RailSafe: Predictive Analytics Models for Mitigating Railroad Accidents

Authors: Ethan Morgan(morgan.et@northeastern.edu), Anshul Rao(rao.ans@northeastern.edu)

INTRODUCTION

Motivation and Background

In the United States, railroad accidents have been a recurring issue, resulting in 893 rail-related fatalities in 2021, the highest recorded since 2007. This concerning trend underscores the pressing need to take a proactive stance in identifying and managing the risk factors that contribute to such incidents. By leveraging current data and developing a predictive model for rail accidents, we can better anticipate potential hazards and reduce the likelihood of accidents occurring in the future. The overarching goal of building a predictive model for railroad accidents is to ultimately save lives and prevent future tragedies. A safer and more sustainable rail system would have a positive impact on both passengers and the wider community.

Previous efforts have been made using comparable datasets concerning railway accidents, though the time frame and data origin have differed. However, no work pertaining to this specific dataset has been conducted on Kaggle. Our intention is to analyze this dataset anew, relying on our expertise to derive novel and compelling insights that have not yet been uncovered.

Dataset

The dataset has been taken from Kaggle[1]. The dataset encompasses railway accidents and incidents spanning from 1975 to 2022, with approximately 160 measurements and 216k observations.

The dataset includes several different types of measurements, spanning location, time, weather, counts of people killed, injured, and evacuated, and other attributes of the trains and tracks themselves. There are also variables regarding severity, such as different costs of accidents and cause of accidents. Finally, there is a link to the entire PDF of the standardized form used to report each accident.

One interesting factor of this dataset is the identifying IDs. Each accident can be reported up to three times: once by each involved party (one or two different rail companies if it was a crash between two trains) and separately by the maintenance company. This required some work to accurately remove duplicates, since each report may not be complete and has to be joined with the other reports dealing with the same accident.

METHODOLOGY

Low Risk

Our focus for the low-risk objectives was mainly on conducting exploratory data analysis, which involved exploring the dataset and gathering insights through visualizations. This included univariate and multivariate analysis, extracting correlations and distributions, and creating visualizations of insights.

Additionally, we performed some feature engineering. There were a large number of redundant features, for instance, categoricals that had both a name and a category number. Some computed features had to be added, for instance, normalizing cost fields for inflation due to the length of time represented in this data.

Medium Risk

For medium-risk objectives, we had two supervised learning tasks related to predicting various attributes. First, we aimed to predict the overall cost of accidents, based on several input features. The target variable was Total Cost, normalized to 2022 inflation using the *cpi* library. This cost comprises several categories, namely equipment damage, track damage, and other costs. Since we used a combination of numerical and categorical features, several transformations were required. The categorical features were one-hot encoded (using empty fields as a separate category), and the numerical fields were normalized to values between 0 and 1. We tried several regression models, including the *sklearn* implementations of Gradient Boosting and State Vector Machines.

The second objective was to predict the cause of accidents. Each observation can have one or multiple causes, which are reported using a hierarchical structure consisting of one letter and up to three numbers. For example, ‘E’ represents a mechanical or electrical failure, whereas ‘E72L’ specifically represents a crank case or air box explosion. Using the same features as the regression task, we used sklearn’s Gradient Boosting and

Logistic Regression classifiers to predict various granularities of the cause codes.

High Risk

For the high-risk objectives, we had planned to identify the safest and most hazardous transit routes, however, the dataset only comprises information about accidents, which poses a challenge due to selection bias. The dataset does not include information on near-misses and may exclude many accidents that have occurred on safer or more dangerous routes that were not reported or recorded. Also, the dataset does not take into account other factors that may impact the safety of transit routes, such as the behavior of drivers and passengers. Hence, we could not accomplish this goal due these limitations in the dataset and it proved to be an excellent failure. Our second high risk objective was to conduct time-series forecasting by constructing an autoregressive moving average model, predicting the frequency of accidents in the future. We were able to accomplish this goal and used the statsmodels module in Python for building the model. Our first step was to check for stationarity and we made the time-series stationary by applying the first difference. Next, we determined the order of the ARMA model by analyzing AIC and BIC scores. Once we identified the appropriate order, we fitted the model and used it for forecasting future values.

RESULTS

EDA

1. Accident frequency across US states: *Most of the railroad accidents occur in Texas and Illinois.*

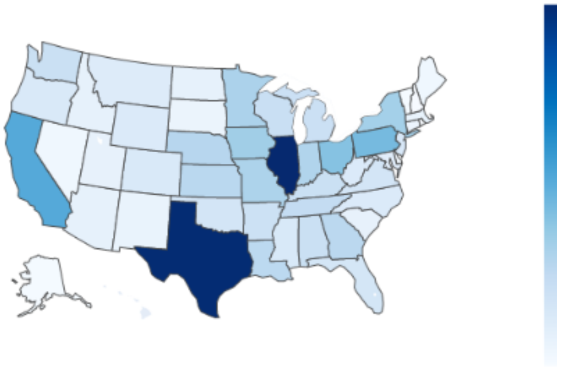

Figure 1

2. Distribution of train speed for different accident types: *Highway rail crossing accidents seem to have maximum speed whereas side collision has lowest speed.*

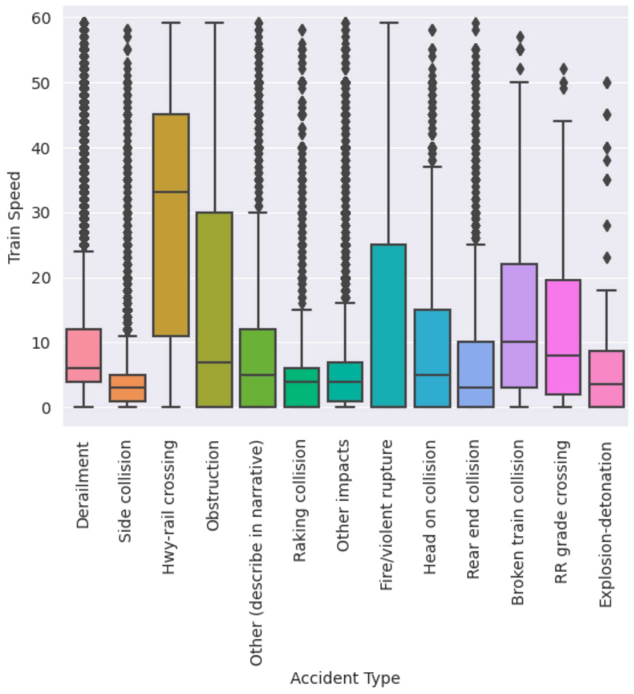

Figure 2

For more plots, refer the notebook.[2]

Total Cost Regression

The results from the SVM model were very poor, but the GB regressor was able to reasonably predict the total cost. The best results achieved after some tuning was a R2 of 0.764 on the training set, with R2 of 0.651 and RMSE of 315260 on the test set. Note that the RMSE seems very high, but it is in units of dollars, and the dataset has accidents costing in the tens of millions. Additionally, we used two methods to extract the feature importances from the base learners. These are shown below in Figure 3, and show that the important features are reasonable.

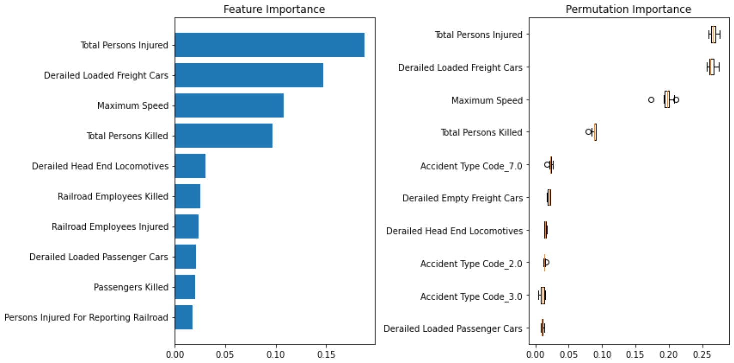

Figure 3

Primary Cause Code Classification

We were able to achieve reasonable results on predicting the first level of the accident cause code hierarchy. With the GB classifier, we achieved training accuracy of

0.663, with 0.641 accuracy and 0.63 F1 on the test set. The LR model achieved similar results, and a full breakdown of the results for each class can be found in the appendix. The confusion matrix for these top-level classes is shown below in Figure 4.

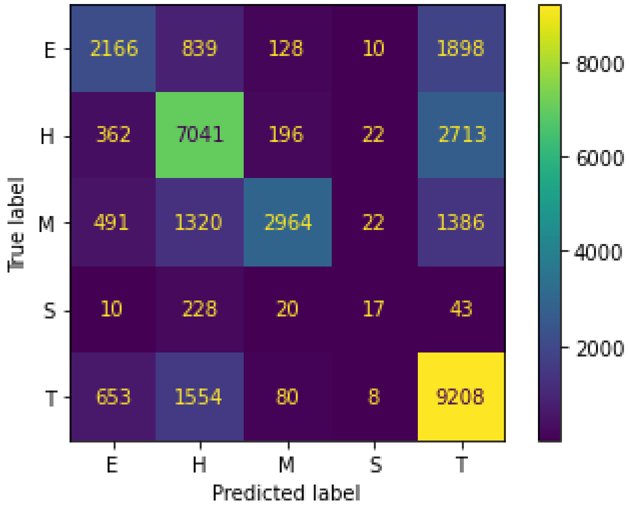

Figure 4

We were not able to achieve good results on the more granular bucketing of cause codes. This seems to be due to several reasons, including the difficulty of extreme multi-class classification and the power-law distribution of cause codes, with a small number of codes dominating the dataset. I think that a more sophisticated technique would be necessary for this task.

Time-Series Forecasting

After inspecting the stationarity of the time series through rolling mean and standard deviation, it was apparent that the values fluctuated over time, indicating that the time series was non-stationary (refer to Figure 5). Consequently, we utilized the first difference approach to transform the time series into a stationary one (see Figure 6).

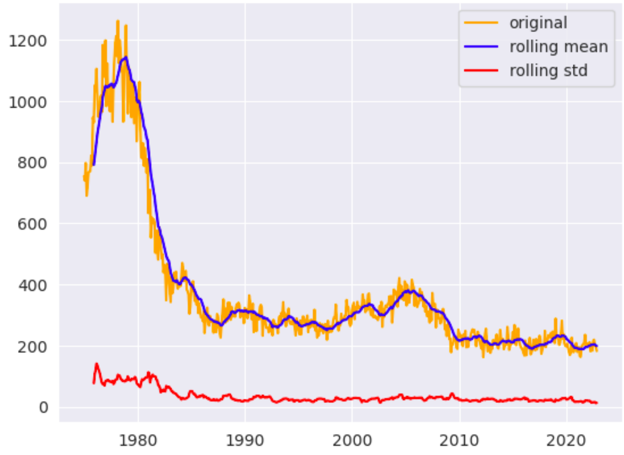

Figure 5

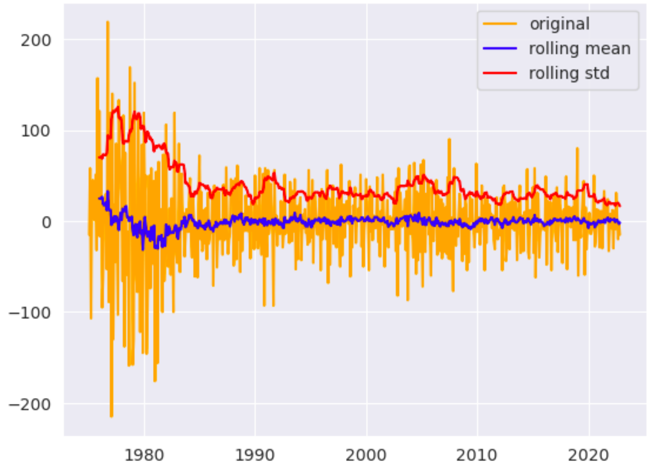

Figure 6

We utilized AIC and BIC scores to determine the model's order. The model containing two auto-regressive terms and three moving average terms was selected.

$$R_t = \mu + \sum_{i=1}^{2} \phi_i R_{t-i} + \epsilon_t + \sum_{j=1}^{3} \theta_j R_{t-j}$$

Then, finally we fit the ARMA(2, 3) model and used it to forecast future values (refer Figure 7).

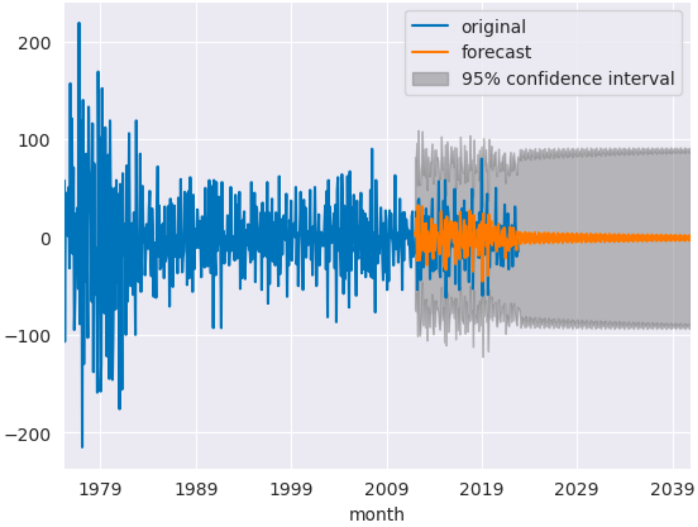

Figure 7

IMPACT

- Increased safety: Based on our findings, rail operators can take steps to minimize accident occurrence and in turn improve overall safety.
- Identifying risk factors: Help allocate their resources more effectively and prioritize

investments in safety measures and training programs where they are most needed.

- Prevention: Appropriate measures can be taken based on results obtained to save lives, prevent injuries, and reduce damage to equipment and infrastructure.
- Better public relations: Railroad accidents can be highly visible and damaging to the government's reputation and hence demonstrating their commitment to safety will improve its public image.

CONCLUSION

During the project's duration, we were able to accomplish the significant milestones that we had set and projected to finish. Our achievements included the development and assessment of models designed to classify accident cause codes, predict accident costs, and forecast accident frequency.

From the cost regression modeling, we were able to reasonably predict the total cost of an accident, given a set of categorical and numerical features. Additionally we were able to extract the most significant features in predicting the cost, which included features such as number of people injured and the number of derailed cars. However, the total cost is made up of several categories, specifically equipment and track damage. Interestingly, there is no data on some of the downstream costs such as lawsuits and disability pay. Given more time, it would be worthwhile to try extracting these costs and building models to predict the costs of each of these factors separately.

For the accident cause classification, we were able to predict the top-level of the hierarchy fairly well, but due to the size of the categorical space, our methodology did not scale well when predicting the more specific codes. A logical next step would be to perform a binary classification for each code at each level. For instance, we would predict, for each of the top-level digits, the probability of the accident having that cause. We would then take the most likely cause, and predict the second digit, then the third, and finally the fourth. However, this was simply out of the scope of this project, and would require more work.

Due to the limited time and scope of the project, a simple ARMA model was developed to forecast future railroad accident frequency and could not be further explored in-depth as it was not the main focus. Nevertheless, it was interesting to work on and there is potential for exploring and contrasting multiple time-series forecasting models in the future.

REFERENCES

1. Kaggle: Railroad Accident & Incident Data, https://www.kaggle.com/datasets/chrico03/railroad-accident-and-incident-data
2. Code: EDA and Time-Series Forecasting, https://www.kaggle.com/anshulrao/rail-accident-and-incident-data
3. Code: Regression and Classification, https://drive.google.com/file/d/12EA-VKx9Ig0FuSoO6F0-Q9Oi0CBFoRST/view?usp=share_link

APPENDIX

Classification Report (GB)

```
          precision    recall  f1-score   support

       E       0.59      0.43      0.50      5041
       H       0.64      0.68      0.66     10334
       M       0.87      0.48      0.62      6183
       S       0.22      0.05      0.09       318
       T       0.60      0.80      0.69     11503

accuracy                           0.64     33379
   macro avg       0.58      0.49      0.51     33379
weighted avg       0.66      0.64      0.63     33379
```

Classification Report (LR)

```
          precision    recall  f1-score   support

       E       0.55      0.40      0.46      5041
       H       0.64      0.63      0.63     10334
       M       0.89      0.45      0.59      6183
       S       0.33      0.00      0.01       318
       T       0.57      0.81      0.67     11503

accuracy                           0.62     33379
   macro avg       0.59      0.46      0.47     33379
weighted avg       0.64      0.62      0.61     33379
```