Analysis of the Effects of Regulation on Railroad Safety

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

FirstName Surname†  
 Department Name  
 Institution/University Name  
 City State Country  
 email@email.com

FirstName Surname  
 Department Name  
 Institution/University Name  
 City State Country  
 [email@email.com](mailto:email@email.com)

Andrew Smith  
 Mechanical Engineering  
 University of Colorado Boulder  
 Boulder, Colorado  
 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.

**Add intro to Location**

**Add intro to intersection characteristics**

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

Add Literature survey.

3 Proposed Work

Add proposed work.

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 data set has interesting attributes such as whether or not the train

**Add interesting information about data set attributes.**

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

We need to decide what evaluation methods we would like to try. (Preferably the same)

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 toolbox, *sklearn*, 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 toolbox, *pyspark*, 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

July 20th: Finish pre-processing data set. Begin implementing and performing data mining algorithms on data set for each of the three questions.

July 27th: Finish data mining and begin to analyze the results and performance of the algorithm. Write Progress Report and submit before August 1st.

August 3rd: Gather all the results and evaluations of performance for each question. Begin to write the final report and presentation.

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

August 9th: Agree to submit all materials and conduct peer evaluations.

REFERENCES

[1] Patricia S. Abril and Robert Plant, 2007. The patent holder's dilemma: Buy, sell, or troll? *Commun. ACM* 50, 1 (Jan, 2007), 36-44. DOI: <https://doi.org/>10.1145/1188913.1188915.

[2] Sten Andler. 1979. Predicate path expressions. In *Proceedings of the 6th. ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages (POPL '79)*. ACM Press, New York, NY, 226-236. DOI:https://doi.org/10.1145/567752.567774

[3] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. DOI:https://doi.org/10.1007/3-540-09237-4.

[4] David Kosiur. 2001. *Understanding Policy-Based Networking* (2nd. ed.). Wiley, New York, NY..

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00