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

ABSTRACT

There have been numerous attempts at utilizing regulations to increase the safety of highway-rail intersections at the cost of increased expenses and labor upkeep. This paper attempts to identify what regulation dictating the specific characteristics of the intersection is useful and how geography and weather might affect the need for regulation. The paper addresses this topic by the utilization of data mining algorithms on a comprehensive data set containing over 240,000 accidents between railroads and highways as maintained by the Federal Railroad Administration. This paper presents three approaches to these questions including the use of decision tree classification, frequent pattern growth, and k-means clustering. The decision tree classification method developed a model with over 71% accuracy of predicting the number of injuries based upon characteristics of an intersection. Frequent patterns revealed associations among incident mortality and casualty related to train car position, as well as an association between location and track type, and a weak association between intersection signals and accident casualty. K-means clustering created a model with a mean 0.49 silhouette score which revealed that locational climate-related inclement weather was present at many crashes. These results not only develop models for these individual questions but could be used as indicators as to where future regulation is needed or not. Statistical analysis could be paired with the trends identified in this paper to determine significance of certain attributes affecting the severity of injuries. This could led to more efficient allocation of resources and a reduction in injuries during future accidents.

1 Introduction

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 many new safety systems and improvements.

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 effectivity 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 Related Work

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 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 159 attributes within the data set allowing for a wealth of potential factors of causality to be explored. There are 242,021 rows of data, or accidents, during the period and there are a total of 25,448,378 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. Additionally, there is a record of the number of injuries and fatalities that occurred in total, on the train, and within the vehicle hit.

The attributes document geographic locations of incidents, including the city, highway, and railroad line in which each incident occurs. Information is provided regarding land use and the weather conditions that were experienced during the specific incident.

Information about the characteristics of the intersection of incident is recorded, including rail design, road design, illumination, visual obstructions, 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.

The data set can be found at the URL below:

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

4 Main Techniques Applied

The data set contained many rows of unfilled information, several attributes that included textual information, and many coded options for characteristics regarding the intersection. Due to this, cleaning and pre-processing of the data was needed before trying to apply data mining algorithms.

Various mining techniques were chosen to allow for knowledge to be gained through differing perspectives and to be analyzed in such a way as to impact future regulatory decisions. For example, decision tree techniques were applied to be able to understand if certain attributes of intersections would increase the probability of many injuries during railroad accidents. Clustering techniques were applied to understand the effect of geography and weather on the severity of railroad accidents.

Finally, frequent itemset generation was implemented to attempt to understand associations and interesting patterns between attributes that result in higher numbers of injuries in support of both the intersection characteristics and location characteristics and to reveal other potentially interesting patterns related to accident severity. 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. To extract patterns of interest, interesting attributes, called consequents, were used to filter from the list of all patterns to identify those the researchers deemed interesting.

These techniques combined aimed to paint an overall picture of the contributing factors of high-injury accidents to drive future regulation decisions.

4.1 Data Cleaning

There were 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 cleaning in support of the frequent itemset generation involved removing duplicated information in pre-coded attributes and attributes without any data were removed to expedite processing speed and memory consumption. Null values were populated via the mode or mean for numerical attributes, whichever was deemed most appropriate for the attribute. For categorical data, the null values were replaced with the most frequent item in the attribute.

4.2 Data Preprocessing

The dataset needed to be preprocessed in slightly varying ways for each of the three techniques used in this report while attempting to address the two interesting questions. 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 patterns could be found 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.

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.

4.3 Techniques Used for Crossing Location

To prepare crossing locations for visualization, latitude and longitude had to be determined from the county FIPS code data provided. Each FIPS code corresponds to one US county, and a database with FIPS codes and county centroids was used to estimate the location of each crash. Using the longitude and latitude values allowed a K-means clustering based on temperature and weather to be visualized by plotting each data point by its coordinates.

4.4 Techniques Used for Intersection Study

Pre-processing the data for the study on the effects of the intersection characteristics required the crossing surface age to be calculated from the difference of the installation date and the date of accident. There were many attributes with ID's that needed 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.

4.5 Techniques Used for Frequent Pattern Generation

For frequent pattern generation, continuous numerical data was binned into at most 10 bins which were mapped to a series of strings to make the results easier to interpret. The data was then pivoted from horizontal to vertical format to increase the processing and memory consumption of the algorithm used.

4.6 Evaluation Methods

Evaluation of each of the questions was performed using three classic data mining methods for classification – Decision Tree classification, K-means clustering, and FP-Growth classification. The goal of these methods was to yield results which could easily be interpreted to generate clear action plans to reduce future rail accidents.

Evaluation of the FP Growth model was performed using a selection of the following metrics dependent on the performance on the data: support, lift, and confidence. Thresholds for these measures were empirically determined after evaluating the full list of results.

In addition to FP-Growth, a Random Forest Decision Tree approach was used to generate a classification around each question's target label, after which rules were extracted. The rules will provided an easily interpreted understanding of the potential cause and known effect which will be communicated to industry experts to improve future rail accident outcomes. The Random Forest was generated using an 80/20 train/test split with sampling without replacement using the built-in sklearn python library. Like the FP-Growth implementation, this decision tree implementation was 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.

For K-means clustering, the primary evaluation method was the silhouette score, which evaluates a combination of inter- and intra- cluster similarity.

4.7 Tools Used

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

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

The Python toolbox, *sklearn*, was used to alter attributes to one-hot encoding to prepare the data for the machine learning algorithms. It was also used to split the data into training and testing groups to evaluate the performance of the random forest and k-means methods. Finally, was used for implementing the random forest method by using the functionalities for the Random Forest techniques.

The Python toolbox, *pyspark*, was used to mine frequent itemsets using the FP-growth algorithm. Given the different implementation for data structures and processing, this library was also used in each step of the cleaning and preprocessing for the frequent itemset generation as the objects were of a different object type required for the parallelized implementation. A different implementation from the

python *mlxtend* library was initially explored, but the implementation could not handle the large volumes of data in this dataset.

For visualization, the Python *matpyplot* toolbox was used to generate a scatterplot for the K-means clustered data.

5 **Key Results**

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.

The frequent pattern generation yielded mostly intuitive results, with relationships between high mortality and casualty linked to accidents involving the front of a train, while location-based analysis showed frequent patterns with city and track type, and intersection characteristics weakly showed mortality and casualty were lower at intersections that involved more safety signaling.

The visualization for K-means clustering showed that the two biggest clusters of crashes, based on temperature and weather, were under conditions of freezing snow in the north and cold rain in the south. Two clusters were found representing cold sun and warm sun, both of which contained few members. The model indicates that inclement weather, dependent on local climate, may be a contributing factor to many highway-rail crashes.

5.1 **Intersection Characteristic Results**

Before the algorithm was implemented, a basic exploration of the data set was conducted to understand the frequency of different characteristics. Figures 1, 2, and 3 depict various frequency charts developed for attributes that describe intersection characteristics that could potentially drive the increase in severity of accidents. They also help to develop an understanding for the general types of intersections observed from the data set.

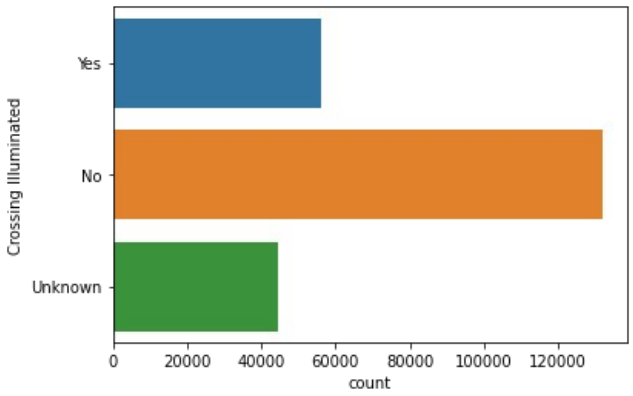

Figure 1. Frequency chart of number of intersections that were illuminated during incident.

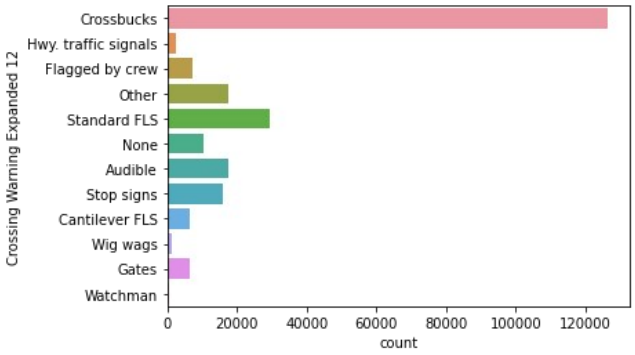

Figure 2. Frequency chart of number of intersections that used each style of crossing warning during the incident.

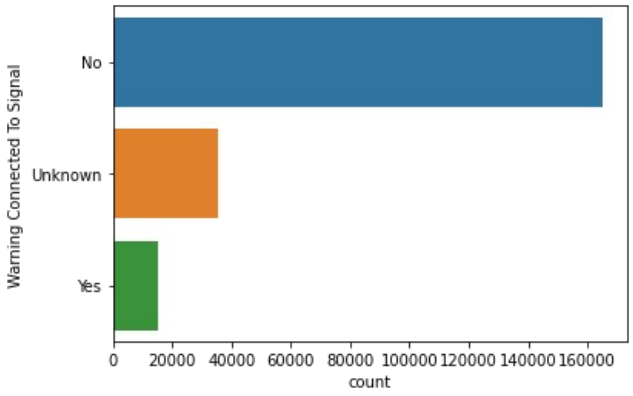

Figure 3. Frequency chart of number of intersections that had the warning connected to the signal.

The decision tree type classification algorithm was applied to attributes describing the characteristics of the railroad-highway intersection. The classification

algorithm used a target variable of total injured as the means of sorting. There were 4 classification categories to sort between including 0, 1, 2, or 3 or more injuries.

The algorithm was applied as both a decision tree and random forest algorithm. The attributes were modified to either include instances of unknown information or to exclude them.

Initially a decision tree algorithm was implemented to develop an image of the decision tree to be able visualize the decisions that classified the instances. Using this decision tree, it was possible to estimate the percent of instances that resulted in the number of injuries. These results were used to determine potential courses of action in developing or maintaining regulation regarding railroad-highway intersection design.

A random forest model was created to increase the accuracy of the decision tree model in predicting the number of injuries. Table I shows a confusion matrix with the number of entries for each prediction and each actual classification. Table II shows the performance characteristics including precision, recall, F1 score, and support for each classification category. The accuracy, macro, and weighted averages are depicted for each score category as well.

Table I. Confusion Matrix with Unknown Attributes

		Predicted Class			
		0	**1**	**2**	**3+**
Actual Class	**0**	33867	834	89	34
	1	9889	272	35	9
	2	1924	64	6	0
	3+	851	43	4	0

Table II. Performance Characteristics of Random Forest including Unknowns

	Precision	Recall	F1 Score	Support
0	.7278	.9725	.8326	34824
1	.2242	.0267	.0476	10205
2	.0448	.0030	.0056	1994
3+	.0000	.0000	.0000	898
Accuracy			.7125	47921
Macro Average	.2492	.2505	.2215	47921
Weighted Average	.5785	.7125	.6154	47921

To improve the accuracy of the model, attributes with unknown rows were excluded from the results. The confusion matrix in Table III showed slightly higher correct predictions than the original in Table I. The random forest method led to a slight increase in performance characteristics as shown in Table IV compared to Table II.

Table III. Confusion Matrix without Unknowns

		Predicted Class			
		0	**1**	**2**	**3+**
Actual Class	**0**	33884	826	84	30
	1	9889	277	32	7
	2	1924	63	7	0
	3+	855	38	5	0

Table IV. Performance Characteristics of Random Forest without Unknowns

	Precision	Recall	F1 Score	Support

0	.7279	.9730	.8328	34824
1	.2301	.0271	.0486	10205
2	.0547	.0035	.0066	1994
3+	.0000	.0000	.0000	898
Accuracy			.7130	47921
Macro Average	.2532	.2509	.2220	47921
Weighted Average	.5802	.7130	.6158	47921

5.2 Intersection Location Results

The K-means clustering model with four clusters selected on temperature and weather presented four clusters: chilly (56.5 F) and raining, freezing (22.9 F) and snowing, hot (74.6 F) and clear, and cold (33.2 F) and clear. Freezing and snowing crash conditions were predominantly in the northern US, with the highest density in the Midwest. Chilly and raining crash conditions were predominantly in the southern US, with the highest density in Georgia. Clear weather clusters were not locationally centralized.

K-means clustering used temperature and weather as the selection variables, but longitude and latitude for plot visualization as seen in Figure I. The data showed that climate-related weather conditions strongly impacted crash conditions.

The mean silhouette score was 0.492, indicating that the clustering is fairly compact and valid with some overlap.

5.3 Frequent Pattern Results

Frequent patterns were evaluated in relation to accident severity where severity was indicated as having higher casualty and mortality numbers. FP Growth results showed that accidents with high mortality were most closely associated with high injury and involvement with trains with a low number of cars with support=0.063, lift=1.063, and confidence=0.786, likewise for accidents with high casualties were most closely associated with a low number of train cars with support=0.063, lift=1.062, and confidence=0.786. Logically this makes sense, as circumstances with high mortality likely have high injury and vice versa, and anecdotally, most accidents occur in rail yards where trains are likely to have a lower number of cars.

When examining location, patterns containing the consequent ‘State Name’ were filtered and revealed that the most commonly co-occurring attributes associated with ‘State Name’ were the city and the type of track with support=0.542, lift=1.128, and confidence=0.826. As previously mentioned, many accidents occur on private yard tracks, and when those tracks cross highways – especially in rural areas – it is likely that accidents in each state in this highway database are localized to these areas.

With consequents of ‘Warning Connected To Signal’ and ‘Crossing Illuminated’ related to intersection characteristics, the most commonly co-occurring attributes, though weakly associated, were the mortality and casualty being very low, with support=0.139, lift=1.005, and confidence=0.921 indicating that there is a frequent pattern between highly signaled highway rail crossings and lower accident severity.

6 Applications

The applications found in this report were focused on the context of highway-rail intersection incidents, however, they could similarly be applied to many transportation topics. These applications include the effect of weather and location on the probability and severity of traffic-related incidents.

6.1 Applications of Decision Tree Analysis

Upon deeper investigation into the decision tree developed by the algorithm, certain attributes seem to have a greater causality than others. For example, 3.43% of accidents in which the train was moving 60 MPH or faster resulted in 3 or more injuries. Overall,

1.87% of accidents resulted in 3 or more injuries. Diving deeper, we find that if the view of the intersection is obstructed and the train is between 60 and 70 MPH, it raises the probability of an accident with 3 or more injuries to 4.92%. If those same traits are paired with a lack of illumination, the probability raises to 5.21%.

These results might suggest that train speeds should be limited to below 60 MPH through intersections with highways. Additionally, efforts to remove obstructions from intersections and to illuminate the intersections could reduce the likelihood of injury. Statistical analysis could be conducted to see if the reduction of injuries would be significant or if it would be worth the financial investment.

More importantly, we have created a model to be able to predict the number of injuries expected for an incident between a vehicle and train given the characteristics of the intersection. This is a powerful tool that could be used to analyze specific intersections to understand the degree of risk it has. This can also be used for emergency mitigation plans or as a tool to flag high risk intersections.

Finally, this model could be used to understand where to invest resources to decrease the likelihood of numerous injuries during incidents. It could be used in conjunction with expert opinion to provide suggestions for improvement.

These applications could bring about intelligent improvements in intersection infrastructure. It could help distribute resources in a planned way and could also help determine varying methods to reduce incident severity. Intersections could have differing combinations of warnings and signals that could have a similar result in reducing the number of injuries experienced during accidents.

6.2 Applications of Intersection Location

K-means clustering on intersection location and climate showed that the largest clusters of crashes occurred in inclement weather: freezing snow in the north and cold rain in the south. This model indicates that more work may need to be done to improve safety measures in inclement weather all across the US, with a focus on blizzards in the north and rainstorms in the south. Temperature alone did not appear to significantly increase crash risk, as both hot and cold clear weather crashes were widely but thinly distributed.

6.3 Applications of Frequent Pattern

Given the results from frequent pattern generation for accident severity, it could be recommended that further investigation be pursued related to the factor of train car position involvement and current safety regulations in the pursuit of reducing accident casualties and mortality. It is possible that there are types of accidents such as head on and track obstacles that occur that are not fully prevented using the current safety mechanisms (if any are present at the intersection). Given the relationship between location and city and track types, specific investigation could be performed to support local regulations or mitigations related to specific problematic intersections, such as a rail yard track that crosses a highway in a particular location. With intersection safety characteristics relating to crossing illumination and warning connected to signal, further work could be done to verify these results and support broader regulation regarding improving signaling at intersections between highways and railways. This aligns with the FRA's 2008 legislation seeking to improve crossing safety on class I lines.

Additionally, future work on frequent pattern mining for US highway railway accidents can be improved by investigating feature scaling, more complex coding of null values based on incident type, and identifying and addressing outliers. This additional analysis could improve the strength of patterns as severe accidents are, thankfully, quite rare.

7 Visualizations

The decision trees were plotted using the *graphviz* library and saving it as an image file. The decision tree not only depicts the decision node criteria but also the number of values for each category at each node. This helps to give a visual understanding of

how the tree algorithm is categorizing the data. The data was sorted into 11,658 leaves and has a depth of 34. Due to the limitations of human reception, small sections of the decision tree at limited depths were needed to process the tree. Figure 4 depicts such a portion of the decision tree.

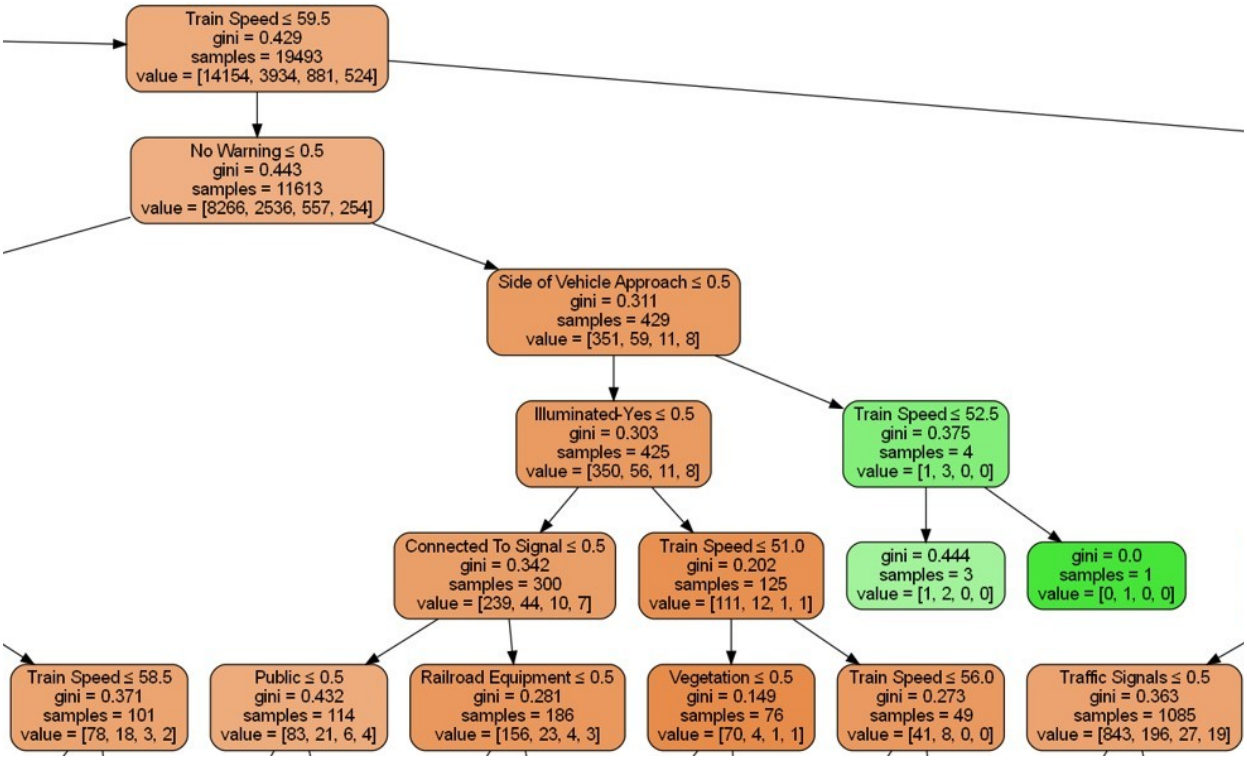

Figure 4. Portion of the decision tree generated from the intersection characterization study showing the decision nodes based upon the Gini impurity index. There were four injury categories in which the tree was sorting into, 0, 1, 2, and 3 or more injuries. The tree allows for the proportion of each of the categories in each node to be calculated.

Matplotlib was used to visualize the K-means clustering information.

Cluster Means:

Red: Temperature 74.6 F, Weather 1.1

Green: Temperature 33.2 F, Weather 1.2

Blue: Temperature 56.6 F, Weather 2.6

Yellow: Temperature 22.9 F, Weather 5.8

Weather Codes:

1 = Clear

2 = Cloudy

3 = Rain

4 = Fog

5 = Sleet

6 = Snow

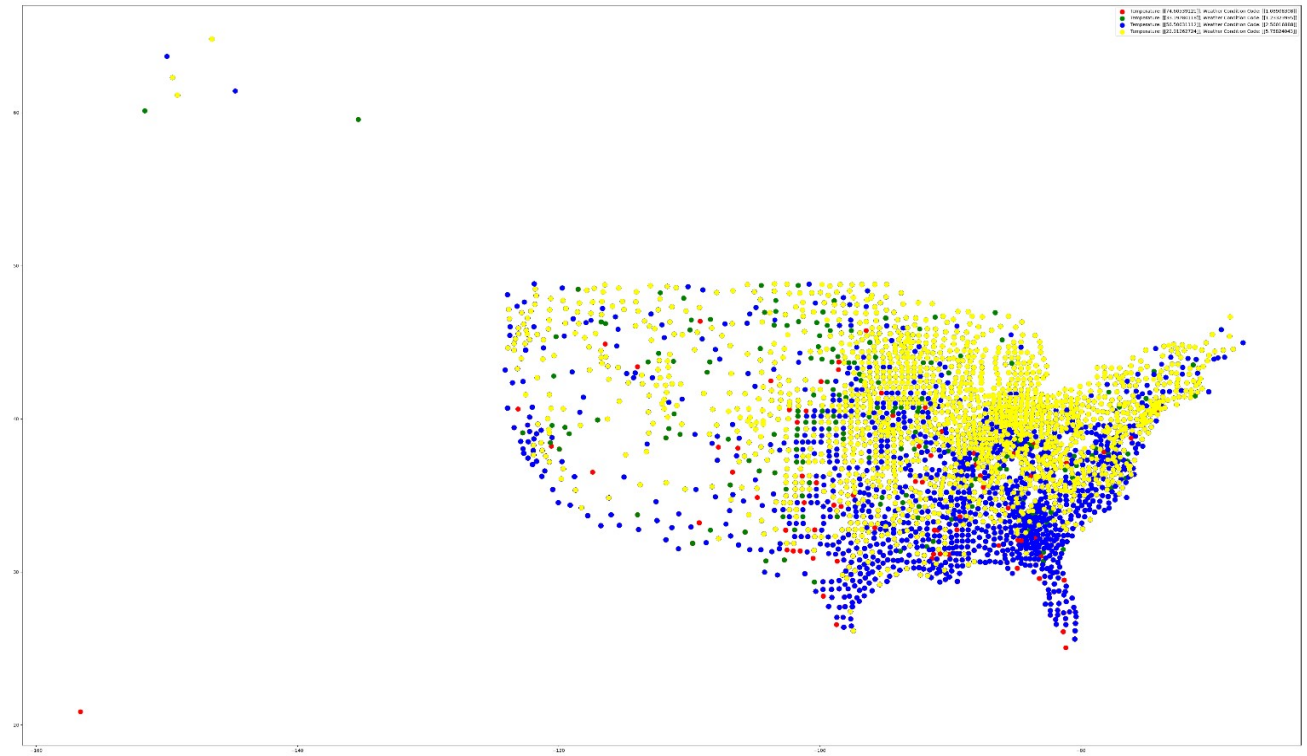

Figure 5. The results of the K-Means clustering for location-based attributes as they relate to weather associated with the incident. As indicated in the results section, incidents associated with snowy weather tended to occur in the northern US, while incidents associated with chilly and rainy conditions primarily occurred in the southern US.

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.