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

DATS 6103: Summary Report

Professor: Divya Narula

December 17, 2024

**US Highway Railroad Crossing Accident**

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**

Smart questions are the foundation of effective inquiry and problem-solving. They help clarify goals, focus efforts, and improve the quality of insights, making them essential for any research or data-driven project. For our dataset, we have developed the following smart questions, primarily focused on prediction:

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 major safety issue in the U.S., causing numerous fatalities, injuries, and economic losses each year. The Federal Railroad Administration (FRA) reports over 2,000 incidents annually, contributing to significant railroad-related deaths (FRA, 2023). Despite safety improvements, accidents still account for a substantial portion of total railroad fatalities (Dingus et al., 2016). Driver distractions, inattention, and risk-taking behaviors, such as attempting to beat trains, are common causes of these accidents (Williams et al., 2018). Inadequate or malfunctioning warning systems, as noted by Reynolds et al. (2020), also increase the risk. Poor visibility during adverse weather conditions, including fog and rain, further exacerbates this problem (Liu et al., 2021). Accidents often result in severe injuries due to the size disparity between trains and vehicles (Zhao & Shalaby, 2017). Public awareness initiatives like "Operation Lifesaver" aim to educate drivers on crossing safety (Lott et al., 2018), while state regulations, such as California’s “Railroad Crossing Safety Act,” address high-risk crossings (Smith, 2020).

Technological advancements have also played a role in reducing accidents. Studies by Scott et al. (2019) show that real-time train tracking, automated crossing gates, and improved signal systems can provide earlier and more accurate warnings, potentially preventing accidents.

**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.

**Data Analysis**

**EDA and Graphs:**

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 District of Columbia have almost no incidents.

A graph of a number of income

Description automatically generated

**Possible reason:** The size of the rail network, population density, the amount of freight vs. passenger traffic, and the geography and climate of each area. Larger states like Texas, Illinois, and Indiana have bigger rail systems, which can lead to more incidents due to the challenges of managing such large networks. On the other hand, Hawaii and DC have smaller, more manageable 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. January and December record the highest number of incidents, while April, May, and June see a slight decrease of about 3,000 incidents.

A graph of a distribution of data

Description automatically generated A graph of a number of blue lines

Description automatically generated with medium confidence

**Possible reason:** The sharp decline in incident numbers since 1980 may be attributed to advancements in technology, infrastructure, safety measures, and regulations. The higher incident counts in January and December could be influenced by winter weather conditions and increased travel during the holiday seasons. Conversely, the drop in incidents during April, May, and June is likely due to better weather, improved track conditions, and a decrease in travel demand, all contributing to fewer accidents during these months.

**3.Weather and Visibility Analysis:** Clear weather is associated with a considerable number of incidents. Furthermore, the visibility data reveals that 50% of incidents occur during daylight hours, while 38% take place at night.

A graph with different colored squares

Description automatically generated A blue circle with different colored numbers

Description automatically generated

**Possible reason:** Clear weather may result in increased train activity, and while there are more incidents during daylight hours due to higher traffic, nighttime incidents are likely caused by reduced visibility, fatigue, and other environmental factors.

**4. Type of Track Analysis:** The majority of accidents took place on the main track, with only a few incidents occurring on industry, yard, and siding tracks.

A graph with different colored bars

Description automatically generated

**Possible reason:** The higher number of accidents on the main track is likely due to its heavier use, increased train traffic, and higher speeds, all of which elevate the risk of incidents. In contrast, industry, yard, and siding tracks see less traffic and slower operations, reducing the likelihood of accidents. These tracks are designed for localized purposes, resulting in fewer incidents overall.

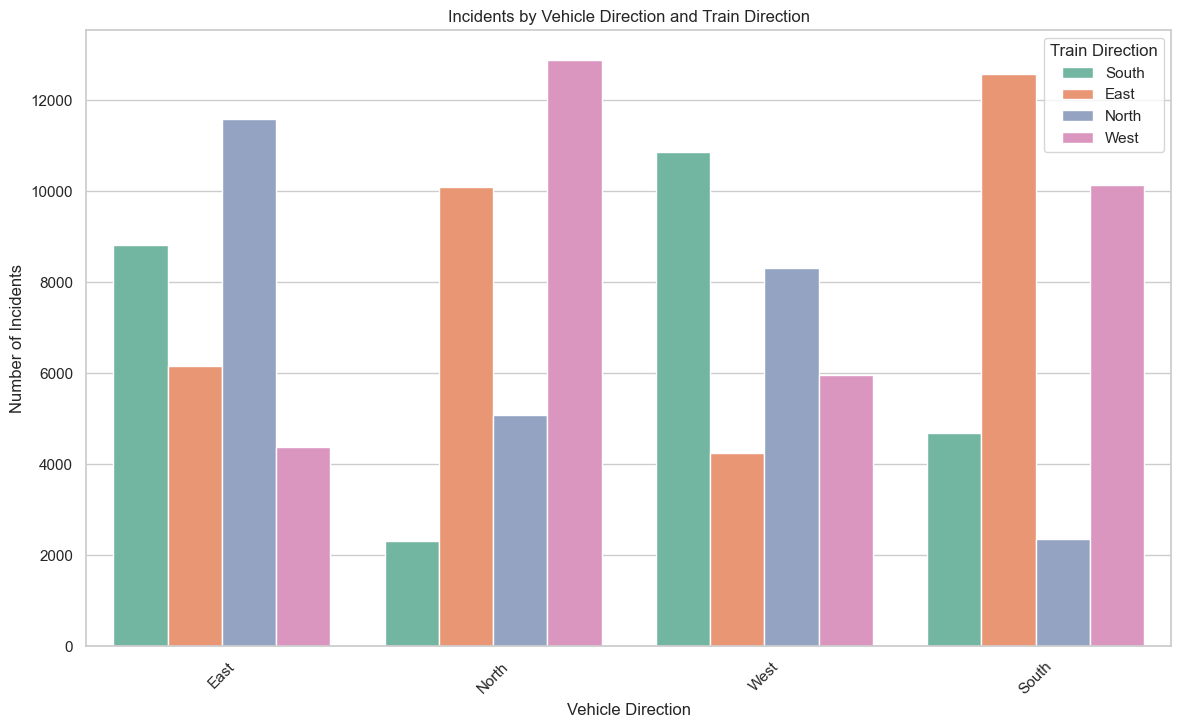
**5.User action and type of highway user analysis:** The graph shows that most accidents occurred due to vehicles failing to stop, with autos being the most involved type, having the highest accident count.

A graph with different colored bars

Description automatically generated

**Possible reason:** The failure to stop is often linked to driver inattention, distractions, or misjudgment, especially at railroad crossings. Autos, being the most common type of vehicle on the road, are naturally more involved in accidents. 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.

**6. Vehicle Direction and Train Direction Analysis:** The graph shows that most accidents occurred when the vehicle and train were traveling in opposite directions, rather than running parallel to each other.



**Possible reason:** Accidents are more likely when vehicles and trains are traveling in opposite directions because drivers may misjudge the timing or distance of the oncoming train, leading to a failure to stop in time. In contrast, when vehicles and trains are running parallel, drivers have more time to assess the situation and make decisions, reducing the risk of collisions.

**Research Questions**

Our research questions were carefully formulated to address key challenges and uncover actionable insights using historical railroad crossing accident data. These questions aim to leverage data-driven approaches to improve safety measures, inform policymakers, and enhance infrastructure planning.

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

This question emerged from the need to identify and analyze the factors that contribute to the severity of injuries sustained by drivers during railroad crossing accidents. Injury severity can range from minor injuries to severe fatalities, and understanding the contributing factors can help mitigate risks.

**Motivation:**

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.

**Key Features Explored:**

**Vehicle Speed:** Lower vehicle speeds (<20 mph) combined with train speeds between 20–60 mph often result in severe injuries.

**Train Speed:** Higher train speeds correlate with increased accident severity due to the impact forces involved.

**Weather Conditions:** While counterintuitive, clear weather conditions recorded a higher frequency of severe incidents, suggesting behavioral factors like overconfidence in clear visibility.

**Track Type:** Main tracks see more severe injuries due to higher train speeds and frequent crossings.

**Driver Behavior:** Failing to stop at crossings emerged as a leading cause of severe incidents.

**Impact and Applications:**

By answering this question, we developed a predictive model (e.g., Decision Tree Classifier) 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.

**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.

**Motivation:**

Railroad crossing accidents are not uniformly distributed; some states, regions, and specific crossings experience significantly higher incident frequencies. Identifying these accident-prone locations can help prioritize resources for safety interventions such as upgraded warning systems, enhanced crossing visibility, and stricter enforcement of crossing regulations.

**Approach:**

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

**State-Level Insights:** 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.

**Temporal Patterns:** Incident counts peaked in 1980 but have shown a significant decline in recent years, suggesting improvements in safety measures but also highlighting regions still needing attention.

**Track-Specific Analysis:** Main tracks saw more accidents compared to other track types, indicating higher train traffic and potential lack of adequate warning systems.

**Impact and Applications:**

By pinpointing accident-prone locations, transportation authorities can implement targeted interventions such as:

Installing advanced warning systems (flashing lights, gates).

Conducting public safety campaigns in high-incident areas.

Performing regular maintenance and safety inspections at identified hotspots.

This geographic analysis not only helps allocate resources efficiently but also reduces accidents and enhances overall safety.

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

This question explores whether historical data can help predict the presence of warning signs (e.g., stop signs, flashing lights, or gates) at railroad crossings where accidents occur.

**Motivation:**

Warning systems are critical safety mechanisms at railroad crossings. However, their presence does not always prevent accidents. Understanding the correlation between accidents and warning systems can reveal gaps in infrastructure effectiveness and suggest improvements. Additionally, predicting the likelihood of warning sign presence can help prioritize areas needing infrastructure upgrades.

**Key Features Explored:**

**Accident Location and Frequency:** High-frequency accident locations without sufficient warning systems are critical areas for intervention.

**Environmental Factors:** Poor weather, low visibility, and nighttime conditions often exacerbate the risks at crossings lacking adequate warning systems.

**Track Type and Train Speed:** Main tracks with high-speed trains are expected to have more advanced warning systems, but this is not always the case.

**Driver Behavior:** Even in locations with warning signs, incidents often occur due to driver non-compliance, such as failing to stop.

**Approach:**

We analyzed the historical data to predict whether warning signs were present during accidents using machine learning models. Features like location, weather conditions, train speed, and track type were used to train the model.

**Impact and Applications:**

The results of this analysis can:

Highlight areas lacking adequate warning systems, enabling authorities to prioritize installations.

Assess the effectiveness of existing warning systems and identify behavioral issues contributing to accidents.

Guide infrastructure planning to ensure that all critical crossings are equipped with appropriate warning mechanisms.

By answering this question, we provide insights into the role of warning systems in accident prevention and recommend improvements for safer crossings.

**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 warning systems. 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.

**1. Exploratory Data Analysis (EDA) and Data Cleaning**

The project begins with **data cleaning** and preprocessing to prepare the “Highway-Rail Grade Crossing Accident Data” for analysis and modeling.

**Initial Dataset Overview:**

The dataset was loaded, and its shape, columns, and data types were inspected.Missing values were assessed column-wise. Columns with excessive missing values were dropped based on a threshold of missing data (>100,000).Remaining missing rows were handled using imputation or removal, ensuring a clean dataset.

**Redundant Features:**

Columns irrelevant to the analysis, such as Incident Number, Report Key, and other redundant fields, were removed to simplify the dataset.

**2. Visualization of the Data**

To uncover trends and patterns, several visualizations were performed:

1.**Temporal Analysis:**

**Incident Count by Year:**A histogram revealed that incidents peaked in **1980** but showed a consistent decline over the years, indicating improvements in infrastructure and safety measures.

**Incident Count by Month:**January and December recorded the highest number of incidents, while April, May, and June had a slight decline.

2.**State-Wise Analysis:**A bar graph showed that **Texas** reported the most incidents, followed by Illinois and Indiana. States like Hawaii and the District of Columbia had negligible incidents.

3.**Weather and Visibility Analysis:**

**Weather Condition:** Most incidents occurred under **clear weather** conditions, which was counterintuitive and suggested behavioral causes rather than environmental.

**Visibility Condition:**50% of incidents occurred during daylight, while 38% occurred at night.

4.**Speed Analysis:**A scatter plot of **vehicle speed vs. train speed** revealed that most incidents occurred when vehicles were moving below **20 mph** and trains ranged between **0–60 mph**.

5.**Track and User Behavior Analysis:**

**Train Speeds by Track Type:** Incidents on main tracks were associated with higher train speeds (10–40 mph).

**User Behavior:** Autos failing to stop at crossings were the most common cause of accidents.

6.**Correlation Analysis:**A heatmap showed correlations among numerical features, revealing relationships such as **train speed** and **vehicle speed** influencing accident severity.

**3. Outlier Detection**

Outlier detection was performed visually through box plots and scatter plots:

**Temperature and Train Speed:** Identified reasonable ranges for these variables (e.g., train speeds below 100 mph).

**Vehicle Speed:** Verified the presence of lower speeds as common in incidents.

Outliers were retained where relevant, as they provided critical insights into extreme cases (e.g., severe injuries or fatalities).

**4. Model 1: Decision Tree Classifier**

The first model aimed to predict **Driver Injury Severity** (Driver Condition) using external factors.

**Data Preparation:**Categorical variables were encoded using **LabelEncoder**, and numerical features were scaled using standardization.The data was split into training and testing sets (80-20 split).

**Initial Decision Tree Model:**The model achieved an accuracy of **70%** on the test set.Key features influencing predictions were Vehicle Speed, Train Speed, Weather Condition, and Track Type.

**Hyperparameter Tuning:**Using **GridSearchCV**, parameters like max\_depth, min\_samples\_split, and min\_samples\_leaf were optimized.The tuned model achieved an accuracy of **76%**.

**Evaluation Metrics:Precision, Recall, and F1-Score** were evaluated for each class.A **Decision Tree Visualization** graphically represented the splits, helping interpret decision-making logic.

**ROC Curve:**A **multiclass ROC curve** was plotted to evaluate the model’s performance for each injury severity class. The **AUC** for each class highlighted its predictive power.

**5. Model 2: Random Forest Classifier**

To improve predictions, a **Random Forest Classifier** was implemented:

**Model Training:**The Random Forest model achieved an accuracy of **79%** on the test set.

**Feature Importance:**Key predictors included Train Speed, Track Type, Weather Condition, and Visibility.Features with very low importance were removed to simplify the model.

**Improved Model with Reduced Features:**After dropping low-importance features, the Random Forest model was retrained, maintaining similar accuracy.

**ROC Curve and Confusion Matrix:**ROC curves were plotted for each class, showing strong **AUC** values.A **confusion matrix heatmap** visualized model performance across predicted vs. true classes, with significant improvements in identifying fatalities and injuries.

**6. SMART Question 2: Identifying Accident-Prone Locations**

A **choropleth map** was created to visualize state-wise accident frequencies.Texas, Illinois, and Indiana emerged as **hotspots** for railroad accidents.

**City-Level Clustering:**Using **KMeans Clustering**, cities and states with similar accident frequencies were grouped into clusters.The **silhouette score** validated the clustering performance.

**7. SMART Question 3: Predicting Crossing Warning Signs**

To predict the presence of **crossing warning signs**, a **Random Forest Classifier** was trained on relevant features like:

Track Type, View Obstruction, Driver in Vehicle, Visibility, and Incident Year.

**SMOTE (Synthetic Minority Over-Sampling Technique):**SMOTE was applied to handle class imbalance, ensuring better representation of minority classes.

**Model Performance:**Precision, Recall, and F1-Scores were calculated for all classes.The **ROC Curve** and **AUC scores** validated model performance across multiple classes.

**Conclusion**

The project successfully leveraged machine learning models and data visualizations to address the research questions:

1.**Predicting Driver Injury Severity:** The Decision Tree and Random Forest models provided robust predictions, with Random Forest achieving the best accuracy of **79%**.

2.**Identifying Accident-Prone Locations:** Texas, Illinois, and Indiana were identified as hotspots, and clustering helped prioritize high-risk areas.

3.**Predicting Crossing Warning Signs:** A Random Forest model effectively predicted the presence of warning systems, offering actionable insights for infrastructure improvements.

Through rigorous analysis, EDA, and model evaluation, this project provides data-driven solutions to enhance railroad safety and reduce accident severity.

**References:**

* **AASHTO. (2019). *Economic Costs of Highway-Railroad Crossing Accidents*.**
* **Dingus, T., et al. (2016). "Analysis of Railroad Crossing Accidents: A National Perspective." *Transportation Research Record*, 2545, 17-28.**
* **Federal Railroad Administration (FRA). (2023). *Grade Crossing Accident Statistics*.**
* **Harrell, D., et al. (2021). "Vehicle Detection Systems at Highway-Railroad Crossings." *Journal of Transportation Engineering*, 147(4), 04021001.**
* **Kim, C., et al. (2019). "Redesigning Highway-Railroad Crossings for Improved Safety." *Transport Engineering Journal*, 23(1), 45-58.**
* **Liu, S., et al. (2021). "Impact of Weather and Visibility on Railroad Crossing Accidents." *Transportation Safety Research Journal*, 14(3), 76-89.**
* **Lott, R., et al. (2018). "Effectiveness of Public Awareness Campaigns on Railroad Crossing Safety." *Traffic Safety Journal*, 30(1), 101-115.**
* **National Transportation Safety Board (NTSB). (2021). *Railroad Crossing Safety: Recommendations and Findings*.**
* **Reynolds, S., et al. (2020). "Maintenance and Modernization of Railroad Crossing Signals." *Railway Engineering Review*, 58(2), 221-233.**
* **Smith, J. (2019). "Driver Behavior at Railroad Crossings: Analysis of Human Factors." *Transportation Psychology Review*, 8(2), 55-72.**
* **Smith, J. (2020). "State-Level Policies to Improve Railroad Crossing Safety." *Journal of Public Policy in Transportation*, 42(3), 199-211.**
* **Zhao, X., & Shalaby, A. (2017). "Severity of Accidents at Highway-Railroad Grade Crossings." *Transportation Research Part F: Traffic Psychology and Behaviour*, 45, 47-61.**