Analysis of the Effects of Regulation on Railroad Safety

Implementation of Data Mining Algorithms to explore Causality and Trends in Railroad Accidents

Kai Liao
Electrical Engineering
University of Colorado Boulder
Broomfield, CO, USA
feli3513@colorado.edu

Katrina Siegfried
Computer Science
University of Colorado Boulder
Boulder, CO, USA
katrina.siegfried@colorado.edu

Andrew Smith
Mechanical Engineering
University of Colorado Boulder
Boulder, CO, USA
ansm6548@colorado.edu

1 Problem Statement/Motivation

Accidents at highway-railroad intersections cause tremendous losses of lives and resources. For example, on June 27th, 2022, an Amtrak passenger train struck a dump truck in rural Missouri. The truck was crossing a passive intersection with no crossing bars, lights, or bells. Several trains and locomotives derailed causing the death of 4 individuals and over 150 injuries. It is estimated that there are over 130,000 passive railroad crossings in the US. The implementation of active restraints on an intersection like the one described in the accident would cost around $400,000 [1]. Due to the loss of life and high expenses of derailment, this case has again brought up debate about whether the investment in railroad restraints or alteration to regulations between highways and railroads is needed.

Regulation has been one avenue of effort to minimize the number of accidents between highways and railroads. Governments have invested in putting barricades between the intersections during crossings, adding signs, and multiple types of indicators. These are improvements to the intersections themselves. However, certain regulations have also been geared to improving the safety of trains themselves.

The Rail Safety Improvement act of 2008 mandated the implementation of Positive Train Control (PTC) systems on Class 1 railroads and all main lines over which intercity or commuter rail passenger transportation is provided. PTC systems attempt to automatically reduce accidents by only permitting a train to move if it has positive authority to do so. This contrasts with typical train operation in which a train has authority to move unless given a stop signal. This system is intended to prevent head-to-head collisions and prevent trains from going into control or restricted zones to potentially avoid collisions.

Beyond regulations, infrastructure is also an important aspect of railroad operations. Trains are designed to carry various cargoes through various locations. The development, zoning, and transportation schedules of each location has a large impact on the risk of accidents. Transport routes are determined by the resource requirements of each location, and if a location has high need for transport, it may cause conflicts between rail and highway transportation systems that result in higher risk of crashes. Such conflicts could be diminished after they are discovered by careful civil planning, such as rezoning and rerouting.

There has been much work in creating physical barriers and auditory and visual warning signals to alert vehicle drivers of an oncoming train. However, these systems are costly to implement and are sometimes viewed as irritating to neighbors of railroads. Therefore, it is important to understand the degree of efficacy of these measures. Additionally, there are unique topological attributes of the interaction between the highway and railroad that could potentially create a higher likelihood of accidents. These characteristics also could be changed through modifications which could increase safety at the expense of construction projects.

This project aims to consider the impacts of new regulations, locations of intersections, and the characteristics/topography of intersections to determine which features promote safety and which features do not.

2 Literature Survey

The Federal Railroad Administration attempted to develop a model for predicting accidents and their severity using data mining techniques using statistical analysis [2]. Though this work does not inherently use data mining techniques, it has relevance due to its exploration of the same data set and the ways it chose to select data and include variables. It also developed a frequency variable called exposure that normalized intersections based upon the amount of exposure to accidents they have. This statistical analysis can also be used as a baseline for the results we discover in this paper to compare to.

Data mining techniques have been successfully applied to investigate the FRA dataset as well as similar datasets generated in other countries. Liu et al utilized chi-squared analysis to look at the causes of rail accidents in the FRA dataset from 2001 to 2010 and the effects of those causes on accident rates [3]. A similar dataset has been generated by the Iranian Railway (RAI) which utilized association rules to identify if-then relations between rail accidents and their potential causes [4]. A survey paper by Bala et al gives a good overview of the literature as it relates to data mining of rail accident data sets [5]. A similar paper by Lu et al compares various generalized linear and data mining models for crash prediction and compares their effectiveness using a test dataset [6]. After a thorough review of rail accident data mining research, it has been determined that none have considered use of a random forest classification or frequent pattern growth (FP-Growth) in their analyses.

In addition to rail, data mining has been utilized in similar large datasets composed of accident reports, mainly around the topic of occupational safety. Several variations of frequent pattern generation including temporal, elevated severity, and high impact were performed by Singh et al using a proprietary data set generated from a steel manufacturing plant in India [7]. Another group led by Khosrowabadi used association rules and K-means clustering to identify the factors affecting occupational safety in industrial paint halls in Tehran [8]. These additional studies help demonstrate that more advanced frequent pattern techniques and classification techniques are promising future areas of investigation for accident analysis.

3 Proposed Work

General cleaning will be applied across the entire dataset prior to any specific preprocessing work for each of the specific investigatory questions. General cleanup of misspellings will be corrected, as well as handling duplicate encodings for attributes and duplicate entries. Missing data will be addressed with either identical values or inferred data where possible as dictated by the nature of the attribute. Where needed, categorical attributes will be numerically encoded. Expert selections to exclude attributes which are clearly unrelated or too noisy or sparse to add value will be performed if required.

3.1 Proposed Work for Crossing Location

Location identifiers and branch/segment information are varied and will need to be one-hot encoded for mining data access. Absolute coordinates can either be estimated or actual, and will be smoothed to reach an equal amount of precision among all data values. Multiple “1 = Yes, 2 = No” attributes will be modified to “0 = No, 1 = Yes” for standardized binary representation. Values that represent “N/A” will be standardized to “-1” as this value is already used to represent “N/A” for ID values in the dataset.

3.2 Proposed Work for PTC Analysis

In the investigation of PTC implementation, to reduce the number of attributes and improve final pattern discovery and classification, concept hierarchies will be identified to mine at different abstraction levels. Data values will be selectively smoothed and discretized iteratively as required to improve results. One-hot encoding and a vertical reformatting will be performed for encoded enumerated attributes if model run time proves to be too slow.

3.3 Specific Work for Intersection Study

Pre-processing the data for the study on the effects of the intersection characteristics will require the crossing surface age to be calculated from the difference of the installation date and the date of accident. There are many attributes with ID's that will need to be changed to one-hot encoding for use in the FP-growth algorithm. The binary recording system for multiple attributes will also be changed from 2 means no to 0 means no. This will allow for uniformity within the data set.

For the sake of focusing this study on intersection characteristics, all attributes that are specific to the location, train characteristics, or datasheet administrative categories will be removed prior to implementing the algorithms.

After implementing the cleaning and pre-processing measures, the modified data will be implemented into both Random Forest and FP-growth algorithms and will be evaluated for its performance.

4 Data Set

The chosen data set was selected based on the large number of papers on railroad safety implementing data mining techniques using different categories within the Federal Railroad Administration (FRA) Office of Safety Analysis' database. There are multiple different tables based upon different reporting forms. The FRA requires the reporting of accidents and fatalities using specific forms as defined by the circumstance. The specific grouping of data that was chosen was all the accidents between railroads and highways due to the many variances in causalities that it provides. The data set contains all the reported information from 1970 to May 2022.

There are 186 attributes within the data set allowing for a wealth of potential factors of causality to be explored. There are 436,498 rows of data, or accidents, during the period and there are a total of 42,567,011 non-empty entries within the data set.

The attributes provide comprehensive data about railroad operations at the time and location of each data object representing an individual incident at a highway-railroad crossing.

Attributes include the time of incident, the number and types of advance warnings, and the activation of other warning systems – painting a clear picture of when the incident occurred and what was done to mitigate it.

The attributes document absolute and geographic locations of incidents, including global coordinates and the city, highway, and railroad line in which each incident occurs. Information is provided regarding land use and typical operating information at the location of the incident, including throughput and average passenger count.

Information about the exact characteristics of the intersection of incident is recorded, including rail design, road design, illumination, crossing angle, and surface materials.

The wealth of information in this dataset allows many different relations to be mined regarding temporal and spatial influences on incident risk.

All the group members have successfully downloaded the data set at the URL below:

https://safetydata.fra.dot.gov/OfficeofSafety/publicsite/DownloadCrossingInventoryData.aspx

5 Evaluation Methods

Evaluation of each of the questions will be performed using two classic data mining methods for classification – Decision Tree classification and FP-Growth classification. The goal of these methods is to yield results which can easily be interpreted to generate clear action plans to reduce future rail accidents.

Frequent pattern growth or FP-Growth is an effective and efficient method of finding frequent patterns in very large data sets. These frequent patterns will yield a sequence of attributes that are related to outcomes of interest for each question, and the output is easily interpreted to provide insight into the factors most associated with the outcomes, so recommendations could be devised to improve future outcomes. FP-Growth addresses the large memory limitation required by the Apriori algorithm because it maintains a tree rather than generating a list of all candidates, and it also can be parallelized by partitioning the database. (The pyspark Python library by Apache Spark contains a pre-built function for implementing FP-Growth. If the model run time proves to be slow even with parallelization, a vertical data format can be explored. Evaluation of the model will be performed using a selection of the following metrics dependent on the performance on the data: support, lift, confidence, X^2, Kulczynski measure, and cosine measure. Thresholds for these measures are still to be determined.

In addition to FP-Growth, a Random Forest Decision Tree approach will be used to generate a classification around each question's target label, after which rules will be extracted. The rules will provide and easily interpreted understanding of the potential cause and known effect which can be communicated to industry experts to improve future rail accident outcomes. The Random Forest will be generated using an 80/20 test/train split with sampling without replacement using the built-in sklearn python library. Like the FP-Growth implementation, this decision tree implementation can also be parallelized. Metrics of accuracy, sensitivity, precision, specificity, F1, and Fb will be used to evaluate performance using k-fold cross-validation as per industry standard.

6 Tools

It was decided to implement the data mining methods proposed in this paper using built-in Python toolboxes to simplify the work and to learn commonly used approaches to these problems.

The Python *pandas* toolbox will be used for data cleaning. The *unique()* function can be used to determine all of the unique values in string attributes such as incident descriptions and locations and determine if there are any misspelled instances or instances that should be combined. The functions *isna()* and *isnull()* can be used to determine missing values and in conjunction with the *unique()* can find all forms of null cells.

The Python *numpy* toolbox will be used to transform the data so that it is easier to manage and manipulate.

The Python toolboxes, *sklearn* and *mlxtend*, will be used to alter attributes to one-hot encoding to prepare the data for the machine learning algorithms. It can also be used to split the data into training and testing groups to evaluate the performance of the methods. Finally, it can be used for performing the random tree method by using the functionalities for the Random Forest techniques.

The Python toolboxes, *pyspark* and *mlxtend*, will be used to mine frequent itemsets using the FP-growth algorithm. This toolbox could also be used to perform Pearson's independence test and the correlation for each attribute.

Other tools could be implemented if needs arise.

7 Milestones Completed

Data Cleaning: Since there was interest in knowing if there had been significant reduction in train accident severity after the implementation of PTC, there was a need to separate the month, year, and date from the single date and time entry. A function was created to separate the day, month, and year from the combined date entry that includes all time components in one string. This allowed for individual months, years, and days of the month to be queried, to determine if there has been any significant change in accident severity by month or over time.

There were also text entries that contained geographic information needed that were indicative of the same place but entered the dataset in different ways. To reduce this redundancy, a function was created to abbreviate common text entries, such as address terms like street and avenue, to minimize the number of unique entries. To further reduce redundancies, a function was made to remove all spaces, periods, dashes, and that capitalized all text. This helped to further reduce the number of unique entries and to reduce repeated entries with slight variations. Finally, there was a need to consolidate all the different null space indicators into a single, consistent one. To do this, a function was made that replaced blank entries with "NA" while applying the same text rules above so that there were not multiple ways in which null information was stored.

Data Preprocessing: The dataset needed to be preprocessed in slightly varying ways for each of the three interesting questions this report attempts to address. This was to account for the unique lists of attributes used for each question.

Location data included the state, county, and city of each incident as well as location-relevant information such as weather and visibility, which correlates to local climate. For state, county, and city, the dataset already contained label-encoded values for each which are usable for mining. Some data objects were missing records for city, with the implication that the incident did not happen within city limits. For data preprocessing, these blanks were consolidated into one "No City" value so that we can find patterns among incidents occurring outside of city limits. Values like weather and visibility are label-encoded in the dataset, but are one-hot encoded for data mining purposes due to the compact set of unique values.

In the investigation of the effects of PTC on train accidents, in addition to the cleaning already performed, functions were written to remove previously encoded values which were already represented by categorical data. Further, all text was casted to lower case and transcoded into a one-hot sparse matrix. Specific redundant and unnecessary attributes were removed such as data related to who filed a report and when a report was filed, and multiple encodings of the date or location.

To address how the characteristics of the intersection architecture and design impact the probability of death and injury from an accident, a particular list of related attributes was generated. Attributes that were included the allowed speed of the train, illumination of the crossing, view obstructions, warnings, signals, and type of intersection. For each of these attributes, some entries needed to be cleaned including values that did not stand for any codes or were mistyped values. These were replaced by finding all the unique values and their frequency. If the frequency was low enough, the entry was looked for in the dataset and identified if it was an error. If it was an error, it was replaced by another value. Most of these were assigned to the corresponding unknown value. These attributes were transformed using one-hot encoding from numerical codes to binary options for each of the given choices.

Finally, the dataset needed to be split for training and testing to effectively analyze the performance characteristics of the data mining attempts. For this, the data was split into training set containing 80% of the data and a testing set containing 20% of the data.

8 Milestones To-Do

August 3rd: Gather all the results and generate evaluations of performance for each question. Present results within team to ensure they make sense. Begin to write the final report and presentation.

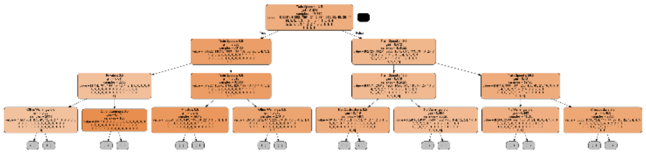
August 8th: First draft of final paper and presentation should be completed.

August 8th: Hold final meeting to go over paper and record presentation.

August 9th: Finalize any updates and edit video. Agree to submit all materials and conduct peer evaluations.

9 Results So Far

Preliminary decision tree induction was performed for the characteristics of the intersection. Most of the early decision points were based upon the speed of the train at the time of the accident. This may suggest that a large potential legislative point could be to restrict the speed of trains through certain intersections.



A first pass for the FPGrowth algorithm was implemented and the code pushed to github. Due to the large number of attributes and data points (even after cleaning) and the increased memory consumption due to one-hot encoding for all values, the initial implementation using the *mlxtend* library resulted in OOM errors. To improve this implementation the data input needs to either be batched or done in parallel. The *pyspark* library also has a FPGrowth algorithm implementation that is inherently parallel, and should be explored before implementation of batching for data input. Additionally, a vertical input may be explored to expedite running time for the FPGrowth algorithm.

Additionally, a second more curated version of the dataset has been found which incorporates not only the basic catalog of the initial dataset but has incorporated other queried data from the FRA database. This dataset provides information related to accident severity that is lacking in the initial dataset. The dataset is currently under review by the group and can be found at:

https://catalog.data.gov/dataset/highway-rail-grade-crossing-accident-data

REFERENCES

[1] Zoe Christen Jones. At least 4 dead, dozens injured after Amtrak train derails in Missouri. *CBS News*, 27 Jun. 2022, https://www.cbsnews.com/news/amtrak-train-derailment-mendon-missouri/

[2] Daniel Brod and David Gillen, Oct. 2020. New Model for Highway-Rail Grade Crossing Accident Prediction and Severity. *U.S. Department of Transportation*, https://railroads.dot.gov/sites/fra.dot.gov/files/2020-12/APS-A.pdf

[3] Ziang Liu, M. Rapik Saat, and Christopher P.L. Barkan. Analysis of Causes of Major Train Derailment and Their Effect on Accident Rates. Transportation Research Record: Journal of the Transportation Research Board, Vol. 2289, No. 1 (2012), 154–163. DOI: //doi.org/10.3141/2289-20.

[4] Ahmad Mirabadi and Sharifian Shabnam. Application of Association Rules in Iranian Railways (RAI) Accident Data Analysis. Safety Science, Vol. 48, No. 10 (2010), 1427–1435. DOI: https://doi.org/10.1016/j.ssci.2010.06.006.

[5] Manju Bala and Bhasin Anshu. A Review on Analysis of Railway Traffic Accident with Data Mining Techniques. International Journal of Computer Sciences and Engineering, Vol. 6, No. 6 (June 2018), 1251–1256, DOI: https://doi.org/10.26438.

[6] Pan Lu, Denver Tolliver, and Zijian Zheng. 2018. *Highway-Rail Grade Crossing Traffic Hazard Forecasting Model.* Mountain-Plains Consortium Technical Report MPC 18-354. North Dakota State University, Fargo, ND.

[7] Kritika Singh and J Maiti. A Novel Data Mining Approach for Analysis of Accident Paths and Performance Assessment of Risk Control Systems. Reliability Engineering & System Safety, Vol. 202 (2020), 107041. DOI: https://doi.org/10.1016/j.ress.2020.107041.

[8] Naghmeh Khosrowabadi and Ghousi Rouzbeh. Decision Support Approach on Occupational Safety Using Data Mining. International Journal of Industrial Engineering & Production Research, Vol. 30, No. 2 (June 2019), 149–164. DOI: https://doi.org/10.22068.