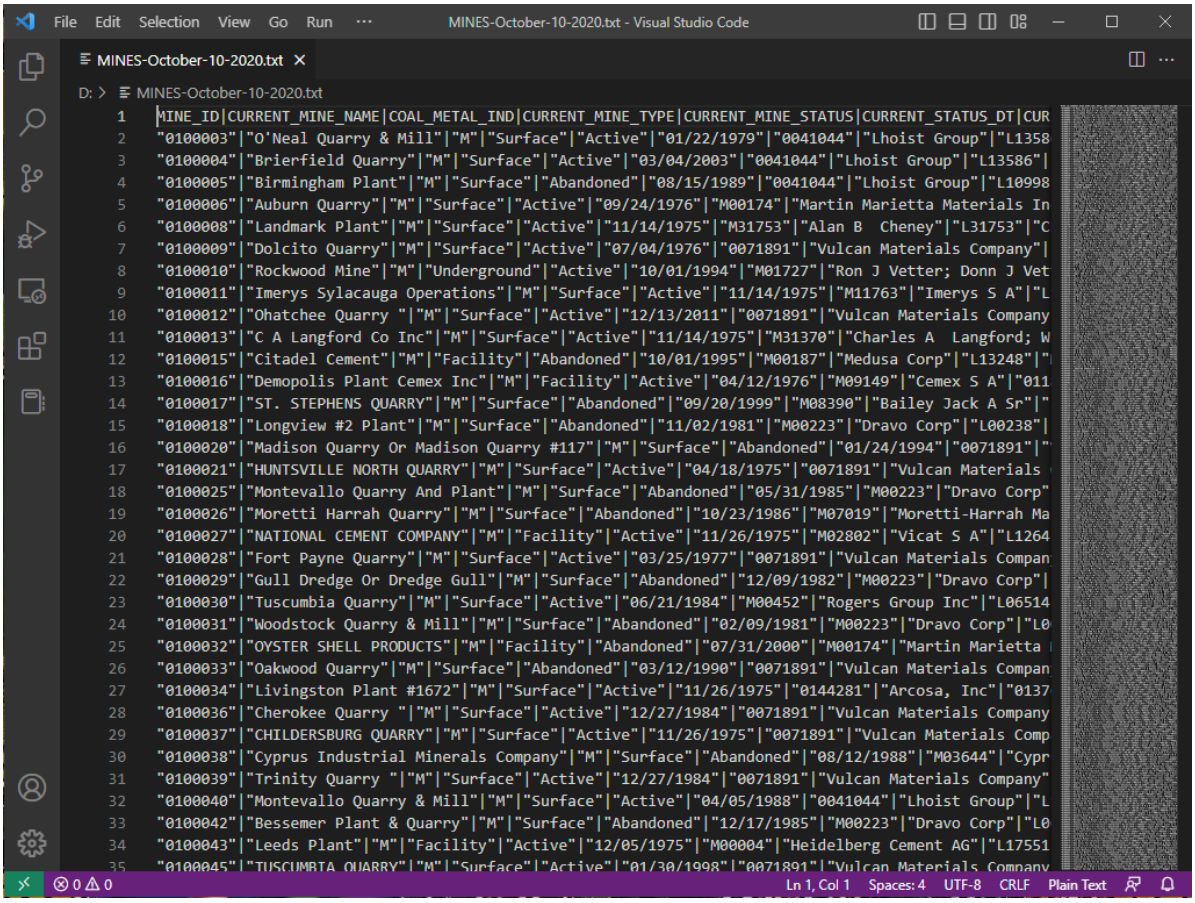

Figure 4: Unprocessed Mines Pipe Deliminated Data

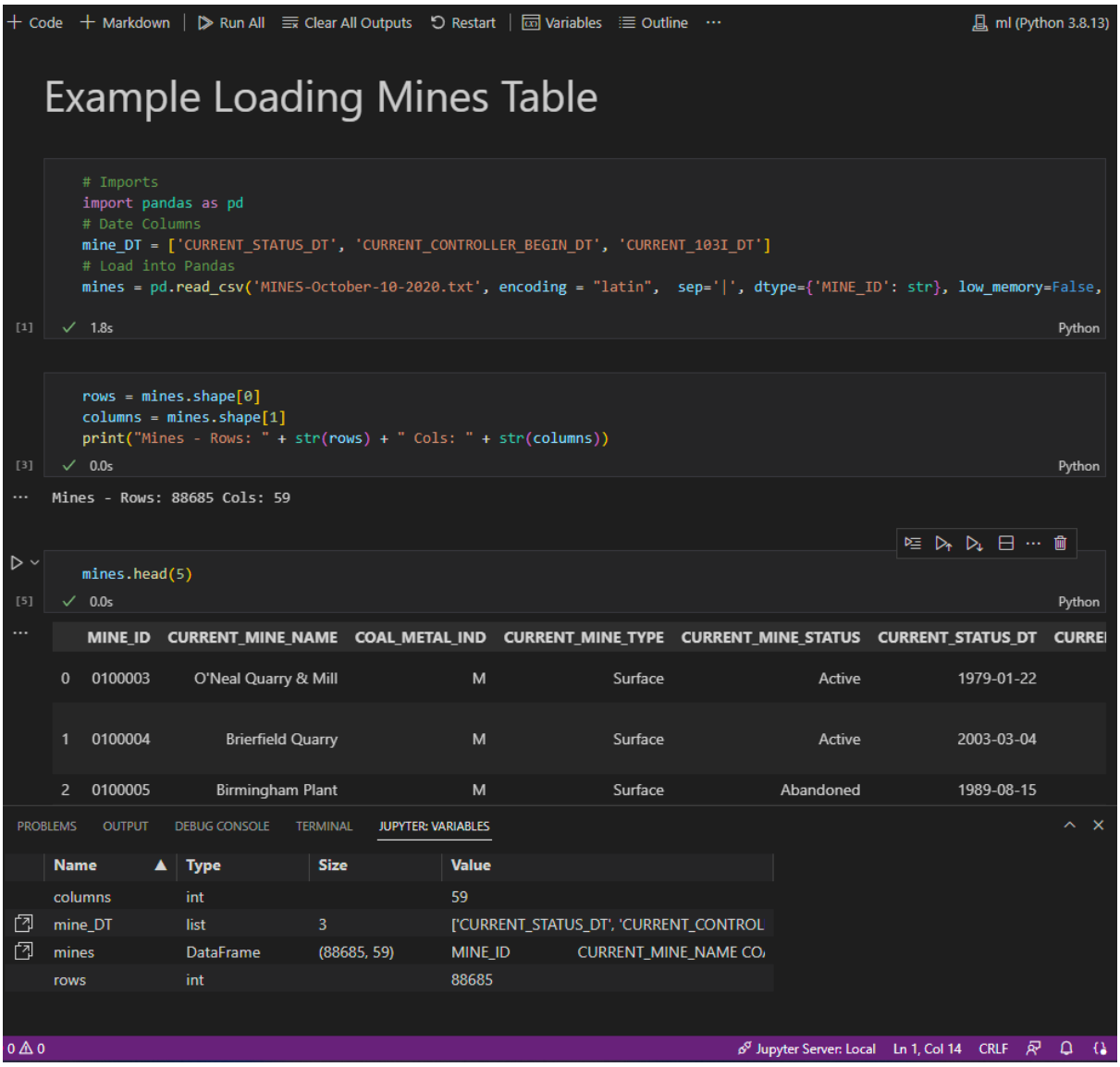

Figure 5: Notebook with Mines Loaded into a Pandas DataFrame

Once each dataset is loaded into a DataFrame, extra calculations are made as part of pre-processing. This is discussed further below. The operations performed on the data include:

- Calculating the SCHEDULE_CHARGE_SUM_35_DAYS for the model to predict.
- Dividing the data into a training, validation, and test set.
- Removing redundant or unimportant columns of data

- Processing all non-numeric data types to numeric forms the model can use. This includes changing fields such as dates and categorical data.
- Identifying missing or null data.

The field SCHEDULE_CHARGE_SUM_35_DAYS is what the model will attempt to predict; however, this field does not exist directly in the dataset. It is calculated as a derived attribute by taking a given violation and summing all accident SCHEDULE_CHARGES that occur within thirty-five days of that violation. We used this attribute as our prediction class, i.e., what the model will try to predict. The idea is that the model will be given a given violation and it will attempt to predict the likelihood if the violation will lead to an accident within thirty-five days. Besides the prediction class, no other attributes from the Accident dataset are used in constructing the prediction model. Procedure for calculating the SCHEDULE_CHARGE_SUM_35_DAYS field:

For each MSHA **violation** record:

- Select the **violation's** MSHA MINE_ID and VIOLATION_OCCUR_DT fields.
- Calculate thirty-five days from the VIOLATION_OCCUR_DT as the temporary field THIRTY_FIVE_DAYS_FROM_VIOLATION_DT.
- Select MSHA **accident** records that occur between VIOLATION_OCCUR_DT and THIRTY_FIVE_DAYS_FROM_VIOLATION_DT for the given MINE_ID.
    - From the selected **accident** records, select each accidents SCHEDULE_CHARGE and sum together to form SCHEDULE_CHARGE_SUM_35_DAYS.
    - Append SCHEDULE_CHARGE_SUM_35_DAYS to the **violation** record.

An example using this procedure may be found in Figure 6: Example SCHEDULE_CHARGE_SUM_35_DAYS Calculation. For Model C, the field SCHEDULE_CHARGE_SUM_35_DAYS was also simplified into three classes of No-Accident (score of zero), Non-fatal Accidents (score greater than zero, but less than 6,000), and Fatal Accidents (scores greater than 6,000).

The data is also divided into a training set, a validation set, and a test set based on the field INSPECTION_BEGIN_DT. The training set is used for model training. The validation set is used to check that the model produced from the training set is behaving as expected. For example, the validation set can be used to determine if the model is overfitting to the training data. The idea is the validation set can be used as a way to benchmark and help drive adjustments to the training model. The final set is the test set, which, for this thesis, is the holdout set. This separate data is used for final evaluation of the model and is not used for calibrating the model. For example, if overfitting is observed in evaluating the validation set, then this might drive different hyperparameters in the model. The test set is only used to evaluate how well the model performs.

The field INSPECTION_BEGIN_DT was used to sort the data and form the different sets based on time. The model is trained on earlier data and attempts to predict later data. The training set was used to train the data and the validation and test set were used to see how well the machine learner made predictions. The dataset was processed this way because it is much easier to predict data around the same time as another sample. This breakdown ensures the model is not predicting a timeframe it has already seen samples for, which makes it more true to its intended use as predictive.

MSHA Data Sets

	Earliest Inspection	*Latest Inspection*	*Samples*	*Percentage of Total*
Train	January 02, 2000	August 07, 2015	2,100,000	81%
Validate	August 07, 2015	August 18, 2016	100,000	4 %
Test (Holdout)	August 18, 2016	October 08, 2020	390,301	15%

Table 10: MSHA Training, Validation, and Test Set Overview

A few redundant columns are also removed, for example, record identifiers or codes that are captured elsewhere as categories. Additionally, the Mines data set is also removed during data preparation. The reason for this is that while the biographical information of the mine is likely predictive, it can also go stale. For example, a violation may have occurred years ago when the mine had less than fifty employees. However, the database only lists current data. So, that mine may have expanded and now has hundreds of employees. The model would then attempt to use incorrect categorization of hundreds of employees in the model, while at the time, it only had less than fifty.

All non-numeric fields must also be converted to data types the model can process. First, non-existent dates (0000-00-00 00:00:00) are converted to NaT (not-a-time) values. Next, additional date parts are added by using the FastAI library on some models. Examples of additional date parts that will be added as columns to the dataset include day of week, day of year, and quarter.

Next categorical data is set to numeric data through using FastAI's train_cats function. This function takes a given category, such as bituminous or non-bituminous coal and converts it to a numerical value such as zero and one. Likewise, ordinals are set as being in the correct increasing order.

Finally, the FastAI helper function ‘proc_df’ is applied on the dataframe to finalize the removal of all non-numeric data and missing values. It also sets a new column on the data frame for missing data. The function also splits the feature that will be predicted, SCHEDULE_CHARGE_SUM_35_DAYS, from the data sets to prepare the data for use in a sci-kit learn Random Forest.

The Random Forest Classifier is selected over a Random Forest Regressor because the schedule charges form a stepwise function. The chart of scheduled charges has a limited number of values, see Table 9 for combinations. For example, a number like 302 for SCHEDULE_CHARGE_SUM_35_DAYS will not occur in the dataset because the schedule charge chart only uses numbers divisible by five. This makes a Random Forest Regressor, not a good choice because the schedule charges essentially act as classes. This is also why one of the models, Model C, is able to easily split into No Accident, Non-fatal and, Fatal categories.

SCHEDULE CHARGE SUM 35 DAYS Example Calculation

MSHA Violation Record

MINE_ID	MINE_TYPE	VIOLATION_OCCUR_DT	PART_SECTION	SECTION_OF_ACT_1	…
1234567	Underground	2003-06-28	75.220(a)(1)	104(a)	

Selecting Relevant Accidents Within Thirty-Five-Days of Violation

MINE_ID	ACCIDENT_DT	SCHEDULE_CHARGE	CLASSIFICATION	…
1234567	2003-07-22	0	POWERED HAULAGE	
1234567	2003-08-02	600	DUST DISEASE OF LUNGS	
1234567	2003-07-20	0	HANDTOOLS (NONPOWERED)	
1234567	2003-07-21	480	DISORDERS (REPEATED TRAUMA)	

Violation Record After Calculation Added

MINE_ID	MINE_TYPE	VIOLATION_OCCUR_DT	SCHEDULE_CHARGE_SUM_35_DAYS	…
1234567	Underground	2003-06-28	1080	

The example violation is 75.220(a)(1), which requires mine operators to have a roof control plan [47]. The accidents that occurred within 35-days of the violation are not necessarily related to roof control. The idea is that some violations are indicative of poor conditions overall. For example, a mine without a roof control plan may be susceptible to accidents of all varieties because such a vital necessity was not properly planned. If a different violation occurred that day, it would also have the same score based on the accidents. With enough records, trends may emerge and violations with unrelated accidents or will have a lower signal than related accidents overall.

Figure 6: Example SCHEDULE_CHARGE_SUM_35_DAYS Calculation

Chapter 4: Analysis and Results

Several Random Forest models were created to explore different possible robust solutions to predict potential safety hazards from MSHA data. The different models highlight the different trade-offs necessary to find a predictive model. For example, Models A and B attempt to predict granularly a specific numerical SCHEDULE_CHARGE_SUM_35_DAYS, whereas Model C attempts to predict a category of No Accident, Non-fatal Accident, or Fatal Accident. This exploration of using different training features, value to predict, and model hyperparameters help produce robust machine learning models.

4.1 Model A: Violation Features with Date Times

Model A is a scikit-learn RandomForestClassifier using five estimators, three samples a leaf, and using half of the features available, without bootstrapping. These hyperparameters were selected to balance speed of model building with accuracy of results. Five estimators, the number of trees used to make up the random forest, were selected so that the model may be constructed within around ten minutes. Predicting the values for validation and test set can occur in an additional ten minutes. Earlier models were constructed using one-hundred estimators or trees and these models took a full day to construct a model. There was a difference in accuracy, with more trees being more accurate. However, there was also a tendency for the Jupyter Notebook kernel to crash and lose all processing work, which requires the code to be re-ran and the model reconstructed. Testing additional estimators and their impact would be key to finalizing the model. The other hyperparameters of using three samples a leaf and half of the features available were selected during early model building based on returning better mean accuracy. Again, final model construction for production use would benefit from a deep investigation to find optimal

hyperparameters. These were selected on the basis of producing generally more accurate results over other hyperparameter options, generally quick execution speed, and overall platform stability.

Model A was trained using most independent variable features present in the MSHA Violations data. A few identification fields were removed from the data such as CONTROLLER_ID because this value is also specified as free text as CONTROLLER_NAME. The full list of features used is presented in Table 11 and Table 12. The dependent variable or predicted variable used is SCHEDULE_CHARGE_SUM_35_DAYS. The possible values for this schedule charge are any value that has appeared in the database before, which generally encompasses MSHA's provided schedule charge table, along with larger values that occur from aggregating the charges. Table 9 has the possible values of SCHEDULE_CHARGE and Figure 6 goes through an example of aggregating these charges. For example, if an accident resulted in loss of the distal phalange on the ring finger (schedule charge of 60) and another separate accident resulted in loss of a distal phalange on a non-great toe (schedule charge of 35) and they occurred within thirty-five days of a violation, then the SCHEDULE_CHARGE_SUM_35_DAYS for that violation (and potentially others) would be 95. The violation and accident are not necessarily causal. The value is more a measure of hazard.

MSHA Violation Features (Independent Variables) Present in Model A, B, C	
AMOUNT_DUE	NO_AFFECTED
AMOUNT_PAID	ORIG_TERM_DUE_TIME
ASMT_GENERATED_IND	PART_SECTION
CAL_QTR	PRIMARY_OR_MILL
CAL_YR	PROPOSED_PENALTY
CIT_ORD_SAFE	REPLACED_BY_ORDER_NO
COAL_METAL_IND_V	SECTION_OF_ACT
CONTESTED_IND	SECTION_OF_ACT_1
CONTRACTOR_ID	SECTION_OF_ACT_2

DOCKET_NO	SIG_SUB
DOCKET_STATUS_CD	SPECIAL_ASSESS
ENFORCEMENT_AREA	TERMINATION_TIME
FISCAL_QTR	TERMINATION_TYPE
FISCAL_YR	VACATE_TIME
INITIAL_VIOL_NO	VIOLATION_ISSUE_TIME
INJ_ILLNESS	VIOLATOR_INSPECTION_DAY_CNT
LAST_ACTION_CD	VIOLATOR_NAME
LATEST_TERM_DUE_TIME	VIOLATOR_TYPE_CD
LIKELIHOOD	VIOLATOR_VIOLATION_CNT
MINE_NAME	WRITTEN_NOTICE
MINE_TYPE	CONTROLLER_NAME
NEGLIGENCE	*Additional NA fields

Table 11: Model A, B, and C Features

Model B and C only use this list of features without datetimes.
See Appendix for MSHA provided field definitions.

* NA fields are used to indicate data was not present before the data cleanup to all numerical values. E.g., data encoded from an empty string (“”) to zero would have a corresponding NA column with an entry of true (value of one) to indicate that the field was originally null and had to be changed to zero. If the original column was zero, then the NA entry would be false (value of zero) because the data field was not coerced into a numerical value.

Datetime Features (Independent Variables) Present Only in Model A	
BILL_PRINT_DT	ORIG_TERM_DUE_DT
CONTESTED_DT	RIGHT_TO_CONF_DT
FINAL_ORDER_ISSUE_DT	TERMINATION_DT
INSPECTION_BEGIN_DT	VACATE_DT
INSPECTION_END_DT	VIOLATION_ISSUE_DT
LAST_ACTION_DT	VIOLATION_OCCUR_DT
LATEST_TERM_DUE_DT	* Additional date and NA fields

Table 12: Model A Additional Datetime Features

* Non-listed date fields are a further breakdown of the “DT” fields, including month end, quarter end, day of week, etc.

The performance and feature importance of Model A was in line with the expectations found in earlier work [7]. The earlier work used a few different techniques to aggregate the data. The

prior work used a SQL database to store and perform the work. Model A's data preparation and model training was entirely created in Python.

The training accuracy of this model is 99.92% and this value is so high that it appears to be an overfit model. However, the validation and test values have an accuracy of 97.16% and 96.73%.

Table 13 has an overview of these results. An overfit model would not result in such high accuracy during validation and testing. An overfit model occurs when the model is so tailored to the training data that it makes decision trees and forests that are only good at predicting the values for the training data itself. Model C explores this issue further with a confusion matrix with limited schedule charge prediction categories later.

As stated in the earlier work, there is concern that a "data leak" is occurring [7]. A "data leak" is when there is a feature that would not be present if the later fact were not already known. For example, in the Mines MSHA datatable, there is a field, CURRENT_103I_DT. (This field was not used in this model; this is only illustrative.) The CURRENT_103I_DT field indicates that:

> "If a mine has experienced an ignition or explosion of methane or other explosive gases that resulted in a fatality or in a permanently disabling injury as defined under 30 C.F.R. § 50.20-6(b)(3)(i) or § 50.20-6(b)(3)(ii) at any time during the previous five years, the mine shall be placed in Section 103(i) status as directed by the Act regardless of total liberation, and a minimum of one Section 103(i) spot inspection of all or part of the mine during every five working days at irregular intervals shall be conducted" [42].

In this example of CURRENT_103I_DT, a mine with a date in this field has already had an accident. If a model were to use this field, it would be much easier to predict historical accidents rather than contemporary accidents. In other words, the model could perform much better in

testing than it would work in application – the "data leak" gives away the answer. No known "data leaks" were present in this model; however, they tend to be challenging to identify.

Model A

	Earliest Inspection	*Latest Inspection*	*Samples*	*Mean Accuracy*	*RMSE*
Train	January 02, 2000	August 07, 2015	2,100,000	0.9992	218.7088
Validate	August 07, 2015	August 18, 2016	100,000	0.9716	510.0146
Test	August 18, 2016	October 08, 2020	390,301	0.9673	513.0333

Table 13: Model A Overview

Model A

	Weighted Avg. Recall	Weighted Avg. Precision	Weighted Avg. F1
Train	0.9992	0.9992	0.9992
Validate	0.9716	0.9572	0.9637
Test	0.9673	0.9582	0.9627

Table 14: Model A Performance Evaluation

Model A's feature importance chart notably has many datetime features. This feature importance chart was created using FastAI library `rf_feat_importance` function which uses the Random Forest Classifiers' `feature_importances_` property. From the documentation, it is calculated using Gini impurity: "The higher, the more important the feature. The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature. It is also known as the Gini importance" [43]. Now, using the many datetime fields could simply be how the decision trees are splitting and not indicative of the feature's predictive power. They also could be introducing a subtle "data leak" too. For example, inspectors could be retroactively updating a violation after an accident. There is no evidence of this, but it illustrates how easily a leak could be introduced for model training. Model B eliminates the datetime fields to gain a better understanding of feature importance. The datetime features were likely used as convenience splits to divide the data.

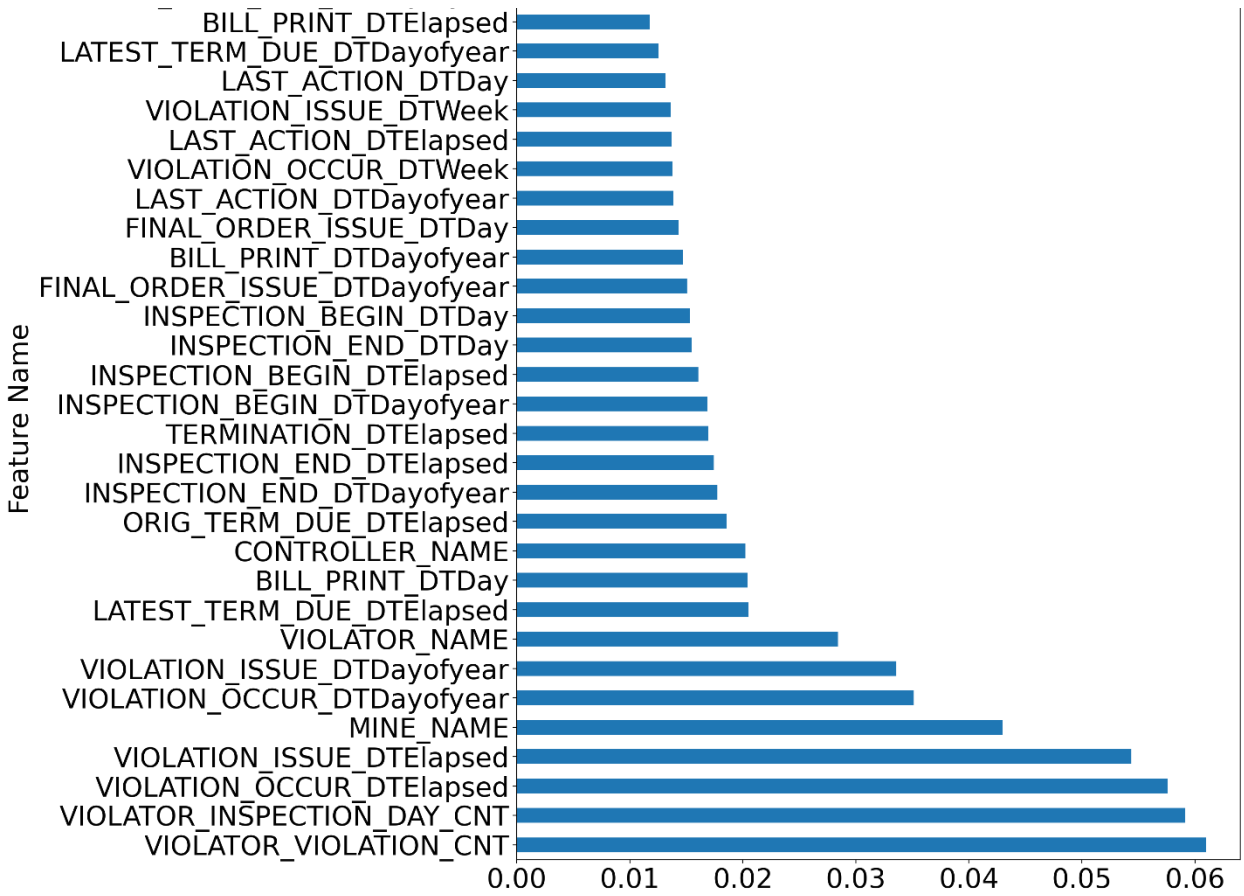

Figure 7: Model A Feature Importance

A confusion matrix was created for Model A; however, due to many unique prediction classes, the resulting matrix was difficult to interpret. (Model C later reduces the available prediction classes to three to generate an interpretable confusion matrix.) There are many prediction classes due to the different ways Table 9: Table of Scheduled Charges in MSHA PC-7014 Appendix C can be combine together over a thirty-five day period. Based on the scatterplots of observed versus predicted, Model A does appear to predict values other than zero (indicating no following accident within thirty-five days), the most common case. In the scatterplot in Figure 8, several striations can be seen in the validation and test scatterplots at the

Model A Observed and Predicted Values

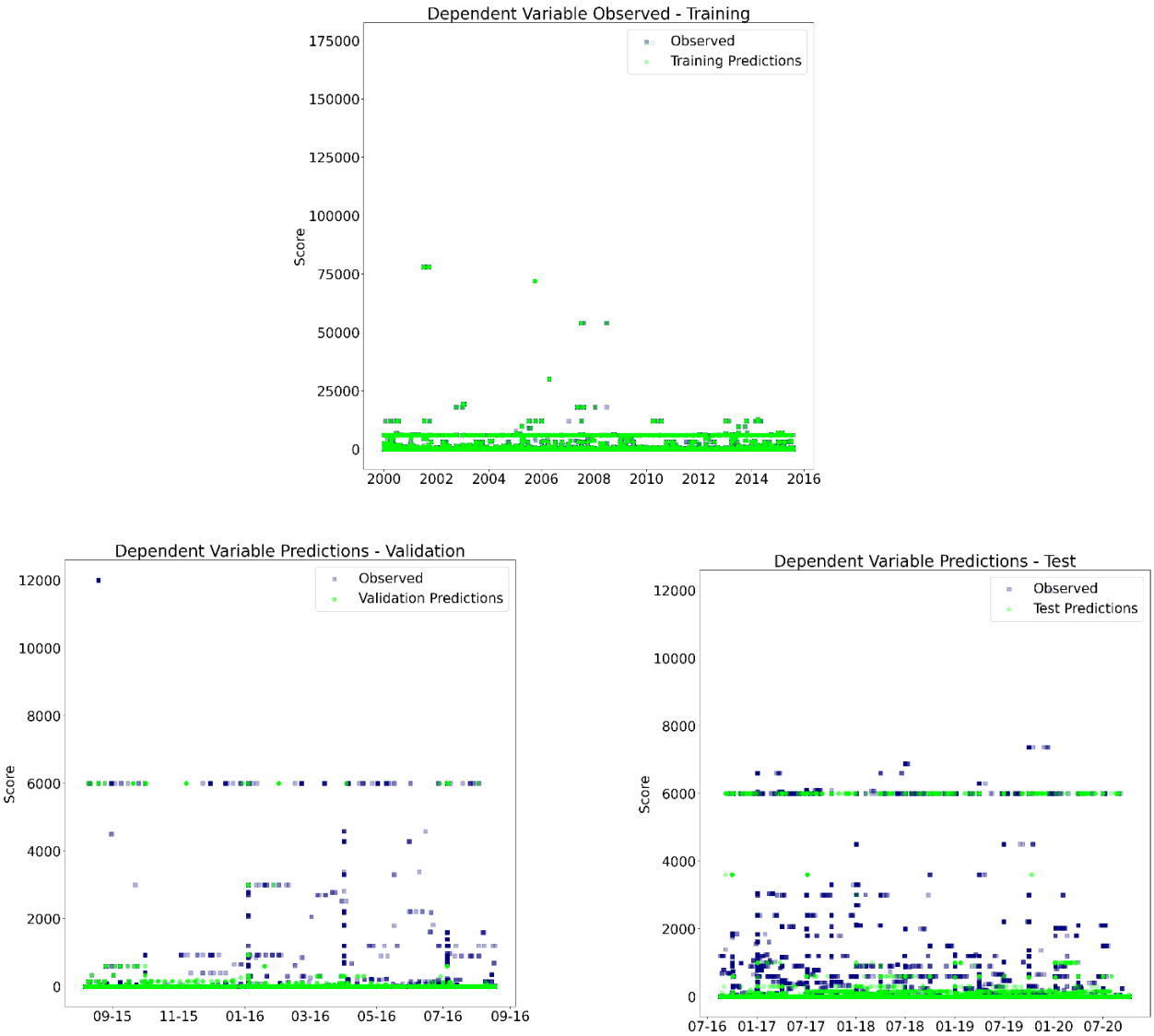

Figure 8: Model A Scatterplot of Observed and Predicted Values

6,000 mark (fatal) and a few values in-between. In later models, where there are more limited classes the model is attempting to predict, the confusion matrix makes it clear what classes are over or underpredicted. This model may be overpredicting a schedule charge of zero, which is the most common case. Later models explore this further by limiting the prediction classes.

Model A is a good baseline to compare the other following models with. The other models attempt to improve this base model. Its primary flaws are too many features, which could be

introducing a “data leak”, and difficulty interpreting confusion matrices due to too many prediction classes. Models B and C both iterate on Model A to explore further. Model A is the most comparable to the prior work [7].

4.2 Model B: Violation Features Without Date Times

The features used to train Model B are the same as Model A, but without datetime features (see Table 11). However, there is some inherit ordering used when training this model because the data was initially sorted by INSPECTION_BEGIN_DT and separated into three different sets before removing the fields. One reason for this change is to better understand the predictive features and avoid potential “data leaks” that date times could introduce without much predictive benefit. A potential “data leak” could be that violations are updated after an accident occurs. An updated date could inadvertently signal that a violation is more important. For example, LAST_ACTION_DT is the “date the last action taken against this violation” [37]. Potentially, if an accident occurred in the future, an MSHA employee could have needed to revise a violation, which inadvertently signals its potential importance.

Model B’s accuracy (Table 15) is comparable to Model A’s (Table 13) Removing the datetime was a prudent choice because it reduces the possibility of a data leak and does not seem to have meaningfully impacted the accuracy of the model.

In this model, VIOLATOR_VIOLATION_CNT, VIOLATOR_INSPECTION_DAY_CNT, MINE_NAME, and VIOLATOR_NAME all rank in the top features based on Gini importance of both Model A, Figure 7, and Model B, Figure 9. It does appear that Model A was only using datetime features to create splits in the dataset.

VIOLATOR_VIOLATION_CNT in the MSHA database is the "total number of assessed violations for this violator at this time during the violation history period. Used in penalty calculation. [sic] Applies to an Operator or a Contractor" [8]. VIOLATOR_INSPECTION_DAY_CNT, likewise, is a similar metric. These two fields that have high feature importance is interesting as it appears that the quantity of violations during the inspection has an impact. It could mean that when an inspector finds many violations, it is an indication that the mine is operating haphazardly.

When designing Model B, another model was made that trimmed the training features further. However, the model stopped becoming predictive of serious and fatal injuries. This alternate model began predicting only zero values. For future work, careful elimination of features is recommended to gain more insight on what features are strictly necessary for a predictive model.

Model B

	Earliest Inspection	*Latest Inspection*	*Samples*	*Mean Accuracy*	*RMSE*
Train	January 02, 2000	August 07, 2015	2,100,000	0.9985	221.9560
Validate	August 07, 2015	August 18, 2016	100,000	0.9744	511.0338
Test	August 18, 2016	October 08, 2020	390,301	0.9707	449.0654

Table 15: Model B Overview

Model B

	Weighted Avg. Recall	Weighted Avg. Precision	Weighted Avg. F1
Train	0.9985	0.9985	0.9985
Validate	0.9744	0.9595	0.9665
Test	0.9707	0.9574	0.9640

Table 16: Model B Performance Evaluation

Model B Feature Importance

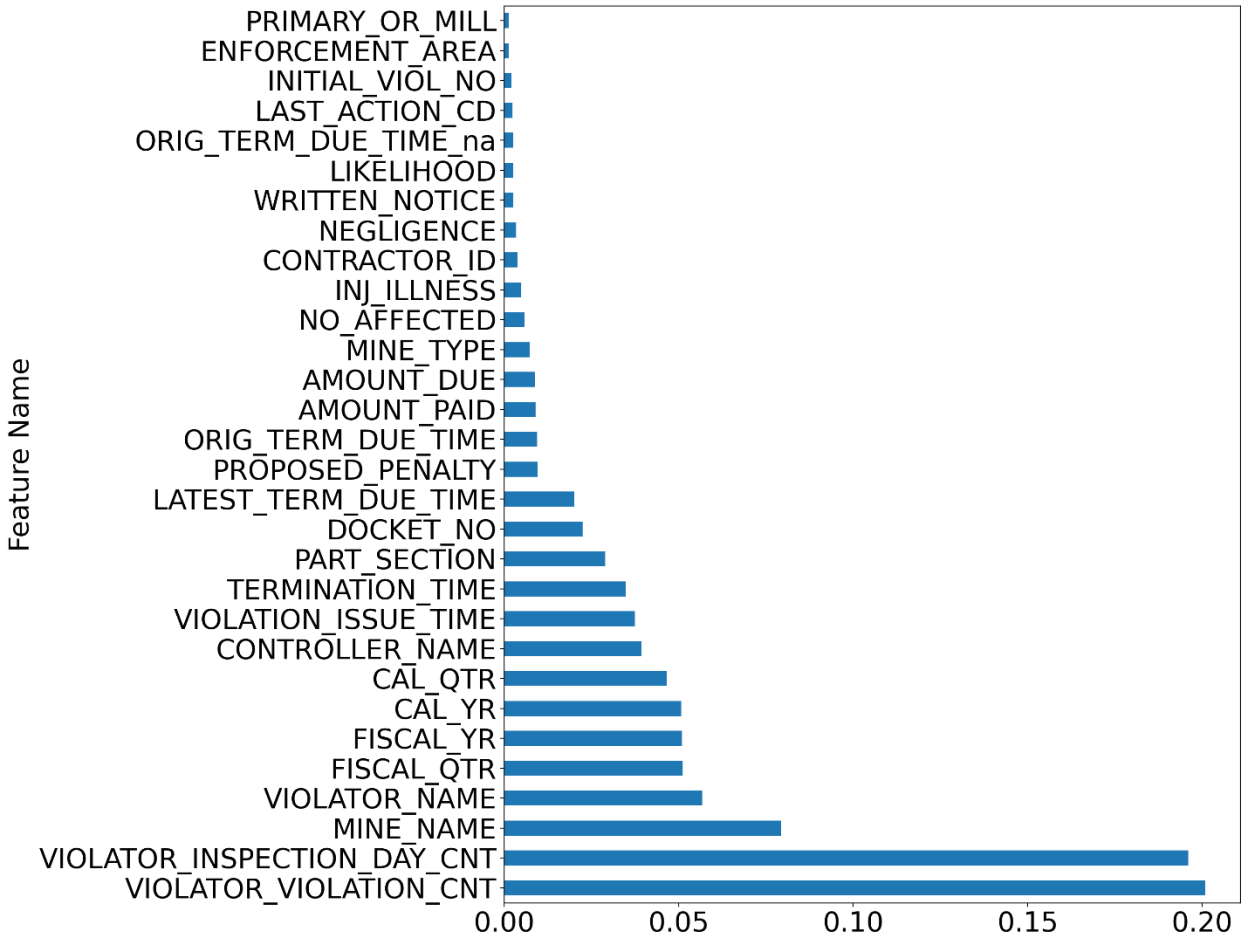

Figure 9: Model B Feature Importance

Model B Observed and Predicted Values

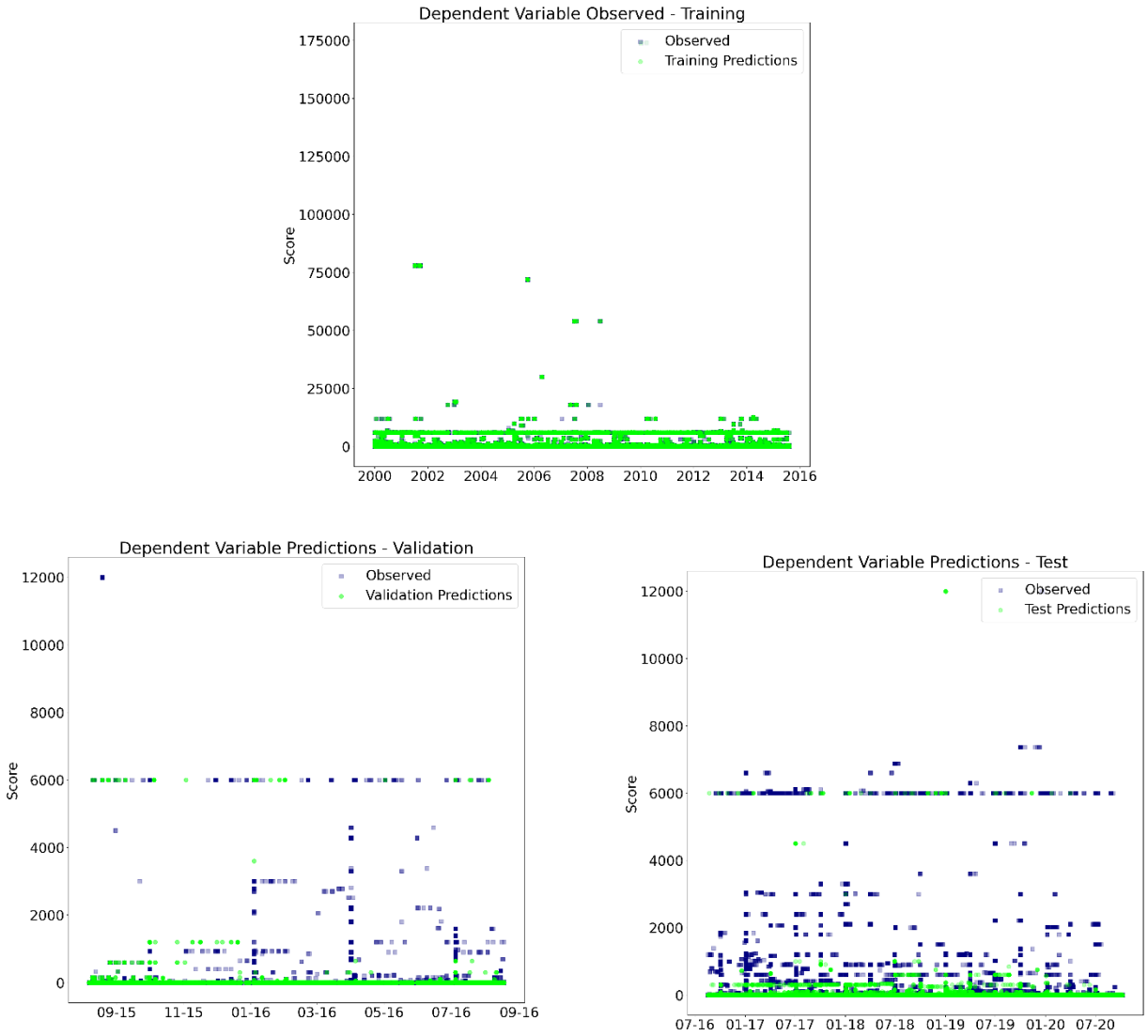

Figure 10: Model B Scatterplot of Observed and Predicted Values

Model B still shows that it is predicting values other than zero for predicting. Like Model A, there are too many features to effectively examine a confusion matrix of this model's performance.

4.3 Model C: Violation Features Without Date Times and Simple Schedule Charge

Model C uses the features listed in Table 11, the same features as used in Model B. Model C differs in that it simplifies the dependent variable SCHEDULE_CHARGE_SUM_35_DAYS into three classes, instead of any value that has appeared in the dataset. The simple schedule charge classes are No Accident (calculated charge of 0), Non-fatal Accident (charge greater than 0 and less than 6,000), and Fatal Accident (greater than 6,000). These values are more representative of the seriousness of the accident than true determinations. The way the values were aggregated, a "Fatal Accident" may indeed be two 3,000 schedule charge accidents aggregated together over the thirty-five-day period; however, the value is indicative of the serious nature of the accident or accidents (a near miss of a fatal accident). The purpose of this division is to allow for better confusion matrix examination; these target variables are also useful for decision-making.

Model C Schedule Charges	
Class	**Value**
No Accident	0
Non-fatal Accident	1
Fatal Accident	2

Table 17: Model C Classes

Model C has three varieties: one with unchanged sampling similar to Models A and B (C.1), one with increased sparse samples (C.2), and one with sampling minimizing by weight (C.3). Model C.1 is most akin to Models A and B. Models A, B, and C.1 can oversample the no-accident classes and subsequently predict the most common outcome of no accident or 0. Model C.2 attempts to improve the prediction capability of Non-fatal and fatal accidents but does not perform well through increasing the amount of accident samples present during training. Finally, Model C.3 begins to show promise as a usable model.

4.3.1 Model C.1: Unchanged Features

Model C.1 is a good baseline for how Models A and B behave. Notice the accuracies are similar (Model A: Table 13, Model B: Table 15, Model C.1: Table 18) of around 99% accuracy for training, 97% for test, and 96-97% for validation. The RMSE is different in Model C.1 from Models A and B because the possible prediction values are between 0 and 2 for C.1 (Table 17), whereas the possible values for Models A and B are between 0 and 174,000. The 174,000 SCHEDULE_CHARGE_SUM_35_DAYS maximum value is the Upper Big Branch Mining disaster.

Model C.1 Unchanged Features - Overview

	Earliest Inspection	*Latest Inspection*	*Samples*	*Mean Accuracy*	*RMSE*
Train	January 02, 2000	August 07, 2015	2,100,000	0.9987	0.0508
Validate	August 07, 2015	August 18, 2016	100,000	0.9717	0.2097
Test	August 18, 2016	October 08, 2020	390,301	0.9743	0.2008

Table 18: Model C.1 Unchanged Features Overview

	Weighted Avg. Recall	*Weighted Avg. Precision*	*Weighted Avg. F1*
Train	0.9987	0.9987	0.9987
Validate	0.9717	0.9599	0.9647
Test	0.9743	0.9651	0.9692

Table 19: Model C.1 Unchanged Features Performance Evaluation

Model C.1's training confusion matrix shows that the model may be predicting the class No Accident as occurring too much. The false negative rate for fatal accidents is very high with 98.05% of the violation test set and 100% of the test set missing fatal accidents (see Table 22). Likewise, the Non-fatal false positive rate is very poor at 93.25% and 79.94% (see Table 21). The goal of the sample minimizing models, C.2 and C.3, is to limit false negatives better.

Model C.1 Unchanged Features - No Accident

Data	True Positive Rate	True Negative Rate	False Positive Rate	False Negative Rate
Train	0.9998	0.9398	0.0602	0.0002
Validate	0.9934	0.0569	0.9431	0.0066
Test	0.9933	0.1654	0.8346	0.0067

Table 20: Model C.1 Unchanged Features, No Accident Confusion Matrix

Model C.1 Unchanged Features - Non-fatal Accident

Data	True Positive Rate	True Negative Rate	False Positive Rate	False Negative Rate
Train	0.9456	0.9998	0.0002	0.0544
Validate	0.0675	0.9937	0.0063	0.9325
Test	0.2006	0.9936	0.0064	0.7994

Table 21: Model C.1 Unchanged Features, Non-fatal Accident Confusion Matrix

Model C.1 Unchanged Features - Fatal Accident

Data	True Positive Rate	True Negative Rate	False Positive Rate	False Negative Rate
Train	0.9219	0.9999	0.0001	0.0781
Validate	0.0195	0.9998	0.0002	0.9805
Test	0.0000	0.9996	0.0004	1.0000

Table 22: Model C.1 Unchanged Features, Fatal Confusion Matrix

Model C.1 Unchanged Features Confusion

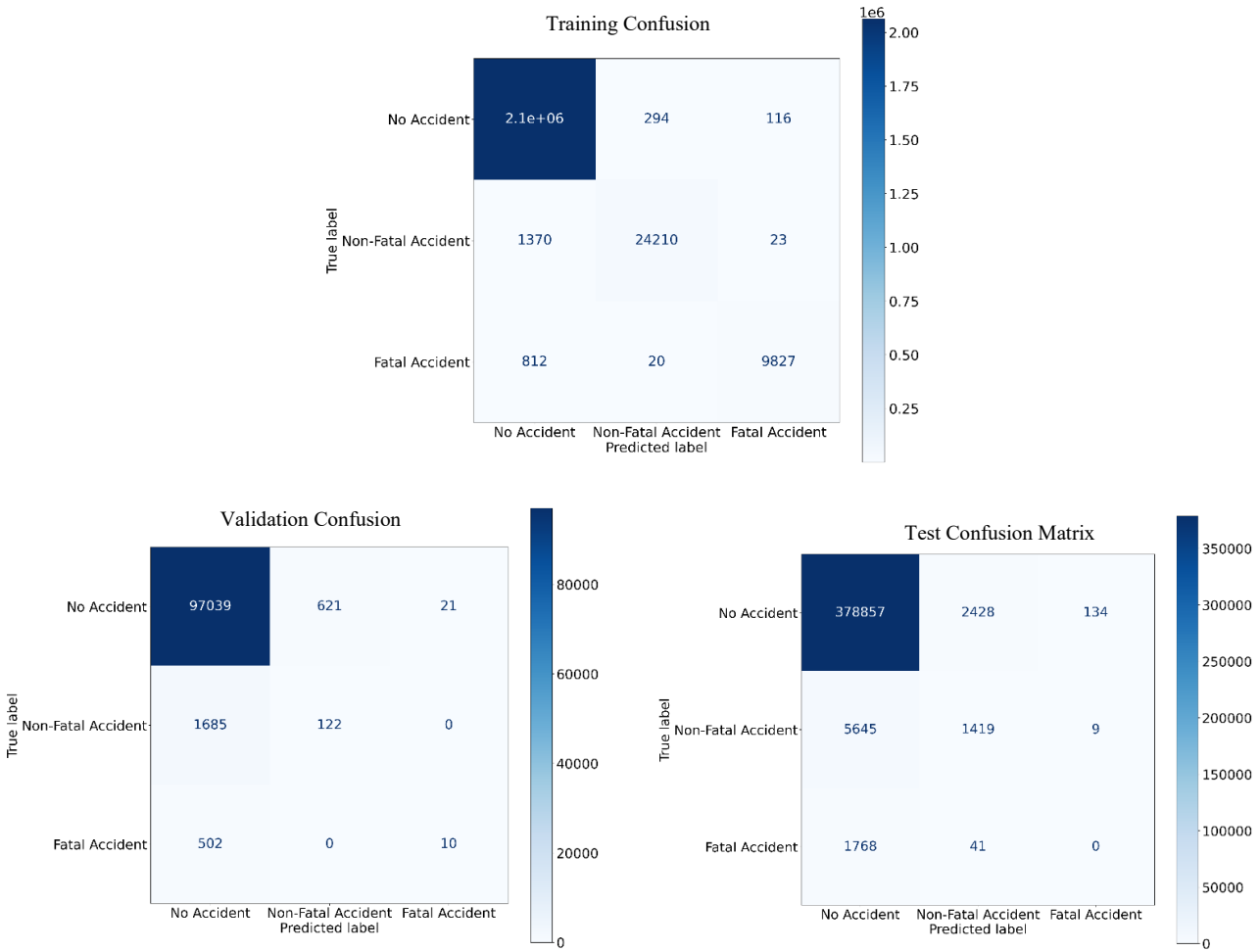

Figure 11: Model C.1 Unchanged Features Confusion Matrices

The unchanged features confusion matrix has an extremely high number of false negatives for fatal and non-fatal accidents. The model appears to be overpredicting the dominate class of no accident. The increased sparse classes and sample minimizing models attempts to correct this issue.

4.3.2 Model C.2: Increasing Sparse Samples

Model C.2 has the same setup as Model C.1, except for Model C.2, non-fatal schedule charges are additionally sampled ten times in the training set and the fatal charges twenty times. Unfortunately, simply increasing the sparse samples of non-fatal and fatal accidents did not improve the false negative rate for non-fatal accidents nor fatal accidents. Notice the 100% false negative rate in the fatal accident test set (Table 27). The next model, which includes sample minimizing, shows movement in these values.

One interesting improvement over model C.1 is the training confusion matrix. In C.1's confusion matrix for training (Figure 11), notice the model does not fit closely to Non-fatal Accidents and Fatal Accidents, i.e., during training, it is not fitting those samples well in the model. However, in C.2, the confusion matrix (Figure 12) shows improvement in model fitting for Non-fatal and Fatal Accidents. This is promising for better fitting the sparse class. Unfortunately, this better fit did not translate when running the model on the validation and test sets, which still have poor prediction levels whenever the true value is Non-fatal or Fatal Accidents.

Model C.2 Increasing Sparse Samples - Overview

	Earliest Inspection	*Latest Inspection*	*Samples*	*Mean Accuracy*	*RMSE*
Train	January 02, 2000	August 07, 2015	2,569,210	0.9985	0.0550
Validate	August 07, 2015	August 18, 2016	100,000	0.9727	0.2095
Test	August 18, 2016	October 08, 2020	390,301	0.9731	0.2076

Table 23: Model C.2 Increased Sparse Samples Overview

	Weighted Avg. Recall	*Weighted Avg. Precision*	*Weighted Avg. F1*
Train	0.9985	0.9985	0.9985
Validate	0.9727	0.9610	0.9661
Test	0.9731	0.9638	0.9680

Table 24: Model C.2 Increased Sparse Samples Performance Evaluation

Model C.2 Increasing Sparse Samples

	True Positive Rate	True Negative Rate	False Positive Rate	False Negative Rate
Train	0.9981	1.0000	0.0000	0.0019
Validate	0.9939	0.0944	0.9056	0.0061
Test	0.9927	0.1415	0.8585	0.0073

Table 25: Model C.2 Increased Sparse Samples, No Accident Confusion Matrix

Model C.2 Increasing Sparse Samples - Non-fatal Accident

	True Positive Rate	True Negative Rate	False Positive Rate	False Negative Rate
Train	1.0000	0.9988	0.0012	0.0000
Validate	0.0968	0.9944	0.0056	0.9032
Test	0.1695	0.9935	0.0065	0.8305

Table 26: Model C.2 Increased Sparse Samples, Non-fatal Accident Confusion Matrix

Model C.2 Increasing Sparse Samples - Fatal Accident

	True Positive Rate	True Negative Rate	False Positive Rate	False Negative Rate
Train	1.0000	0.9995	0.0005	0.0000
Validate	0.0195	0.9992	0.0008	0.9805
Test	0.0000	0.9991	0.0009	1.0000

Table 27: Model C.2 Increased Sparse Samples, Fatal Accident Confusion Matrix

Model C.2 Increased Sparse Classes Confusion Matrix

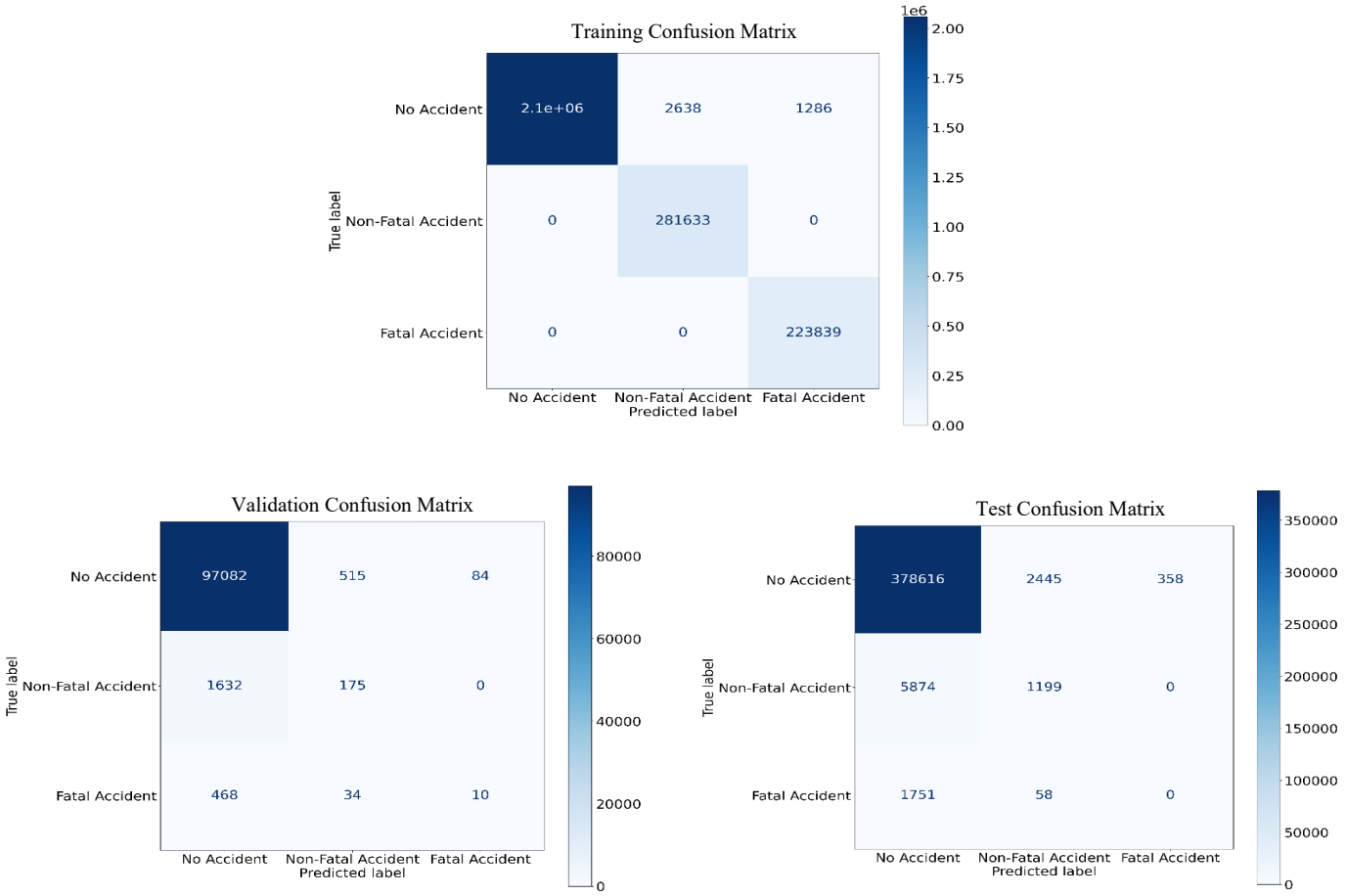

Figure 12: Model C.2 Increased Sparse Class Confusion Matrices

The increase sparse classes greatly improved the training fit for non-fatal and fatal accidents. However, this model does not abstract out to cover the validation and test sets. Notice the false negatives are still high.

4.3.3 Model C.3: Sample Minimizing with Weights

The best-performing model is Model C.3 with sample minimizing with weights. This model had Fatal Accidents sampling increased by **forty** times and Non-Fatal Accident sampling increased **ten** times as appeared in the training set. In addition, the sampling rate of training data was reduced to a weighed pseudo-random sampling at **5% of the original training data**. SIG_SUB violations were used as the weight for sampling with non-Significant & Substantial (S&S) violations weighed as 3, S&S violations weighed at 100, and non-entered values as 1.

The pandas 'sample' function was used to conduct the sampling [44]. In addition, a new hyperparameter was added to the Random Forest Classifier for scikit-learn to use "balanced" class weights – this "mode uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data as n_samples / (n_classes * np.bincount(y))" [43].

Pseudo-random Sampling using S&S Violations as Weights

S&S Type	Weight
Yes, a S&S Violation	100
Not a S&S Violation	3
No Value	1

Table 28: Model C.3 S&S Weights for Random Sampling

One feature that increased in feature importance for the model was SIG_SUB. This criterion was used when sampling, and it was the second most important feature in this model version.

The false negative rate for Fatal Accidents is improved but is still less than ideal. For the test set, 84.96% and 94.14% (Table 33). However, the model shows promise in the Non-Fatal

Accidents false negative rate of 66.85% and 32.08% (Table 32). When tuning the model, the false negative rate for non-fatal accidents values was sometimes better when sampling more of this accident class. This did not make it into final tuning because it worsened the fatal false negative rates.

The overall accuracy rate of this model has fallen to 73.35% for the validation set and 70.09% for the test set (Table 29). However, the improvements in predicting the sparse class of Non-fatal and Fatal Accidents is very important. The false positive rate for No Accidents is 54.29% and 33.49% (Table 31). The test set has many more samples, so it looks like it may be even lower in practice. Balancing this with the false negative rate is critical. On the one hand, missing a Fatal Accident is a much worse outcome, so optimizing false negatives are important. However, overly sensitive models tend to erode user confidence in the model. In future work, achieving this balance would be important for a real-life model.

Another area of improvement for this model is random sampling. This introduces entropy on how the model is trained. Determining what separates good sampling from poor sampling to create a more static sampling function would be ideal.

Model C.3 Sample Minimizing with Weights – Overview

Set	Earliest Inspection	Latest Inspection	Samples	Mean Accuracy	RMSE
Train	January 02, 2000	August 07, 2015	787,390	0.9978	0.0657
Validate	August 07, 2015	August 18, 2016	100,000	0.7335	0.6987
Test	August 18, 2016	October 08, 2020	390,301	0.7009	0.7418

Table 29: Model C.3 Sample Minimizing with Weights Overview

Model C.3 Sample Minimizing with Weights – Performance

Set	Weighted Avg. Recall	Weighted Avg. Precision	Weighted Avg. F1
Train	0.9978	0.9978	0.9978
Validate	0.7335	0.9608	0.8285
Test	0.7009	0.9676	0.8059

Table 30: Model C.3 Sample Minimizing with Weights Performance Evaluation

Model C.3 Sample Minimizing with Weights – No Accident

Data	True Positive Rate	True Negative Rate	False Positive Rate	False Negative Rate
Train	0.9834	1.0000	0.0000	0.0166
Validate	0.7440	0.4571	0.5429	0.2560
Test	0.7043	0.6651	0.3349	0.2957

Table 31: Model C.3 Sample Minimizing with Weights, No Accident Confusion Matrix

Model C.3 Sample Minimizing with Weights - Non-fatal Accident

Data	True Positive Rate	True Negative Rate	False Positive Rate	False Negative Rate
Train	1.0000	0.9978	0.0022	0.0000
Validate	0.3315	0.8161	0.1839	0.6685
Test	0.6792	0.7865	0.2135	0.3208

Table 32: Model C.3 Sample Minimizing with Weights, Non-fatal Accident Confusion Matrix

Model C.3 Sample Minimizing with Weights - Fatal Accident

Data	True Positive Rate	True Negative Rate	False Positive Rate	False Negative Rate
Train	1.0000	0.9984	0.0016	0.0000
Validate	0.1504	0.9262	0.0738	0.8496
Test	0.0586	0.9177	0.0823	0.9414

Table 33: Model C.3 Sample Minimizing with Weights, Fatal Accident Confusion Matrix

Model C.3 Sample Minimizing with Weights Confusion

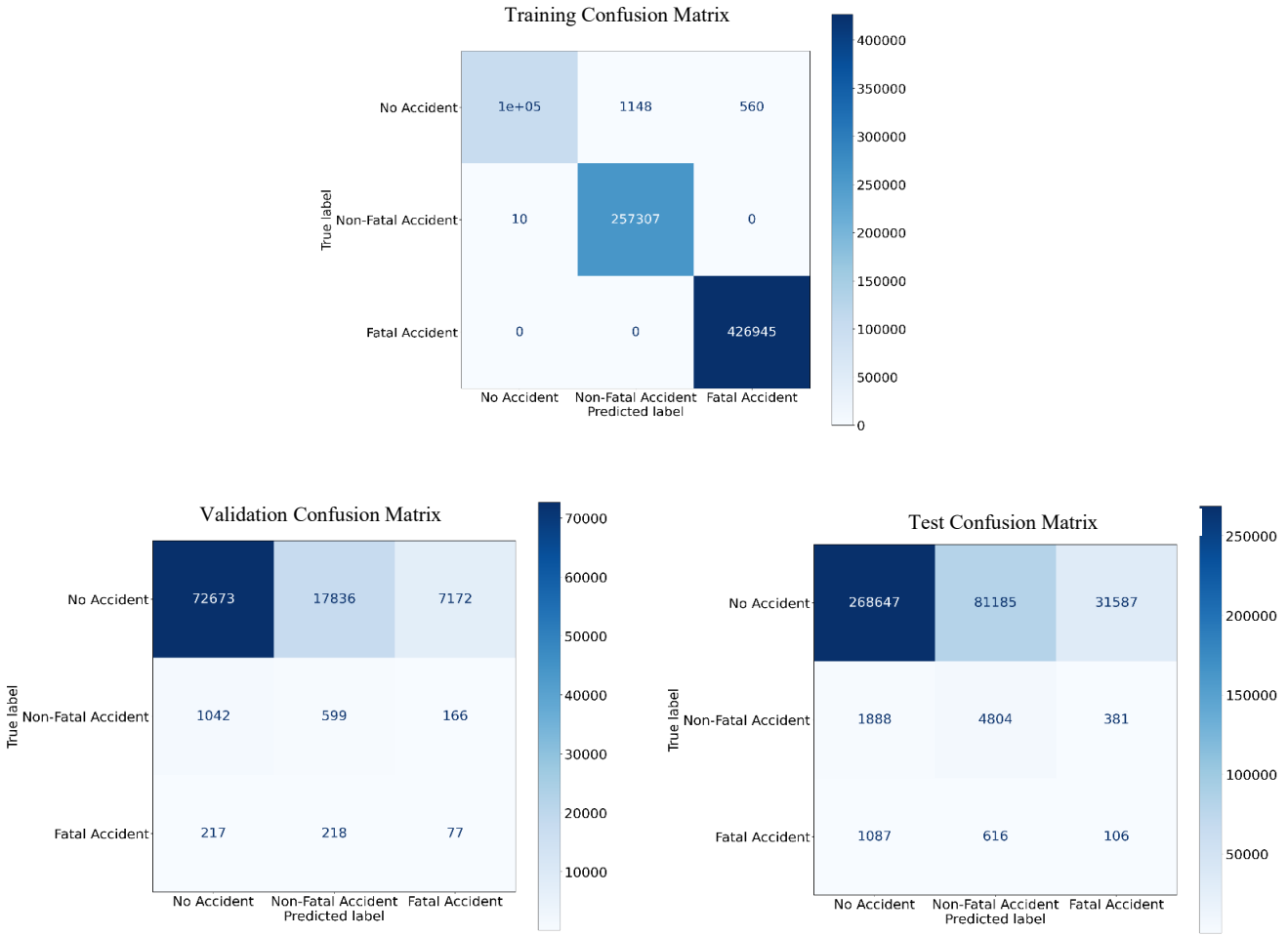

Figure 13: Model C.3 Sample Minimizing with Weights Confusion Matrices

Sample minimizing with Weights helps improve the false negative rate but increases false positives. Finding the appropriate balance is critical for real-world use of the model.

Chapter 5: Conclusion and Future Work

Creating a predictive machine learning model using random forests on MSHA violation and accident data is possible. Finding an ideal model to predict possible mining accidents based on violation data involves balancing many different factors. Identifying the best machine learning method, timescale, independent and dependent variables, and model hyperparmeters all are required to create a robust model that can be used by mining stakeholders.

5.1 Model Feature Improvements

MSHA provides a great wealth of information on U.S. mines. Identifying additional predictive features and creating additional calculated features could be beneficial in improving model robustness and predictive capability. Identifying additional safety-values to predict could also improve the model. There are other MSHA values or calculated values that could be used to determine accident severity or type. For example, worker days lost incidence rates and MSHA inspector severity. Further exploring other potential values to predict would also be another possible future improvement. For example, SCHEDULE_CHARGE is only one of a few other MSHA fields that indicate severity of an accident. DAYS_LOST could also be a good candidate for predicting accident severity. Using them in conjunction could help create a more holistic model.

Possible future work would be exploring the smallest predictive number of features further and identifying additional predictive features, as MSHA has vast quantities of available candidate model feature data. Finding the smallest predictive number of features would give good insight into what parts of the data are most important. Certain thresholds on false positives and false negatives would also need to be established. This core of required features could then

be used as a baseline when adding in new features. Exploring new feature data could include adding in existing MSHA data, such as environmental samples, to attempt to improve the model. Other potentially useful data sources could include coal production amounts, publicly traded mining operator's financial reports, and coal commodity trading prices.

Exploring models and features to predict accidents in small versus large, versus medium mines also would be valuable. A common trend in this area of research is that large mines have fewer accidents than small mines. Adding a variable that accurately indicates mine size would likely improve the model. One difficulty in using the MSHA mines overview data source, which has information such as number of employees and shifts, is that this data is only the current number of employees and does not track changes in operation size. Using other public data, such as SEC reports could be used to find past number of employees. Other data, such as MSHA's employment production table, has yearly employee hours and production. Using this data could provide valuable insight to size-of-mine and how it impacts safety.

5.2 Model Tuning Improvements

In this thesis, many potential models were presented that used different parameters and variables. The model with the most potential was Model C.3, which limited the number of features and narrowed down the predictive classes of accident types. The fatal accident false negative rate needs to be improved before this model could be put into use, as well as the no accident false positive rate.

For future work, improving Model C to limit false negatives and false positives is imperative. Using multiple models may be a solution to this issue. For example, creating multiple binary classification models to assist in violation categorization could be a possible way to improve these rates. These models would each have the options of no accident versus other outcomes,

non-fatal accidents versus other outcomes, and fatal accidents versus other outcomes. This could then be used to gain more granularity in a specific violation's risk. This approach would also offer further transparency to mining decision makers.

Testing other ideas for separating out models could improve performance too. For example, different models and features for large mines versus small mines and underground versus surface mines. The reasoning behind different models depending on mine size are that different features might be disproportionately important for smaller or larger mines. For example, it is likely that the price of coal will impact smaller mines more than large mines with the assumption that smaller mines will shut down or seriously curtail production at lower prices. Additionally, creating models based on certain accident categorizations could be of benefit. MSHA Accidents attributed to methane as an injury source are much more critical and severe than accidents attributed to sunburn, for example. Focusing on injury source or the nature of the injury the accident produces could also be of use.

5.3 Model Improvements

Continuing to find alternate ways to increase the non-fatal and fatal accident signal relative to no accidents is also important for creating *predictive* models that do not simply predict the most common case, in this instance no accidents. As seen on some of the earlier models produced in this thesis, the model can appear predictive, but is in-fact only over predicting the most common case. Using the SIG_SUB (Significant and Substantial) column showed to be a good factor to weigh training on to increase the sparse class sample of an accident. There are other columns that could also be used to improve the signal. For example, the NEGLIGENCE or LIKILIHOOD column could also be used as a factor in selecting the training set values. Identifying especially important features could improve feature sampling.

Additionally, presenting stakeholders with visualizations of the model would be imperative for real world use. Knowing how the model arrived at the solution is critical for building a usable solution. This could be achieved through drawing decision trees and how certain the model is at the answer. A workable tool needs to be transparent to users. Dashboards complete with visualizations such as histograms, line charts, and tables could be prepared for stakeholders to research potential issues before they occur.

5.4 Identifying Top Mining Incidents

In the future, our work will also involve scanning the ultra-large MSHA data repository, which contains numerous tables, to create traceability among them and identify all the incidents that have occurred. This process will help us to determine the most frequently occurring and the most fatal incidents. We firmly believe that identifying the top incidents in the mining industry is crucial for several reasons.

- **Safety**: The mining industry is inherently dangerous due to the nature of the work and the environment in which it takes place. Identifying the top incidents helps to identify the risks associated with mining operations, which can help to reduce the occurrence of accidents and injuries.
- **Compliance**: Mining companies are required to comply with safety regulations and standards set by regulatory bodies. Identifying top incidents can help companies to understand which safety regulations they are failing to comply with and take measures to rectify this.
- **Reputation**: Incidents in the mining industry can have a negative impact on a company's reputation. Identifying top incidents and taking measures to address them can help to improve a company's reputation and maintain stakeholder trust.

- **Financial**: Incidents in the mining industry can be costly in terms of damage to equipment, loss of production, and compensation claims. Identifying top incidents can help companies to take measures to reduce the financial impact of incidents.

Therefore, identifying top incidents as part of future work can help mining companies to comply with safety regulations, maintain their reputation, and reduce the financial impact of incidents. Thereby, enhancing safety and promoting best practices in the mining industry.

References

[1] Mine Safety and Health Administration (MSHA), "Upper Big Branch Mine-South (UBB) Executive Summary Report," [Online]. Available: https://www.msha.gov/sites/default/files/Data_Reports/Fatals/Coal/Upper%20Big%20Branch/ExecutiveSummary.pdf.

[2] Mine Safety and Health Administration (MSHA), "Mine Disaster Investigations Since 2000," [Online]. Available: https://www.msha.gov/data-reports/mine-disaster-investigations-2000.

[3] Mine Safety and Health Administration (MSHA), "Report of Investigation - Fatal Underground Coal Mine Explosion Darby Mine No. 1," 2007. [Online]. Available: https://www.msha.gov/sites/default/files/Data_Reports/FTL06c2731_0.pdf.

[4] Mine Safety and Health Administration (MSHA), "Report of Investigation Fatal Underground Coal Mine Fire Aracoma Alma Mine #1," 2007. [Online]. Available: https://www.msha.gov/sites/default/files/Data_Reports/FTL06c1415total.pdf.

[5] Mine Safety and Health Administration (MSHA), "Report of Investigation Fatal Underground Coal Mine Explosion Sago Mine," 9 May 2007. [Online]. Available: https://www.msha.gov/sites/default/files/Data_Reports/ftl06C1-12wa.pdf.

[6] U.S. Department of Labor, "Mine Inspections," [Online]. Available: https://www.msha.gov/compliance-enforcement/mine-inspections.

[7] O. Milam, M. Haroon and S. Surber, "Digital canaries: identifying hazardous patterns in MSHA data using a machine learner," *Procedia Computer Science*, vol. 177. Elsevier BV, pp. 227–233, 2020 [Online]. Available: http://dx.doi.org/10.1016/j.procs.2020.10.032

[8] Mine Safety and Health Administration (MSHA), "Forms and Online Filing," [Online]. Available: https://www.msha.gov/compliance-and-enforcement/forms-online-filing.

[9] A. B. Szwilski, "Economic environment of coal mining operations in Appalachia, United States," *Mining Science and Technology*, vol. 5, no. 1, pp. 1-10, 1986.

[10] U.S. Department of Labor, "Mining Industry Accident, Injuries, Employment, and Production Statistics and Reports," [Online]. Available: https://arlweb.msha.gov/ACCINJ/accinj.htm.

[11] Mine Safety and Health Administration (MSHA), "Reports: Part 50 Reports," Department of Labor, [Online]. Available: https://www.msha.gov/data-reports/reports.

[12] "Notification, Investigation, Reports and Records of Accidents, Injuries, Illnesses, Employment, and Coal Production in Mines, Code of Federal Regulations. Title 30, section 50," National Archives and Records Administration, 2023. [Online]. Available: https://www.ecfr.gov/current/title-30/chapter-I/subchapter-I/part-50.

[13] U.S. Department of Labor, "Mine Injury and Worktime, Quarterly January-December 2021 Final Report," U.S. Department of Labor, 2021.

[14] U.S. Securities and Exchange Commission, "Mine Safety Disclosure Final Rule 33-9286," 27 January 2012. [Online]. Available: https://www.sec.gov/rules/final/2011/33-9286.pdf.

[15] U.S. Securities and Exchange Commission, "SEC Adopts Dodd-Frank Mine Safety Disclosure Requirements," 21 December 2011. [Online]. Available: https://www.sec.gov/news/press/2011/2011-273.htm.

[16] H. Kinilakodi and R. L. Grayson, "Citation-related reliability analysis for a pilot sample of underground coal mines," *Accident Analysis and Prevention*, no. 43, pp. 1015-1021, 2010.

[17] Mine Safety and Health Administration (MSHA), "Pattern of Violations (POV)," UNITED STATES, April 2021. [Online]. Available: https://www.msha.gov/compliance-enforcement/pattern-violations-pov. [Accessed 08 2022].

[18] F. Molaei, E. Rahimi, H. Siavoshi, S. G. Afrouz and V. Tenorio, "A comprehensive review on Internet of Things (IoT) and its implications in the mining industry," *American Journal of Engineering and Applied Sciences*, vol. 13, no. 3, pp. 499-515, 2020.

[19] C. Zhou, N. Damiano, B. Whisner and M. Reyes, "Industrial Internet of Things: (IIoT) applications in underground coal mines," *Mining Engineering*, vol. 69, no. 12, pp. 50-56, 2017.

[20] L. Chong-mao, N. Rui and Q. Xiang-yan, "Forecast and prewarning of coal mining safety risks based on the internet of things technology and big data technology," *Electronic Journal of Geotechnical Engineering (EJGE)*, vol. 20, no. 20, pp. 11579-11586, 2015.

[21] J. M. Gernand, "Machine learning classification models for more effective mine safety inspections," in *ASME 2014 International Mechanical Engineering and Exposition*, 2014. [Online]. Available: http://dx.doi.org/10.1115/IMECE2014-38709

[22] L. Breiman, "Random Forests," *Machine Learning* , vol. 45, pp. 5-32, 2001.

[23] scikit-learn, "1.11. Ensemble Methods," [Online]. Available: https://scikit-learn.org/stable/modules/ensemble.html.

[24] scikit-learn, "1.10 Decision Trees," [Online]. Available: https://scikit-learn.org/stable/modules/tree.html.

[25] scikit-learn, "sklearn.tree.DecisionTreeClassifier," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html.

[26] IPython , "IP[y]:IPython Interactive Computing," [Online]. Available: https://ipython.org/.

[27] Microsoft, "Visual Studio Code," [Online]. Available: https://code.visualstudio.com/.

[28] Anaconda, "Anaconda," [Online]. Available: https://www.anaconda.com/.

[29] Numpy, "NumPy," [Online]. Available: https://numpy.org/.

[30] pandas, "pandas," [Online]. Available: https://pandas.pydata.org/.

[31] scikit-learn, "scikit-learn: Machine Learning in Python," [Online]. Available: https://scikit-learn.org/stable/.

[32] matplotlib, "Matplot Library," [Online]. Available: https://matplotlib.org/.

[33] fastai, "Welcome to fastai," [Online]. Available: https://docs.fast.ai/.

[34] pandas, "pandas.DataFrame - pandas 1.4.3 documentation," [Online]. Available: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html.

[35] U.S. Department of Labor, "MSHA Data Sets," [Online]. Available: https://arlweb.msha.gov/OpenGovernmentData/OGIMSHA.asp.

[36] U.S. Department of Labor, "Mines Definition File," [Online]. Available: https://arlweb.msha.gov/OpenGovernmentData/DataSets/Mines_Definition_File.txt.

[37] U.S. Department of Labor, "Violations Definition File," [Online]. Available: https://arlweb.msha.gov/OpenGovernmentData/DataSets/violations_Definition_File.txt.

[38] U.S. Department of Labor, "Citation and Order Writing Handbook for Coal Mines and Metal and Nonmetal Mines," December 2013. [Online]. Available: https://arlweb.msha.gov/READROOM/HANDBOOK/PH20-I-13.pdf.

[39] Mine Safety and Health Administration, "Report on 30 CFR Part 50," December 1986. [Online]. Available: https://www.msha.gov/sites/default/files/Support_Resources/Forms/rptonpart50.pdf.

[40] U.S. Department of Labor, "Accidents Definition File," [Online]. Available: https://arlweb.msha.gov/OpenGovernmentData/DataSets/Accidents_Definition_File.txt.

[41] Apache Arrow, [Online]. Available: https://arrow.apache.org/docs/python/feather.html.

[42] Mine Safety and Health Administration (MSHA), "Program Policy Manual VOLUME I INTERPRETATION AND GUIDELINES ON ENFORCEMENT OF THE 1977 ACT," [Online]. Available: https://arlweb.msha.gov/regs/complian/ppm/pmvol1b.htm.

[43] scikit-learn, "sklearn.ensemble.RandomForestClassifier," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[44] Pandas, "pandas.DataFrame.sample," [Online]. Available: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.sample.html.

[45] L. Yang, G. E. Birhane, J. Zhu and J. Geng, "Mining Employees Safety and the Application of Information Technology in Coal Mining: Review," Systematic Review, vol. 9, 2021.

[46] F. Santibáñez, C. Flores, F. Basso, A. Jiménez, F. Bravo, F. Nuñez, H. Luco, L. Martínez and Á. Benítez, "Mining Accident Detection Using Machine," *16th IFAC Symposium on Automation in Mining*, pp. 31-33, 2013.

[47] "Roof control plan, Code of Federal Regulations. Title 30, chapter I, subchapter O, part 75, subpart C, section 75.220," National Archives and Records Administration, 2023. [Online]. Available: https://www.ecfr.gov/current/title-30/chapter-I/subchapter-O/part-75/subpart-C/section-75.220.

[48] U.S. Securities and Exchange Commission (SEC), "Modernization of Property Disclosures for Mining Registrants: A Small Entity Compliance Guide," [Online]. Available:

https://www.sec.gov/corpfin/secg-modernization-property-disclosures-mining-registrants.

Appendix A: Institutional Review Board Letter

MARSHALL UNIVERSITY
www.marshall.edu

Office of Research Integrity

February 2, 2023

Olivia Milam

Dear Olivia,

This letter is in response to the submitted thesis abstract entitled *"Identifying Hazardous Patterns in MSHA Data."* After assessing the abstract, it has been deemed not to be human subject research and therefore exempt from oversight of the Marshall University Institutional Review Board (IRB). The Code of Federal Regulations (45CFR46) has set forth the criteria utilized in making this determination. Since the information in this study does not involve human subjects as defined in the above referenced instruction, it is not considered human subject research. If there are any changes to the abstract, you provided then you would need to resubmit that information to the Office of Research Integrity for review and a determination.

I appreciate your willingness to submit the abstract for determination. Please feel free to contact the Office of Research Integrity if you have any questions regarding future protocols that may require IRB review.

Sincerely,

Bruce F. Day, ThD, CIP
Director

Appendix B: MSHA Violation Fields Used Data Definitions

Violation Field	MSHA Database Definition Description[1]
AMOUNT_DUE	"The current dollar value of the proposed assessment, reflecting any modifications that have been made since it was originally assessed."
AMOUNT_PAID	"The total dollar value of all payments applied to this proposed assessment to date."
ASMT_GENERATED_IND	"If the violator is an Operator or Contractor, the indicator is N. If the violator is an Agent, the indicator is Y. If the violator is a Miner, the indicator can be Y or N. If none of the above, the indicator is ?"
CAL_QTR	"Calendar Quarter of the date the citation or order was issued by the MSHA inspector."
CAL_YR	"Calendar year during which the citation/order was issued by the MSHA inspector."
CIT_ORD_SAFE	"Specifies the type of Citation: Citation, Order, Safeguard, Written Notice or Notice."
COAL_METAL_IND_V	"Identifies if the record is for a Coal or Metal/Non-Metal mine."
CONTESTED_IND	"Indicates if this violation has been assessed and is being contested (Y or N)."
CONTRACTOR_ID	"Code identifying the contractor to whom the citation or order was issued. May contain null values if the contractor was not cited."
DOCKET_NO	"The Docket Number assigned by the Court to this group of Assessments being contested."
DOCKET_STATUS_CD	"Denotes the current status of this docket: Approved (approved by the court) or Proposed (not yet been approved by the court)."
ENFORCEMENT_AREA	"Specifies the enforcement areas affected by the violating condition or practice constitute a health hazard, safety hazard, both or other type of hazard. May contain null values."
FISCAL_QTR	"Fiscal Quarter of the date the citation or order was issued by the MSHA inspector."
FISCAL_YR	"Fiscal Year of the date the citation or order was issued by the MSHA inspector. MSHAs fiscal year begins October 1 and ends September 30."

[1] Feature description is provided by MSHA's data definition file for the VIOLATIONS table [37].

INITIAL_VIOL_NO	"This is the preceding citation record when there is a need to relate a citation to a previous one. For example this would apply when an order follows a citation. This relationship is needed to calculate the good faith reduction penalty points."
INJ_ILLNESS	"Value assigned to a violation for gravity of injury. Measure of seriousness of violation being cited as measured by severity of the injury or illness to persons if accident were to occur due to the conditions of the violation: Fatal, LostDays, NoLostDays or Permanent."
LAST_ACTION_CD	"Last action taken against this violation such as 1stDemandPrinted, BillingReady, ApprovedforTreasury and Proposed."
LATEST_TERM_DUE_TIME	"Time by which the conditions cited on the citation/order are to be abated."
LIKELIHOOD	"This is a measure of the seriousness of the violation being cited as measured by the likelihood of the occurrence of an accident: Highly, NoLikelihood, Occurred, Reasonably or Unlikely. May contain null values if situation does not apply."
MINE_NAME	"Name of the mine where the violation was issued."
MINE_TYPE	"Mine type of the mine where the violation has been issued: Facility, Surface or Underground."
NEGLIGENCE	"Codes representing the degree of negligence that the Inspector assigned to the violator due to the violation: HighNegligence, LowNegligence, ModNegligence, NoNegligence or Reckless. A high degree of negligence is assigned when the operator was in a position to be aware of the condition that contributed to a dangerous situation and there were no mitigating circumstances, or if there are unique aggravating circumstances associated with the violation, such as repeated past violations of the same standard at the mine."

NO_AFFECTED	"This is a measure of the number of persons affected or potentially affected by the conditions at the Mine due to the violation. Can be zero.' , NEGLIGENCE VARCHAR (20) COMMENT 'Codes representing the degree of negligence that the Inspector assigned to the violator due to the violation: HighNegligence, LowNegligence, ModNegligence, NoNegligence or Reckless. A high degree of negligence is assigned when the operator was in a position to be aware of the condition that contributed to a dangerous situation and there were no mitigating circumstances, or if there are unique aggravating circumstances associated with the violation, such as repeated past violations of the same standard at the mine."
ORIG_TERM_DUE_TIME	"Original time by which the cited condition was to be abated."
PART_SECTION	"Code of Federal Regulations: Part/section of Title 30 CFR violated in format PPSSSSSXXXX where (P) Part, (S) Section and (X) Suffix. Four-digit section numbers are expanded to five within one leading zero. May contain null values."
PRIMARY_OR_MILL	"A code indicating if the Violation was observed in the Primary Mine location or in an associated Mill (Metal/Non-Metal only). May contain null values."
PROPOSED_PENALTY	"The original dollar value of the proposed penalty prior to any modifications such as those possibly resulting from a decision on a contested case."
REPLACED_BY_ORDER_NO	"Order number which replaced the original citation. May contain null values if situation does not apply."
SECTION_OF_ACT	"Section of the Act under which the citation/order was issued. May contain null values."
SECTION_OF_ACT_1	"Primary Section of Act which gives the MSHA Inspector the authority to take the action specified by this Issuance. More than one type of action may be cited."
SECTION_OF_ACT_2	"Secondary Section of Act which gives the MSHA Inspector the authority to take the action specified by this Issuance at Metal/Non-Metal mines only. More than one type of action may be cited."

SIG_SUB	"An indicator as to whether or not the gravity is determined by the inspector to be significant and substantial. If this is Y, the inspector has indicated that based upon the particular facts surrounding the violation there exists a reasonable likelihood the hazard contributed to will result in an injury or illness of a reasonably serious nature."
SPECIAL_ASSESS	"Specifies whether this citation has been designated for Special Assessment based on Special Assessment Review (Y or N)."
TERMINATION_TIME	"Time of day (24 hour) at which the citation/order was terminated. May contain null values if citation has not yet been terminated."
TERMINATION_TYPE	"Code identifying the type of termination: Issued, ReplacedByOrder or Terminated."
VACATE_TIME	"Time of day (24 hour) at which the citation/order was vacated. May contain null values if the violation was not vacated."
VIOLATION_ISSUE_TIME	"Time (24 hour) the citation or order was issued by the MSHA inspector."
VIOLATOR_INSPECTION_DAY_CNT	"Total number of assessed violations for this violator at this time during the violation history period. Used in penalty calculation. Applies to an Operator or a Contractor."
VIOLATOR_NAME	"Name of the operator active at the time the violation was cited. May contain null values if this record pertains to a violation issued to a contractor."
VIOLATOR_TYPE_CD	"Each Violator record represents an entity (Operator, Contractor, Agent or Miner) that has one or more violations at a mine."
VIOLATOR_VIOLATION_CNT	"Total number of assessed violations for this violator at this time during the violation history period. Used in penalty calculation. Applies to an Operator or a Contractor."
WRITTEN_NOTICE	"Indicates if this citation is a result of a Miner or Agent notice of complaint to MSHA (written notice 103(g)): (Y or N). May contain null values."
CONTROLLER_NAME	"Name of the controller active at the time the violation was cited. May contain null values if this record pertains to a violation issued to a contractor."

Datetime Violation Field	MSHA Database Definition Description[2]
BILL_PRINT_DT	"Date the bill was printed. This date always represents the first time the bill was printed."
CONTESTED_DT	"Date of the most recent docket status for this violation."
FINAL_ORDER_ISSUE_DT	"Date that this assessment becomes a Final Order. This date is set when the Certified Return Receipt date (CRR) is set. Note that this can be a projected future date that is set as soon as the CRR is entered."
INSPECTION_BEGIN_DT	"Start date of the inspection (mm/dd/yyyy)."
INSPECTION_END_DT	"Inspection close out date (mm/dd/yyyy)"
LAST_ACTION_DT	"Date the last action taken against this violation."
LATEST_TERM_DUE_DT	"Date by which the conditions cited in the citation/order are to be abated. For Metal mines, this can be the termination due date to which the citation/order is extended."
ORIG_TERM_DUE_DT	"Original date by which the cited condition was to be abated. Original time by which the cited condition was to be abated."
RIGHT_TO_CONF_DT	"Date the operator was advised of his right to a conference (Metal/Non-Metal only). May contain null values."
TERMINATION_DT	"Date on which the citation/order was terminated. May contain null values if citation has not yet been terminated."
VACATE_DT	"Date on which the citation/order was vacated. May contain null values if the violation was not vacated."
VIOLATION_ISSUE_DT	"Date the citation or order was issued by the MSHA inspector."
VIOLATION_OCCUR_DT	"Actual date of occurrence of the violation."

[2] Feature description is provided by MSHA's data definition file for the VIOLATIONS table [37].

Appendix C: Acronymns

CART	Classification and Regression Tree
conda	Anaconda Python Environment Management
CSV	Comma Separated Values
DoL	U.S. Department of Labor
IIoT	Industrial Internet of Things
IoT	Internet of Things
IPython	Interactive Python
matplotlib	Matplot Library
Mine Act	The Federal Mine Safety and Health Act of 1977
MSE	Mean Squared Error
MSHA	U.S. Mine Safety and Health Association
NaN	Not a number in computing
NaT	Not a time in computing
NIOSH	U.S. National Institute for Occupational Safety and Health
numpy	Numerical Python
pandas	Python Data Analysis Library
POV	MSHA Pattern of Violations
R^2	R-squared or coefficient of determination
RA	(Citation-related) Reliability Analysis
RMSE	Root Mean Squared Error
S&S	Significant and Substantial MSHA Violation

scikit-learn	Scikit Learn
SEC	U.S Securities and Exchange Commission
SM	Severity Measure