USA Highway Railroad Crossing Accident Analysis & Prediction

Team 3: Aakash Hariharan, Abhilasha Singh, Trisha Singh, Vishal Fulsundar

DATS 6103: Summary Report

Professor: Divya Narula

December 17, 2024

Introduction

Railroad crossings are vital intersections where highways and railroads meet, creating unique challenges for traffic safety. Accidents at these crossings are a significant concern, often resulting in fatalities, severe injuries, and property damage. Our team selected this topic due to the critical need to enhance safety measures and reduce the frequency of such incidents. Despite advancements in transportation infrastructure, railroad crossing accidents remain a persistent risk, particularly in the United States.

The study of US Highway Railroad Crossing Accidents aims to understand the causes, impacts, and prevention strategies for these incidents. This involves analyzing variables such as crossing infrastructure, highway user behavior, train characteristics, environmental conditions, and the effectiveness of warning systems. These factors play a crucial role in identifying risk patterns and implementing targeted safety solutions.

The **US Highway Railroad Crossing Accident Dataset** serves as an essential resource for this analysis. It offers detailed information, including railroad specifications, incident records, geographic locations, crossing types, user behaviors, train attributes, environmental conditions, and warning system performance. By leveraging this dataset, researchers can uncover the root causes of accidents, evaluate existing safety measures, and design strategies to improve railroad crossing safety. Data visualization further enhances the process, turning complex data into actionable insights that inform decision-making and promote safer outcomes at railroad crossings across the United States.

SMART Questions

For our dataset, we have developed the following smart questions, primarily focused on prediction & analysis:

1. **How can we predict the severity of a driver's injury in a railroad crossing accident using external factors?**
2. **How can we identify accident-prone locations in the USA based on accident frequency over the past 46 years?**
3. **How can we predict the presence of crossing warning signs during railroad accidents using historical data?**

Literature Review

Highway-railroad crossing accidents remain a critical safety concern in the U.S., causing significant fatalities, injuries, and economic losses annually. Research highlights various contributing factors to these accidents, such as driver inattention, risk-taking behaviors, and poor visibility during adverse weather conditions (Liu et al., 2012). The disparity in size and weight between trains and vehicles exacerbates the severity of collisions, leading to serious injuries and fatalities.

Improved technology and safety systems have proven effective in reducing accidents. Automated warning systems, enhanced signal systems, and risk reduction measures, as highlighted by Arthur D. Little, Inc. (1996), play a crucial role in mitigating incidents. Additionally, studies on derailment factors emphasize the importance of infrastructure improvements and risk assessments in reducing hazardous material transportation risks (Barkan et al., 2003).

Freight train operations also influence accident rates. Empirical analyses by Zhang et al. (2022) indicate that derailment rates vary between unit trains and manifest trains, emphasizing the need for tailored risk management strategies. Despite advancements, safety remains a challenge, necessitating ongoing improvements to warning systems, visibility enhancements, and public awareness initiatives.

Dataset Overview

The dataset, titled "Highway-Rail Grade Crossing Accident Data," was sourced from Kaggle and contains 46 years of historical data on railroad crossing incidents across the United States. Initially, the dataset included 239,487 observations and 141 variables. However, many variables were highly correlated, either presenting the same features in different ways or conveying the same information with slight variations. As a result, we removed the redundant columns. The final dataset now contains 120,365 observations and 46 variables, covering information on railroad incidents, location details, highway and crossing specifics, vehicle data, train information, and environmental conditions.

Preparing The Data

As part of preprocessing the data we implemented the following steps:

Handling Missing Values: Missing data can compromise the integrity of the dataset, making it less reliable for analysis. In our dataset, we found no missing values.

Managing Outliers: Outliers can affect model accuracy and statistical analysis. Given that each incident in our dataset is crucial for identifying causes, we retained all data points, including outliers, to preserve valuable information.

Removing Duplicate Rows: Duplicate entries can skew analysis results. We identified and removed four duplicate rows from our dataset.

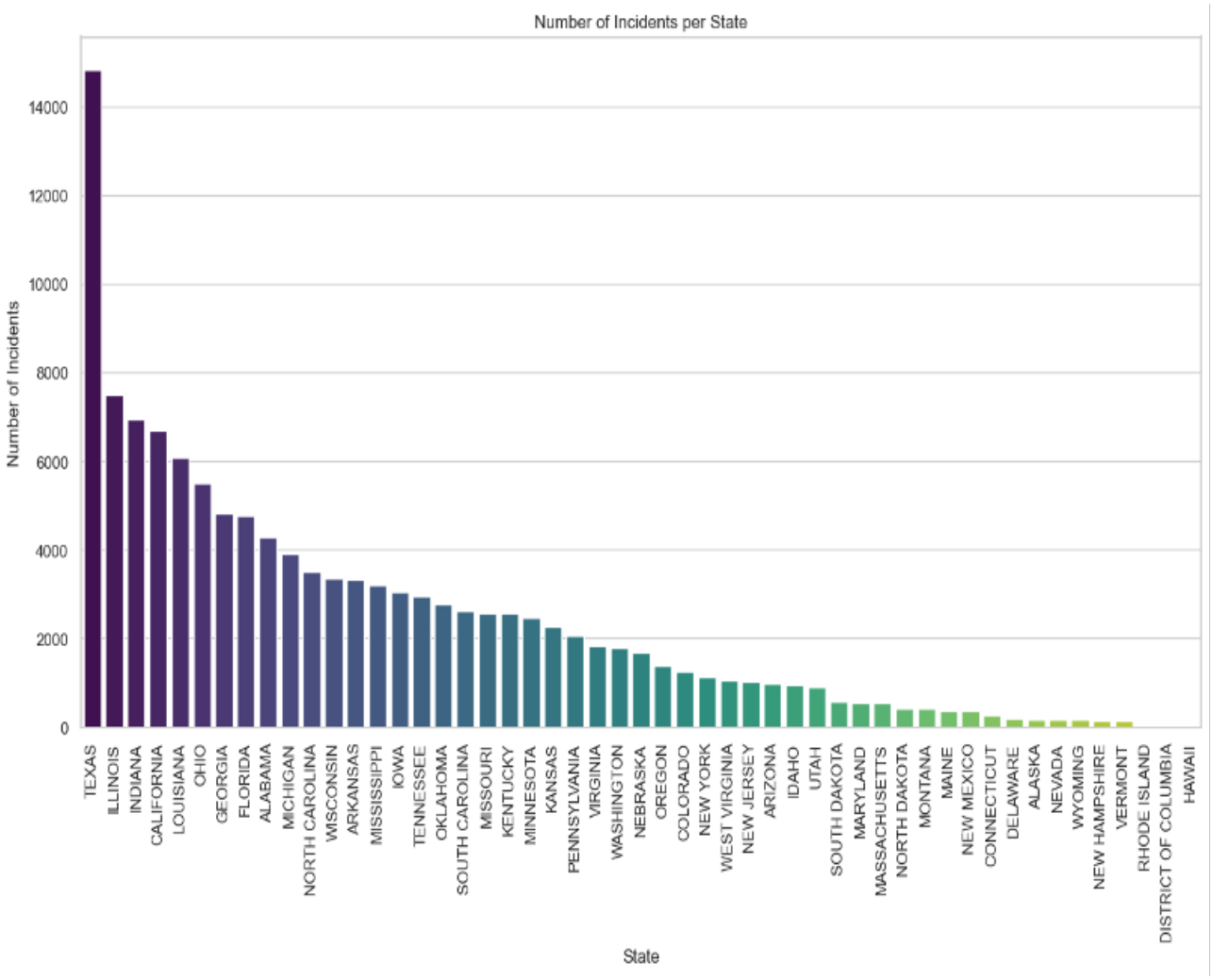
Data Type conversion: We converted the relevant categorical variables into factorial values to facilitate better analysis.

Exploratory Data Analysis

Data visualization serves as a critical tool to transform this dataset into actionable insights. Let’s begin by visualizing the US Highway Railroad Crossing Accident Dataset to uncover meaningful patterns and trends.

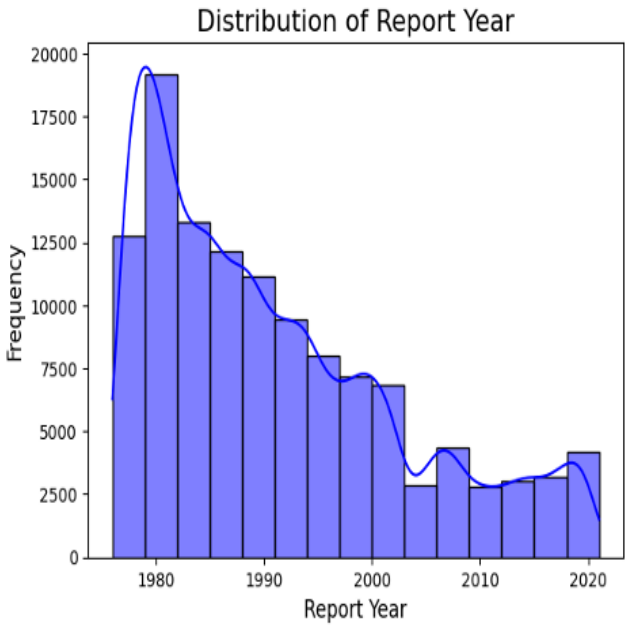
1.State Analysis: Texas has the significantly highest number of incidents, followed by Illinois and Indiana. In contrast, Hawaii and the District of Columbia have almost

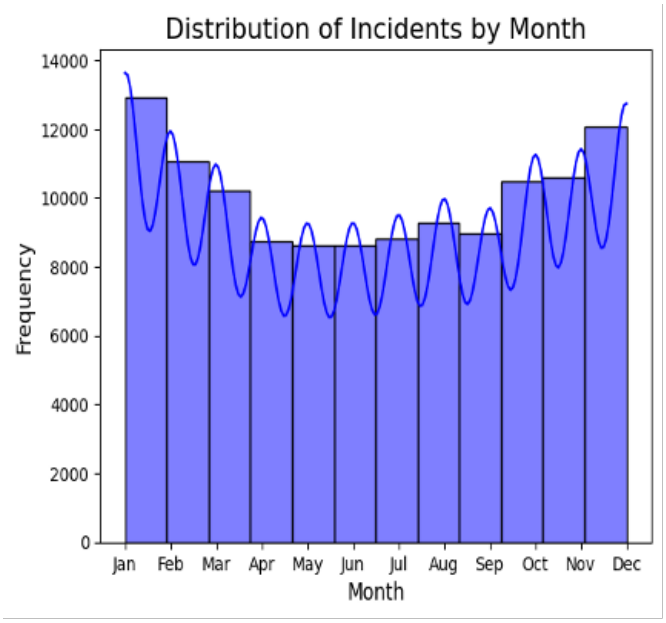
no incidents. Possibly since Hawaii and DC have smaller, more manageable rail systems than Texas with bigger rail systems, which results in fewer incidents.



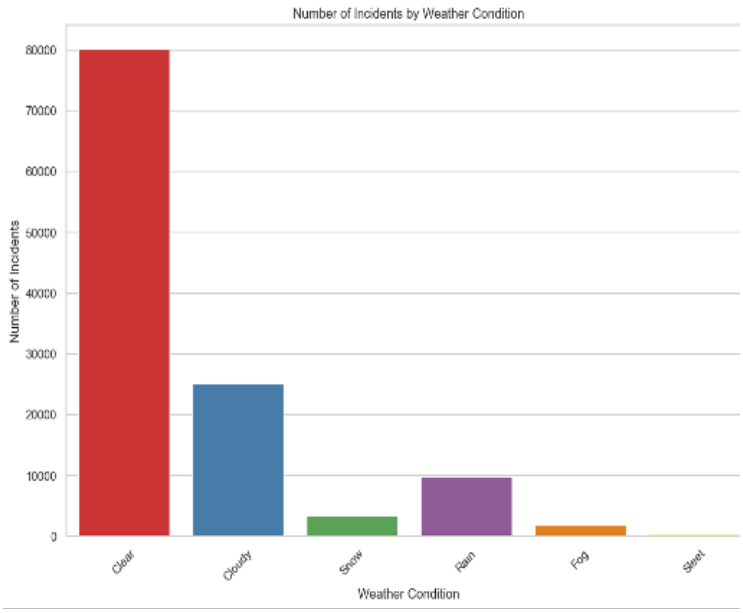
2. Yearly and Monthly Analysis: The number of incidents was significantly higher in 1980, but there has been a notable decline in the following years which may be attributed to advancements in technology, infrastructure, safety measures. Similarly, January and December record the highest number of incidents, while April, May, and June see a slight decrease of about 3,000 incidents which could be

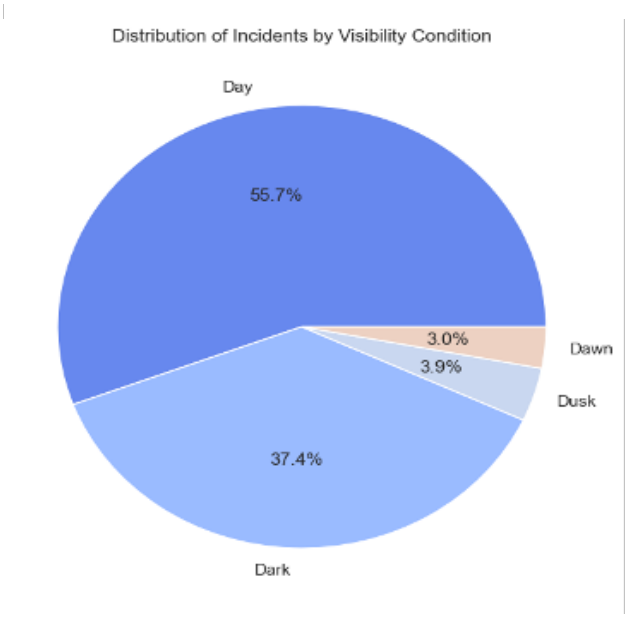
influenced by winter weather conditions and increased travel during the holiday seasons.



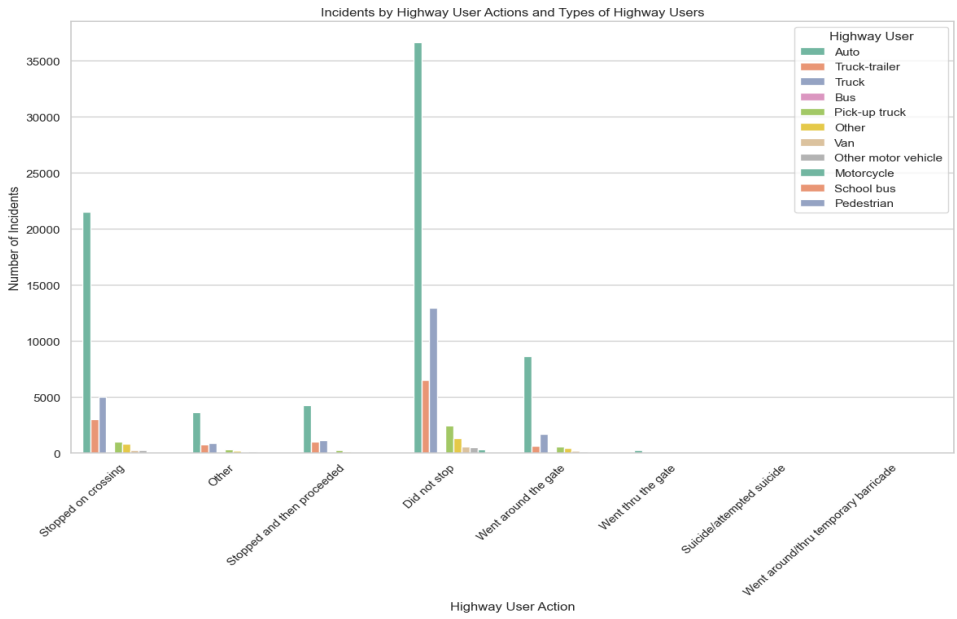


3. Weather and Visibility Analysis: Clear weather is associated with a considerable number of incidents since it could also be the reason for increased train activity. Furthermore, the visibility data reveals that 50% of incidents occur during daylight hours, while 38% take place at night.





4. **User action and type of highway user analysis:** The graph shows that most accidents occurred due to vehicles failing to stop, with autos being the most involved type, having the highest accident count. The failure to stop is often linked to driver inattention, distractions, or misjudgment, especially at railroad crossings. Additionally, drivers of autos may be less aware of crossing signals or may underestimate the danger of approaching trains, leading to a higher incidence of accidents compared to other vehicle types.



5. **Additional Insights:** The data indicated that most accidents took place on the main track, while industry, yard, and siding tracks saw significantly fewer incidents. Furthermore, accidents were more likely to occur when vehicles and trains were moving in opposite directions rather than traveling parallel to one another.

RESEARCH QUESTIONS

Our research questions were carefully formulated to address key challenges and uncover actionable insights using historical railroad crossing accident data.

1. How can we predict the severity of a driver's injury in a railroad crossing accident using external factors?

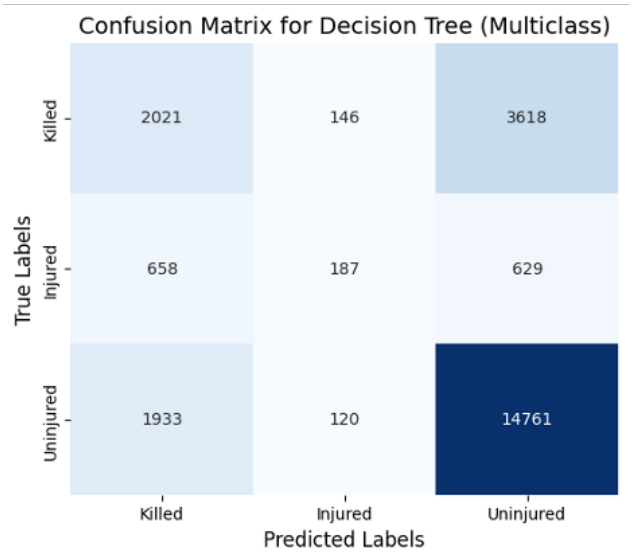
Railroad crossing accidents often occur under complex circumstances involving multiple factors like weather conditions, vehicle speed, train speed, visibility, and driver behavior. By analyzing these factors, we can predict the likelihood and severity of injuries, enabling targeted safety interventions such as speed limits, better signage, and public awareness programs.

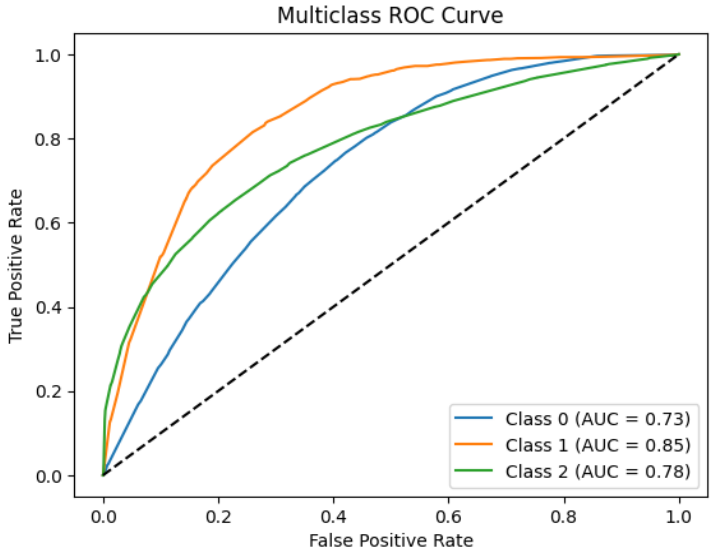
MODELS USED:

To answer this smart question, we have utilized two algorithms: - Decision Tree and Random Forest.

1.Decision Tree Classifier:

The algorithm initially achieved an accuracy of 63%. However, after implementing hyperparameter tuning using GridSearchCV and adjusting parameters like max_depth and min_samples_split, the model's accuracy increased to 70%, along with an overall improvement in performance.

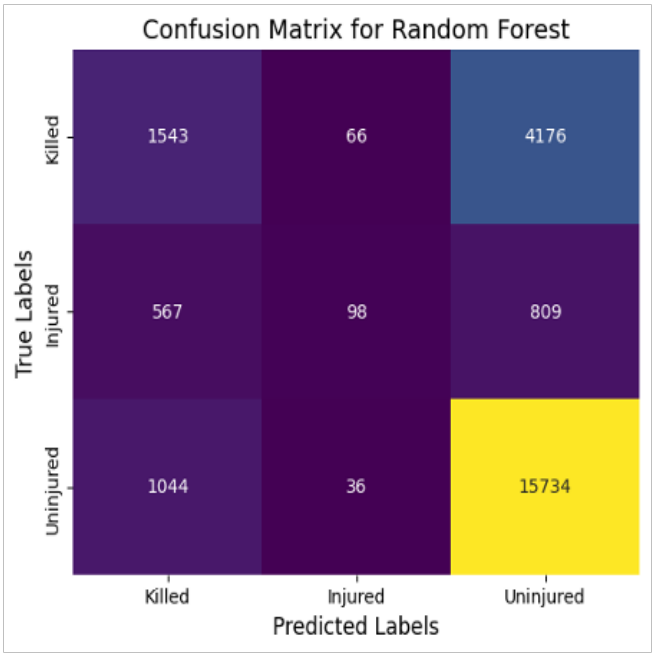


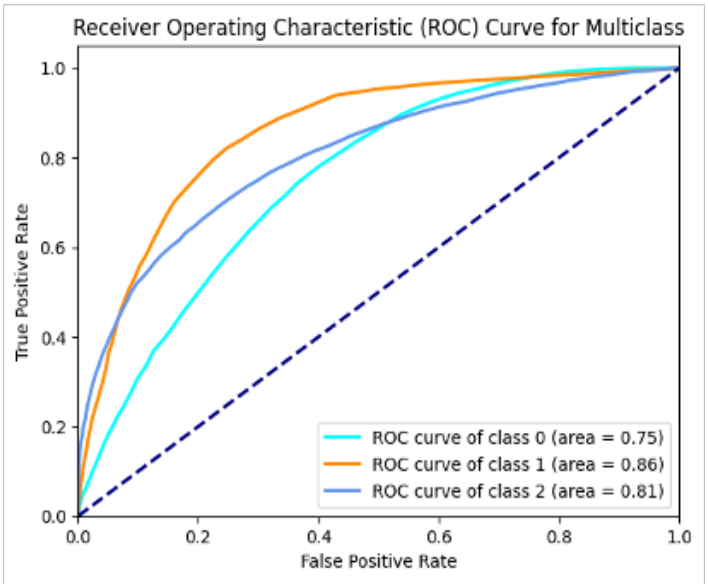


2. Random Forest classifier:

We aimed to enhance the performance of tree-based models by using ensemble methods like Random Forest. In this approach, we trained the model on the same dataset as the decision tree, visualized the feature importances, and then eliminated features with importance weights below

0.01. However, the improvement was marginal, with both models achieving around 72% accuracy—still an improvement over the decision tree model.





Conclusion:

By answering this question, we developed a predictive model to estimate injury severity based on external factors. These predictions can guide authorities in prioritizing high-risk locations and conditions to enforce stricter safety measures, optimize emergency response, and educate drivers on safe practices.

Table 1 SMART Q1 Results

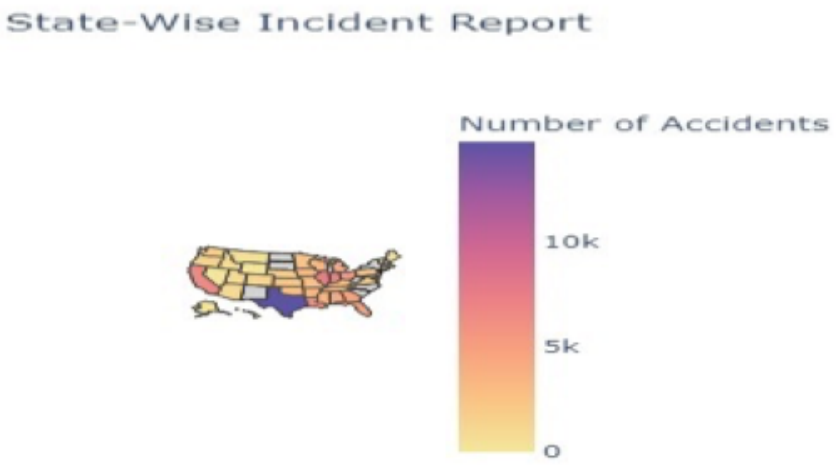
Method Used	**Accuracy**	Precision	Recall	F1 Score
Decision Tree	0.63	0.64	0.63	0.64
Decision tree with GridSearchCv	0.70	0.67	0.70	0.68
Random Forest	0.728	0.68	0.72	0.68
Random Forest (Feature selection)	0.724	0.68	0.73	0.68

2. How can we identify accident-prone locations in the USA based on accident frequency over the past 46 years?

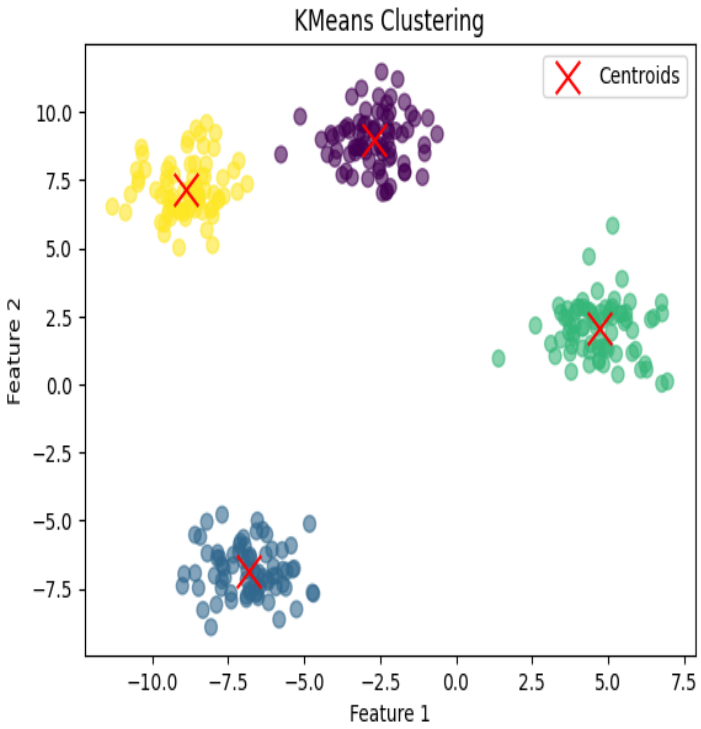
This question focuses on identifying geographical hotspots for railroad crossing accidents across the United States. Understanding where accidents occur most frequently helps authorities concentrate safety efforts in the most vulnerable areas.

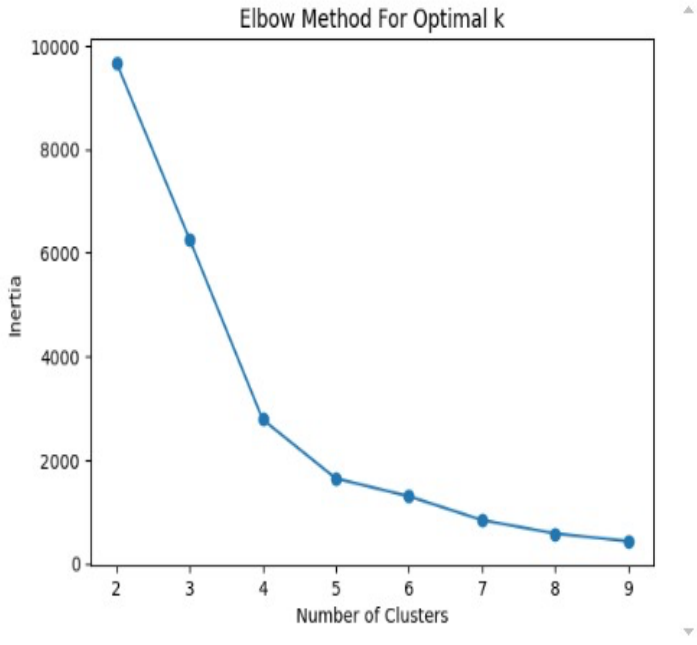
Approach:
We analyzed the dataset to determine trends over 46 years, identifying states and regions with the highest accident frequencies.

State-Level Insights from EDA: Texas emerged as the most accident-prone state, followed by Illinois and Indiana. These states reported consistently high accident frequencies due to their extensive railroad networks and higher crossing volumes.



MODELS USED:
K-Means Clustering was applied to group states and cities into clusters in order to identify which states and cities were more prone to accidents. The Elbow method was utilized to determine the optimal number of clusters. The **silhouette score** is used to validate the clustering performance.





Conclusion:
By pinpointing accident-prone locations, we were able to classify the states and cities based on number of accidents into:

1. Small states and Big cities accidents.
2. Small city and state accidents
3. Large city and state accidents
4. The outliers.

3. How can we predict the presence of crossing warning signs during railroad accidents using historical data?

This question was designed to predict the location of crossing warning signs on railroads at the time of an accident, with the aim of identifying which locations are associated with a higher frequency of accidents. Understanding the correlation between accidents and warning systems can reveal gaps in infrastructure effectiveness and suggest improvements.

Approach:

We discovered that the target variable, 'Crossing Warning Location,' was significantly imbalanced, which impacted the model's performance. To address this, we applied the SMOTE technique to handle class imbalance, ensuring better representation of the minority classes. Additionally, we used label encoding to convert all categorical features into numerical values for modeling purposes.

Model Used:

The Random Forest algorithm was implemented since it can effectively deal with missing or unbalanced data, which is common in historical accident datasets. We noticed that after using the SMOTE methodology the performance of the model improved by a huge margin achieving an accuracy of up to 80%.

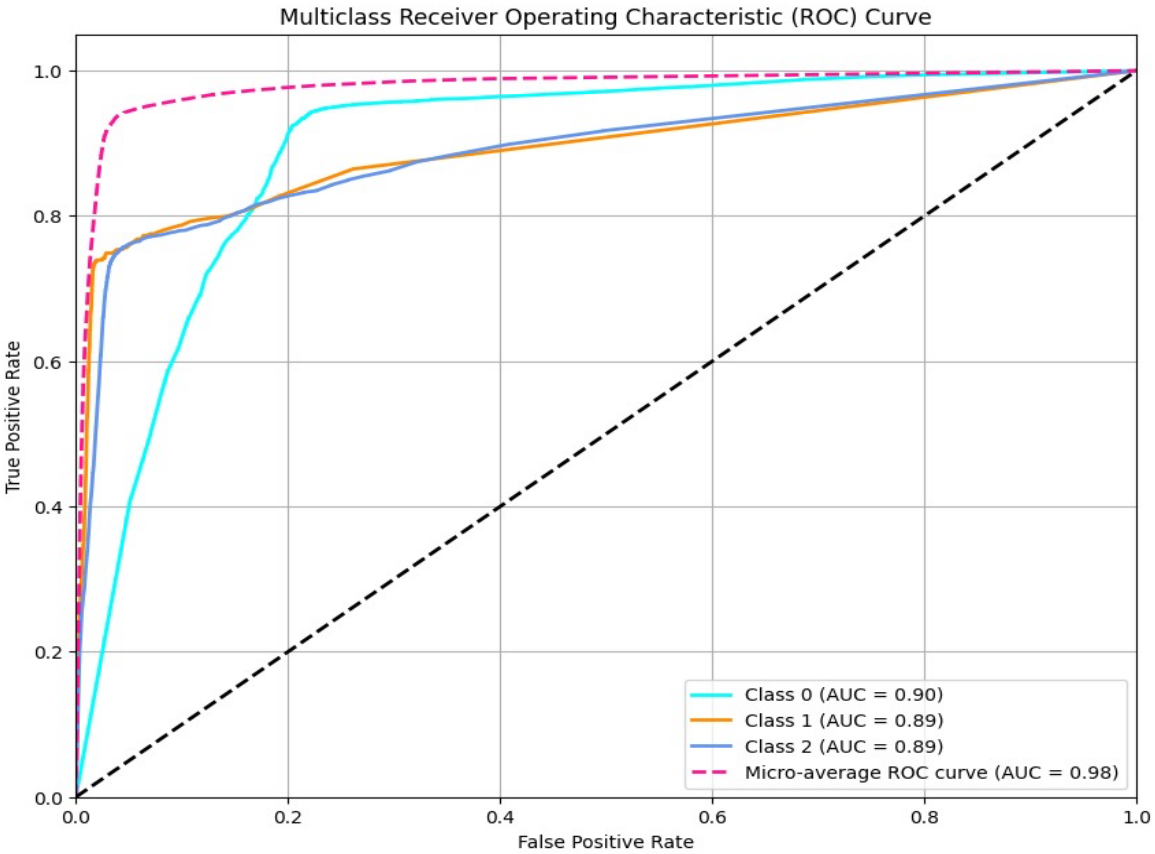


Figure 1 ROC Curve for Random Forest

Conclusion:

Table 2 : SMART Q3 Results

Method Used	Accuracy	Precision	Recall	F1 Score
Random Forest	0.80	0.88	0.80	0.84

In addressing this smart question,

1. We were able to identify locations where accidents are likely due to insufficient or absent warning signs.

2. Suggest high-risk crossings for improvement, such as installing new gates or lights.

Summary of Research Questions

Each of these research questions addresses a specific safety concern: predicting injury severity, identifying accident-prone locations, and understanding the role of crossing warning signs. Together, they provide a comprehensive approach to improving railroad crossing safety through data analysis and machine learning. The insights derived can guide infrastructure improvements, enhance safety regulations, and reduce the frequency and severity of railroad crossing accidents.

References

1. Arthur D. Little, Inc. (ADL). Risk Assessment for the Transportation of Hazardous Materials by Rail, Supplementary Report: Railroad Accident Rate and Risk Reduction Option Effectiveness Analysis and Data, 2nd rev.ADL, Cambridge, Mass., 1996.
2. Barkan, C. P. L., C. T. Dick, and R. Anderson. Railroad Derailment Factors Affecting Hazardous Materials Transportation Risk. *In Transportation Research Record: Journal of the Transportation Research Board*,No. 1825, Transportation Research Board of the National Academies,Washington, D.C., 2003, pp. 64–74.
3. Liu, Xiang & Saat, Rapik & Barkan, Christopher. (2012). Analysis of Causes of Major Train Derailment and Their Effect on Accident Rates. *Transportation Research Record: Journal of the Transportation Research Board*. 2289. 154-163. 10.3141/2289-20.
4. Zhang Z, Lin C-Y, Liu X, et al. An Empirical analysis of freight train derailment rates for unit trains and manifest trains. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit.* 2022;236(10):1168-1178. doi:10.1177/09544097221080615