# A Data-Driven Early Warning System for Mining Accidents

Catherine Zhao

*Chemical Engineering, Computer Science, and Business School Columbia University*

## Abstract

This study analyzes how to use predictive analytics to forecast severe accidents in mines. By using the data from the United States Department of Labor's Mine Safety and Health Administration, also known as MSHA, a model was created to determine if there is a potential fatality or disability in the future quarter. The final model used a conditional logistic effect with a final accuracy of 77.2%.

*Keywords:* Mines, Fix Effect Model, R

## 1. Introduction

Historically, the mine industry has been one of the most dangerous fields to work in. From the long work hours to toxic fumes, theses condition has caused many accidents. However, in recent history, there has been a significant decrease of mine accident per year, as shown in Figure 1. This is in large effect due to the United States Department of Labor's Mine Safety and Health Administration, also known as MSHA. Formed in 1977, they strive to prevent death, illness, and injury from mines by promoting safe and healthful workplaces.[1] To do so, they create regulation and collect records of accidents and violations.

One of the worst accidents in recent history occurred on April 5, 2010, at the Upper Big Branch Mine in Raleigh County, West Virginia. Upper Big Branch Mine Disaster resulted in 29 deaths. Because of toxic gases in the mine, MSHA investigator waited for over two months to start investigation. After the investigation was concluded, it was determined that the "root cause of tragedy" was "unlawful policies and practice",[2] which supports stricter regulation to promote safer mines. Similarly, in 2006, Sago Mine Disaster resulted in 13 miners trapped for 2 day with 1 surviver. Because of this accident, MSHA drew heavy criticism from having lax regulation which urged MSHA to create stricter regulations.

The goal of this research is to prevent systemic risk within the industry. Systemic risk is defined as the risk of collapse of a whole system. In particular, to prevent one off accidents, such as Upper Big Branch and Sago Mine, and develop early warning

[1] https://www.msha.gov/about/mission

[2] https://arlweb.msha.gov/MEDIA/PRESS/2011/NR111206.asp

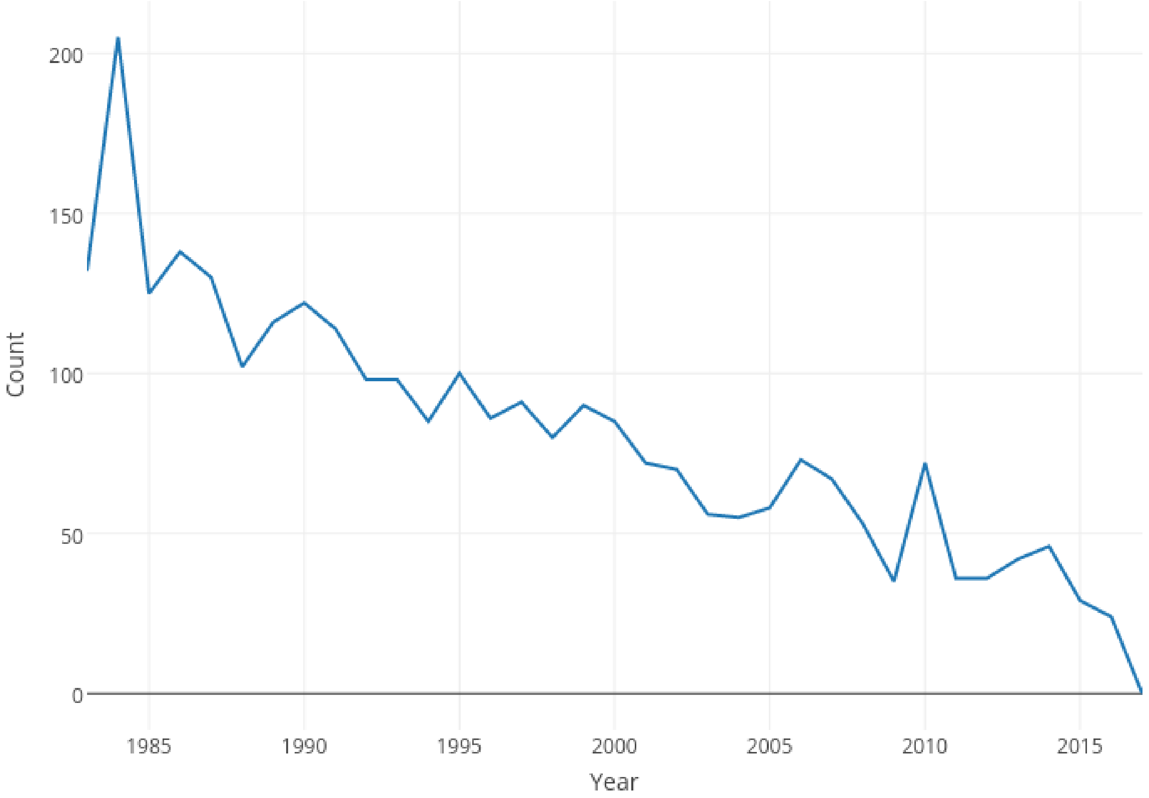


Figure 1: Decrease in Number of Accident Per Year Over Time

systems based on past behaviors. We will be using a method similar to credit scores to calculate the likely chance of a default. Accidents are similar to defaults. Violations are missed payments or late payments.

The raw data contains missing data, human errors, and are not numeric. To complete the research, the data had to be heavily cleaned.

## 2. Data Cleaning and Explorations

### *2.1. Data Cleaning*

This study used three open datasets from MSHA: assessed violation [3], accidents [4], and mines [5]. Due to the decline in accidents over the recent years, accidents and violations between 2000 and 2015 were used. In the raw data, each row represented data for a unique event such as an accident for an individual mine.

In the accident dataset, we furthered filtered the severity of the accident. If the degree injury column contained one of the following: ”days away from work only”,

[3] Retrieved 4/12/2017, from https://arlweb.msha.gov/OpenGovernmentData/DataSets/AssessedViolations.zip

[4] Retrieved 4/12/2017, from https://arlweb.msha.gov/OpenGovernmentData/DataSets/Accidents.zip

[5] Retrieved 4/12/2017, from https://arlweb.msha.gov/OpenGovernmentData/DataSets/Mines.zip

"dys awy frm wrk & restrctd act", "days restricted activity only", "fatality", "perm tot or perm prtl disablty", or "no value found", then it was kept. If the injury degree was "no value found", the number of days lost and number of days restricted was set to zero. In addition, two new columns were created: death and perm_dis. The death column would be set to one if a death occurred. The same procedure was applied to permanent disabilities.

In the Assessed Violation dataset, the date column was extracted. Based on the date, the quarter and year were taken.

Because violation and accidents were taken in the unit of day, all row were consolidated to quarters. This method is used in the banking industry to group defaults and payments to months or quarters. For example, if the mine has one violation per month for three months, this would be consolidated to three for this quarter.

Then from all datasets the following columns were selected:

| Dataset | Column Extracted |
|---|---|
| Assessed Violation | mine_ID, cal_qtr, cal_yr, violation and proposed_penalty_amt |
| Accidents | mine_ID, cal_qtr, cal_yr, days_lost, days_restrict, death, and perm_dis |
| Mines | mine_ID, curr_mine_name, coal_metal_ind, current_mine_type, and no_employees |

Table 1: Columns Extracted

The goal of our study is to predict one quarter into the future. To do this three time measurements were calculated: prior month, 1 year, and 3 years. By using rollover function provided in the RcppRoll package, we summed the previous quarter to the desired amount. For example, to calculate the number of days restricted within 3 years, 12 previous quarters were summed. This was done for days lost, days restrict, death, and permanent disability for accidents, and violation, and proposed penalty amount for assessed violation.

Then for each mine in a given year and quarter, all attributes calculated from above were combined. Then for each mine, the current mine type, number of employees and coal or metal were merged.

Finally, after all attributes were calculated, the sum of the numeric values was taken. If the sum is greater than zero, then quarter was labeled as active. This means for a mean to be active, it needs to have at least one violation or accident within the past three years. This was done to eliminate mines that were not active during the specific time period. Then the cleaned data was filtered to only contain active quarters.

The following is a small selection of the final table. For each row, the mine attributes were shown.

| Mine ID | Mine Name | C/M | Mine Type | Employee | Active |
|---|---|---|---|---|---|
| 100003 | O'Neal Quarry & Mill | M | Surface | 108 | TRUE |

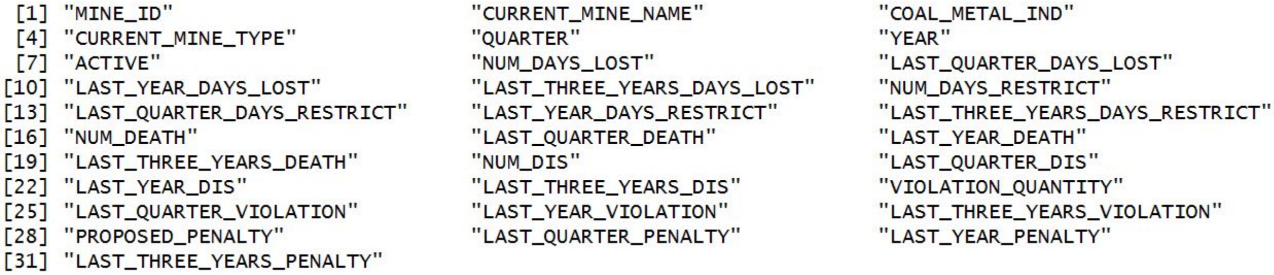
Table 2: Example for Individual Mine

Then for each row, it also shows the calculated attributes. In the example below, the show a mine during the year 2005 with its days lost shown.

| Year | Quarter | Active | Days Lost | Last Quarter | Last Year | Last 3 Year |
|---|---|---|---|---|---|---|
| 2005 | 1 | T | 4 | 33 | 353 | 353 |
| 2005 | 2 | T | 0 | 4 | 37 | 357 |
| 2005 | 3 | T | 0 | 0 | 37 | 357 |
| 2005 | 4 | T | 0 | 0 | 37 | 357 |

Table 3: Mine ID = 100003, Example of number of days lost

The consolidated data contained 461,604 number of rows, 8,877 unique mine and 31 attributes. Each mine contained a maximum of 52 rows. Each row represents data for a unique combination of mine, year, and quarter.



```
 [1] "MINE_ID"                      "CURRENT_MINE_NAME"              "COAL_METAL_IND"
 [4] "CURRENT_MINE_TYPE"            "QUARTER"                        "YEAR"
 [7] "ACTIVE"                       "NUM_DAYS_LOST"                  "LAST_QUARTER_DAYS_LOST"
[10] "LAST_YEAR_DAYS_LOST"          "LAST_THREE_YEARS_DAYS_LOST"     "NUM_DAYS_RESTRICT"
[13] "LAST_QUARTER_DAYS_RESTRICT"   "LAST_YEAR_DAYS_RESTRICT"        "LAST_THREE_YEARS_DAYS_RESTRICT"
[16] "NUM_DEATH"                    "LAST_QUARTER_DEATH"             "LAST_YEAR_DEATH"
[19] "LAST_THREE_YEARS_DEATH"       "NUM_DIS"                        "LAST_QUARTER_DIS"
[22] "LAST_YEAR_DIS"                "LAST_THREE_YEARS_DIS"           "VIOLATION_QUANTITY"
[25] "LAST_QUARTER_VIOLATION"       "LAST_YEAR_VIOLATION"            "LAST_THREE_YEARS_VIOLATION"
[28] "PROPOSED_PENALTY"             "LAST_QUARTER_PENALTY"           "LAST_YEAR_PENALTY"
[31] "LAST_THREE_YEARS_PENALTY"
```

Figure 2: Columns from Final Data Frame

### *2.2. Exploratory Data Visualization*

To understand the data, the summary of the major attributes were calculated, as shown in Table 4.

| | Day Lost | Day Restrict | Death | Disability | Violation | Penalty |
|---|---|---|---|---|---|---|
| Min | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0 |
| 1st Qu. | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0 |
| Median | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0 |
| Mean | 7.41 | 2.21 | 0.00 | 0.00 | 3.17 | 2407 |
| 3rd Qu. | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 127 |
| Max | 2940 | 1665 | 29 | 3 | 515 | 10671517 |

Table 4: Summary of Major Attributes

The data is heavily skewed to the right, with multiple factor having over 75% of their attributes as 0.

In Table 5, then the greatest fatality after 2005 was analyzed. The 2 most dangerous accidents were upper big branch and Sago mine.

| Mine name | Mine ID | Year | Quarter | Number of death |
|---|---|---|---|---|
| Upper Big Branch Mine-South | 4608436 | 2010 | 2 | 29 |
| Sago Mine | 4608791 | 2006 | 1 | 12 |
| Crandall Canyon Mine | 4201715 | 2007 | 3 | 9 |
| Darby Mine No 1 | 1518185 | 2006 | 2 | 5 |
| Gibson Mine | 1202215 | 2007 | 3 | 3 |

Table 5: Top 5 Fatal Accidents Since 2005

Two measurement were used to analyze the accident in a given year: number of violation, and the average penalty per violation in one year. The violation does not indicate the severity of the violations. To capture this, the average penalty was taken by dividing the proposed penalty over the number of violations.

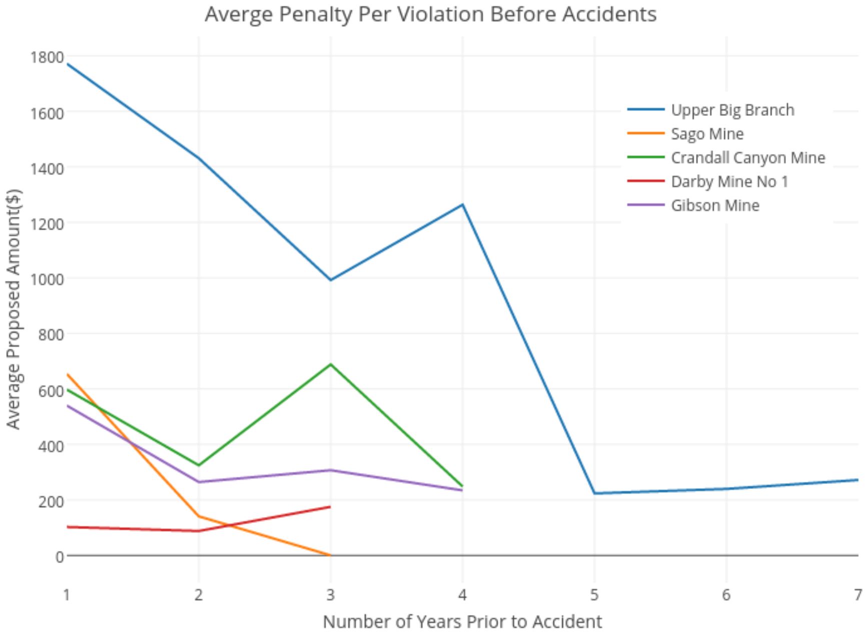


Figure 3: Average Penalty Prior to Top Five Accidents

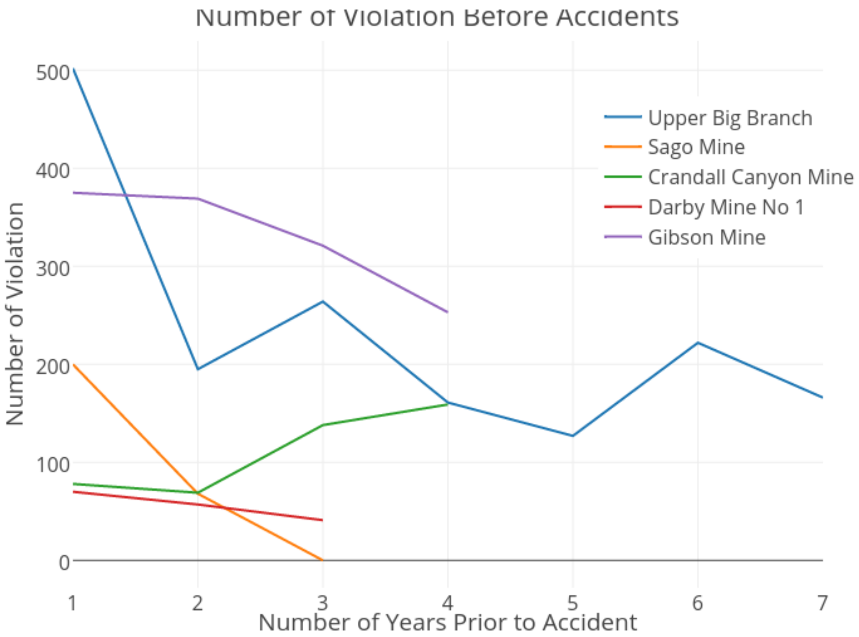


Figure 4: Violation Prior to Top Five Accidents

In Figure 3 and 4, the graphs show the number of violation and average penalty before a major accident. In Figure 3, there is a rising trend as the years get closer to the accident. This indicates the violation are more severe. In Figure 4, the number of violation is also shown to be increasing except for Darby Mine. There is a noticeable increase for both graph. This means the number of violation increase along with the severity.

## 3. Analysis

In our analysis, conditional logistic regression was used. This is found in the survival::clogit in R. Typically used in biostatistics and epidemiology, conditional logistic regression is suitable for panel data, similar to the consolidated data. The conditional, also known as fixed effect, was applied when a mine has certain characteristics. Our fixed effect includes: coal or metal, and type of mine. The data consists of 3 mine types: Facility, Surface, and Underground, and 2 mining material: coal, and medal. Using these parameters the 6 categories of mines were constructed: Facility-Coal, Facility-Metal, Surface-Coal, Surface-Metal, Underground-Coal, and Underground-Metal.

To understand mathematically of this conditional logistic function, a normal logistic function is below.

$$Pr(Y = 1|\mathbf{X}) = \frac{\mathbf{1}}{\mathbf{1 + e^{-(\alpha+\beta X)}}}$$

The $X$ represented an single mine in a given year and quarter along with its other attributes such as last quarter days lost. Then an additional parameter was added which will act as the fixed effect for this given mine.

$$Pr(Y = 1|\mathbf{X}, \mathbf{i}) = \frac{\mathbf{1}}{\mathbf{1} + \mathbf{e}^{-(\alpha_\mathbf{i} + \beta\mathbf{X})}}$$

where $i$ represents the type of category. In Table 2, the type of category would be surface metal mine.

To add a predictive element to the conditional logistic function, a mine in a given quarter and year with at least one death or disability was defined as severe. The percent distribution was shown in Table 6. There was a very small amount of incidents with severe accidents.

| Severe | Number | Percent(%) |
|---|---:|---:|
| FALSE | 363423 | 99.46 |
| TRUE | 1962 | 0.54 |

Table 6: Severe Distribution

Due to the low number of positive, the train and test data were the entire sample. The results are shown in Table 7.

| | Coefficient | P-Value |
|---|---|---|
| Last Quarter Days Lost | 2.470e-05 | 0.905970 |
| Last Year Days Lost | 5.616e-04 | 6.81e-08 |
| Last 3 Years Days Lost | -1.840e-04 | 2.05e-07 |
| Last Quarter Days Restrict | 8.838e-04 | 0.110570 |
| Last Year Days Restrict | -7.996e-5 | 0.779726 |
| Last 3 Years Days Restrict | 3.740e-04 | 7.06e-05 |
| Last Quarter Death | 1.252e-02 | 0.928404 |
| Last Year Death | 2.316e-02 | 0.757046 |
| Last 3 Years Death | 1.311e-01 | 0.002926 |
| Last Quarter Disability | -2.059e-01 | 0.230381 |
| Last Year Disability | 1.636e-01 | 0.076512 |
| Last 3 Years Disability | 3.842e-01 | 5.55e-16 |
| Last Quarter Violation | 4.013e-03 | 0.001912 |
| Last Year Violation | 1.689e-03 | 0.004640 |
| Last 3 Years Violation | 5.095e-04 | 0.018640 |
| Last Quarter Penalty | -4.714e-07 | 0.269156 |
| Last Year Penalty | -1.093e-07 | 0.546883 |
| Last 3 Years Penalty | -3.404e-07 | 0.000255 |

Table 7: In-Sample Model Using Logistic Regression With Fixed Effects

|  | Reference |  |
|---|---|---|
| Prediction | FALSE | TRUE |
| FALSE | 281028 | 827 |
| TRUE | 82395 | 1135 |

Table 8: Prediction VS. Actual of Logistic Regression with Fixed Effect Output

| Accuracy | Sensitivity | Specificity | Precision | F1 |
|---|---|---|---|---|
| 0.772 | 0.578 | 0.773 | 0.014 | 0.027 |

Table 9: Logistic Regression with Fixed Effect Model Accuracy

The prediction outcome is shown in Table 8. The model predicted 23% positive outcome. Since the original data had a 0.5% likelihood, the model over predicted severe accidents. To understand this performance, the accuracy components were taken and shown in Table 8. The accuracy of the model represents the likelihood that the model prediction is correct is 0.772. The sensitivity means given an severe accidents, the likelihood the model will predict is 0.578. The specificity of the model represents given no severe accidents, the likelihood the model will predict no accident was 0.773. The precision of the model means if the model predict an accidents, the likelihood that there is a true accident is 0.014. F1 takes into account the sensitivity and precision to gain by taking the weighted average of them which is 0.027. The best case scenario is 1 and the worse case is 0.

To gain a better understanding of the model, two queries were excited: the top ten worse accidents, shown in Table 10, and the top ten highest probability of an accident,shown in Table 11.

| Mine ID | Mine Name | Time | Death | Dis | Severe | Prob | Pred |
|---|---|---|---|---|---|---|---|
| 4608436 | Upper Big Branch Mine | 2010-2 | 29 | 0 | T | 0.66 | T |
| 4608791 | Sago Mine | 2006-1 | 12 | 0 | T | 0.61 | T |
| **4201715** | **Crandall Canyon Mine** | **2007-3** | **9** | **0** | **T** | **0.42** | **F** |
| 1518185 | Darby Mine No 1 | 2006-2 | 5 | 0 | T | 0.54 | T |
| 1202215 | Gibson Mine | 2007-3 | 3 | 0 | T | 0.84 | T |
| 4601437 | Marshall County Mine | 2003-1 | 3 | 0 | T | 0.92 | T |
| 4608878 | Affinity Mine | 2013-1 | 2 | 0 | T | 0.67 | T |
| 4608801 | Aracoma Alma Mine #1 | 2006-1 | 2 | 0 | T | 0.62 | T |
| **4609086** | **Black Stallion UG Mine** | **2014-2** | **2** | **0** | **T** | **0.22** | **F** |
| 4609066 | Cucumber Mine | 2007-1 | 2 | 0 | T | 0.68 | T |

Table 10: Top 10 Mines with Greatest Number of Death using Logistic Regression with Fixed Effect

In Table 10, it is shown that the model correctly predicted eight out of the ten most deadly accident. The mines were arranged by greatest number of death, then alphabetically by mine name. Since Crandall Canyon Mine contained nine death and was still rejected, we studied it further.

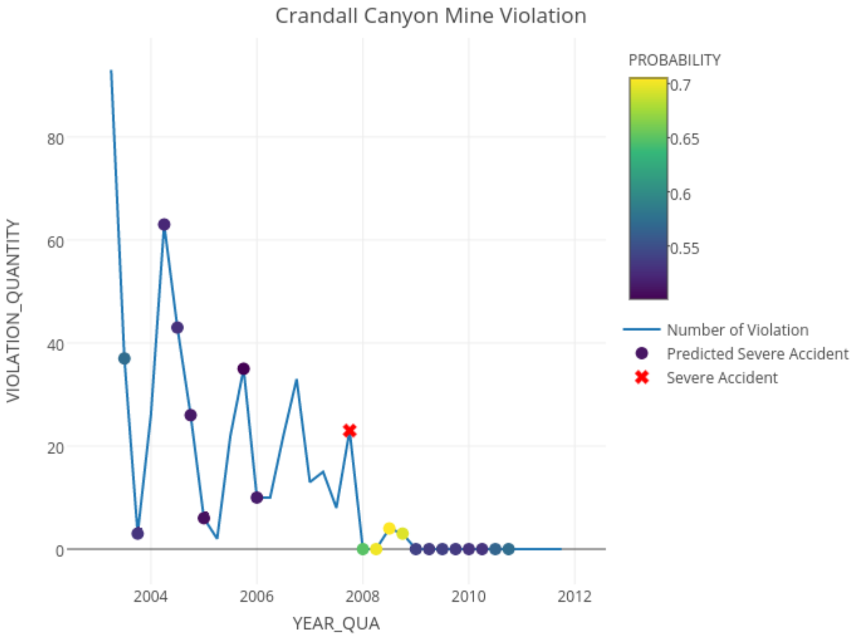


Figure 5: Crandall Canyon Mine Violation

In the Figure 5, the violation quantity was plotted along with its predicted value. The line represented the number of violation in a given quarter. The dots on the graph represents if in the given quarter the model predicted an accident. The color of the dots represent the probability the model predicted an accident. The cross represents an actual accident. In Crandall Canyon Mine, the model predicted the likelihood of an accident a few years before and the actual accident.

| Mine ID | Mine Name | Time | Death | Dis | Severe | Prob | Pred |
|---|---|---|---|---|---|---|---|
| 1102752 | The American Coal Company | 2006-2 | 0 | 1 | T | 0.99 | T |
| 1102752 | The American Coal Company | 2006-1 | 0 | 0 | F | 0.99 | T |
| 1102752 | The American Coal Company | 2009-2 | 0 | 0 | F | 0.99 | T |
| 200024 | Freeport-McMoRan Morenci Inc. | 2015-2 | 0 | 0 | F | 0.99 | T |
| 1102752 | The American Coal Company | 2005-4 | 0 | 0 | F | 0.99 | T |
| 200024 | Freeport-McMoRan Morenci Inc. | 2015-4 | 0 | 0 | F | 0.99 | T |
| 1102752 | The American Coal Company | 2008-3 | 0 | 0 | F | 0.99 | T |
| 1102752 | The American Coal Company | 2009-3 | 0 | 0 | F | 0.99 | T |
| 1102752 | The American Coal Company | 2006-4 | 0 | 0 | F | 0.99 | T |
| 1102752 | The American Coal Company | 2009-4 | 0 | 0 | F | 0.99 | T |

Table 11: Top 10 Mines with Greatest Probability using Logistic Regression with Fixed Effect

In Table 11, the model predicted 2 mines with very likelihood accident rate: The

American Coal Company and Freeport-McMoRan Morenci Incorporation. We further studied American Coal Company.

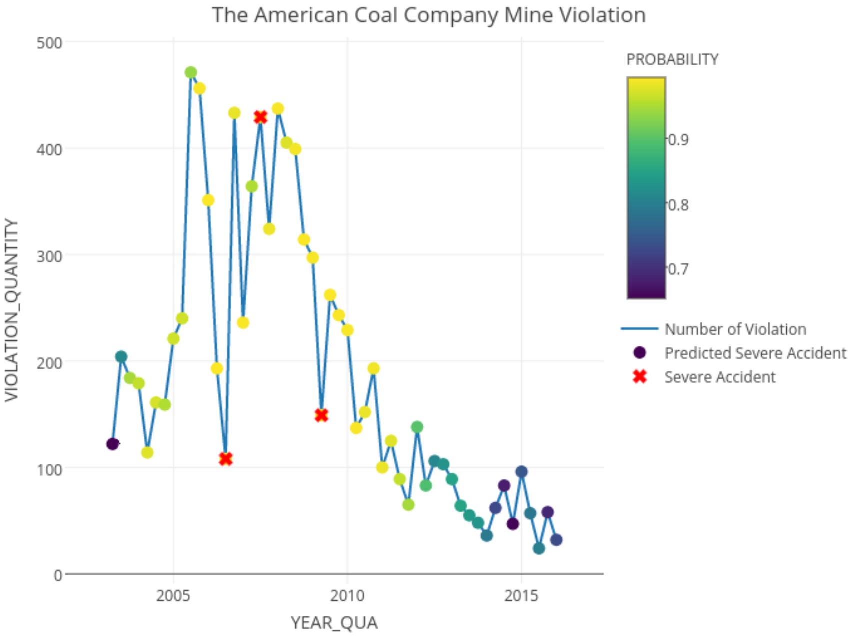


Figure 6: American Coal Company: Number of Severe Incidents VS. Predicted Incidents

In Figure 6, shows a similar graph to Figure 5 for the American Coal Company. the American Coal Company violation was plotter over time. The model predicted every quarter to have a severe accident. There are three actual accidents that occurred. This means the model successfully predicted that this mine is likely to have an accidents. The quarter before the first accidents, there is over 90% of an accident occurring. After the couple of years, there is also a decrease in number of violation for the mine along with its probability for an accident.

From Figure 6 and Figure 5, the model did not fail to give warning signs if a mine has accident. To gain a better understanding how well the modeled predicted, we analyzed the number of severe accident that were predicted and the number of actual sever accident. For example, Upper Big Branch mine accident had 29 deaths, however this will only be labeled as one incidents. Then box plot were plotted to show the number of prediction vs. the number of severe incident, as shown in Figure 7.

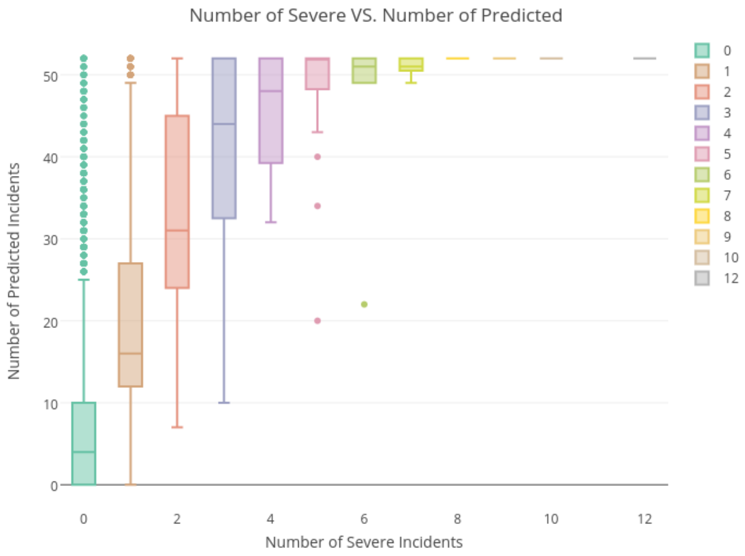


Figure 7: Number of Severe Accident VS. Number of Predicted Accident

It is shown that the number of severe accident has a positive correlation to the number of predicted accident. If there is only one sever accident, the model fails give an warning indication for 14 out of 1,021. Given a mine has at least two accidents, the model predicts at least seven warnings. Given a mine has at least eight accidents, the mine has all quarters labeled at severe meaning this mine is high risk for any given quarter.

Then to measure if the model predicted before an accident, the time of the first severe accident and the first predicted accident was taken. The probability that the model predicted correctly was taken and plotted in Figure 8.

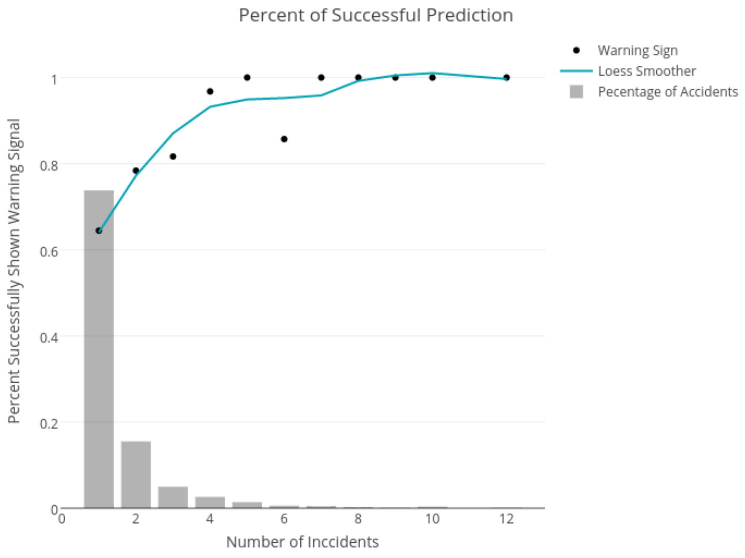


Figure 8: Percent of Successful Prediction.png

The bar graph of the figure shows the percentage of accident. For example, for all mines with an accident, 75.6% of the them have only one accident, which is 1021 out of 1348 mines. There is a significant fewer mines with more than one accident. The black dots represent the success rate that the model predicted an accident before or on the actual accidents. There is an upward trend which shows the model successfully predicting accidents, which shows success in the model.

## 4. Conclusion

Through the efforts of MSHA data collected, there has been a significant decrease overtime the number of accidents. In the final model, the accuracy was 0.774. It also heavily over predicted the number of accidents. Since the goal of the study is to create a warning system, this was furthered studied. The model was able to generate a warning signal to mines prior to the accident somewhat successfully. For example, if the mine had 3 severe accident, the model will predicted 81% before the accident occurred signaling this is a dangerous mine.

## Appendix A. Simple Linear Regression

In the simple linear regression model, the model was fitted to a variable that is a combination of days lost, days restricted, deaths, and disability.

$$num\_lost = num\_days\_lost + 0.5 \times num\_days\_restrict + 300 \times num\_death + 150 \times num\_dis$$

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---:|---:|---:|---:|---:|
| (Intercept) | 0.64 | 0.10 | 6.66 | 0.00 |
| Last quarter days lost | 0.05 | 0.00 | 21.52 | 0.00 |
| Last year days lost | 0.07 | 0.00 | 55.25 | 0.00 |
| Last 3 years days lost | 0.04 | 0.00 | 80.18 | 0.00 |
| Last quarter days restrict | 0.01 | 0.01 | 1.47 | 0.14 |
| Last year days restrict | 0.02 | 0.00 | 6.18 | 0.00 |
| Last 3 years days restrict | 0.03 | 0.00 | 23.49 | 0.00 |
| Last quarter death | 5.22 | 1.84 | 2.83 | 0.00 |
| Last year death | -4.59 | 1.01 | -4.56 | 0.00 |
| Last 3 years death | 1.18 | 0.48 | 2.48 | 0.01 |
| Last quarter disability | 7.72 | 1.63 | 4.75 | 0.00 |
| Last year disability | -2.00 | 0.95 | -2.10 | 0.04 |
| Last 3 years disability | 2.16 | 0.48 | 4.50 | 0.00 |
| Last quarter violation | 0.34 | 0.02 | 21.84 | 0.00 |
| Last year violation | 0.17 | 0.01 | 23.39 | 0.00 |
| Last 3 years violation | -0.02 | 0.00 | -8.35 | 0.00 |
| Last quarter penalty | -0.00 | 0.00 | -4.57 | 0.00 |
| Last year penalty | -0.00 | 0.00 | -5.32 | 0.00 |
| Last 3 years penalty | -0.00 | 0.00 | -7.93 | 0.00 |

Table A.12: Linear Regression Model

Adjusted $R^2$: 0.3405

## Appendix B. Logistic Regression without Fixed Effect

|  | Reference |  |
|---|---|---|
| Prediction | FALSE | TRUE |
| FALSE | 363361 | 1948 |
| TRUE | 62 | 14 |

Table B.13: Prediction VS. Actual of Logistic Regression without Fixed Effect Output

| Mine ID | Mine Name | Time | Death | Dis | Severe | Prob | Pred |
|---|---|---|---|---|---|---|---|
| 4608436 | Upper Big Branch Mine | 2010-2 | 29 | 0 | T | 0.02 | F |
| 4608791 | Sago Mine | 2006-1 | 12 | 0 | T | 0.01 | F |
| 4201715 | Crandall Canyon Mine | 2007-3 | 9 | 0 | T | 0.00 | F |
| 1518185 | Darby Mine No 1 | 2006-2 | 5 | 0 | T | 0.01 | F |
| 1202215 | Gibson Mine | 2007-3 | 3 | 0 | T | 0.06 | F |
| 4608878 | Affinity Mine | 2013-1 | 2 | 0 | T | 0.02 | F |
| 4608801 | Aracoma Alma Mine #1 | 2006-1 | 2 | 0 | T | 0.01 | F |
| 4609086 | Black Stallion UG Mine | 2014-2 | 2 | 0 | T | 0.00 | F |
| 4609066 | Cucumber Mine | 2007-1 | 2 | 0 | T | 0.01 | F |
| 1517165 | D-14 Stillhouse | 2005-3 | 2 | 0 | T | 0.01 | F |

Table B.14: Logistic Model without Fixed Effect Prediction Against Top 10 Worst Accidents

## Appendix C. Logistic Regression with Fixed Effect Model With 5-fold Cross Validation

Given the small sample of severe accident, the testing data ran for entire dataset. The average probability from each validation was taken. There is no significant change with running the model using cross validation.

| | Reference | |
|---|---|---|
| Prediction | FALSE | TRUE |
| FALSE | 281144 | 825 |
| TRUE | 82279 | 1137 |

Table C.15: Prediction VS. Actual of Logistic Regression with Fixed Effect and 5-Fold Cross Validation Output

| Accuracy | Sensitivity | Specificity | Precision | F1 |
|---|---|---|---|---|
| 0.77 | 0.58 | 0.77 | 0.01 | 0.03 |

Table C.16: Logistic Regression with Fixed Effect and 5-fold Model Accuracy

## Appendix D. Accuracy Model Calculations

| | Formula |
|---|---|
| Accuracy | $\frac{True\_pos+True\_neg}{True+False}$ |
| Sensitivity | $\frac{True\_pos}{True}$ |
| Specificity | $\frac{True\_neg}{False}$ |
| Precision | $\frac{True\_pos}{True\_pos+False\_pos}$ |
| F1 | harmonic mean of sensitivity and precision |

Table D.17: Logistic Regression with Fixed Effect Model Accuracy