Railroad incidents analysis

Table of Contents

[**List of Figures** 2](#_Toc165493211)

[**Acknowledgements** 3](#_Toc165493212)

[**Abstract** 4](#_Toc165493213)

[**1 Introduction** 4](#_Toc165493214)

[**2 Literature Review** 5](#_Toc165493215)

[**3 Data** 6](#_Toc165493216)

[3.1 Description of the dataset 6](#_Toc165493217)

[3.2 Descriptive Statistics of interest 8](#_Toc165493218)

[3.3 Exploratory Data Analysis 13](#_Toc165493219)

[3.4 Outliers and missing values 19](#_Toc165493220)

[**4 Methodology** 20](#_Toc165493221)

[**5 Results** 23](#_Toc165493222)

[5.1 Model Evaluation after Feature Selection 26](#_Toc165493223)

[5.2 Stepwise Regression Technique 30](#_Toc165493224)

[**6 Conclusions** 31](#_Toc165493225)

[6.1 Logistics Regression Model 31](#_Toc165493226)

[6.2 Neural Network Model 31](#_Toc165493227)

[6.3 New Questions and Future Exploration: 33](#_Toc165493228)

[**10 References** 34](#_Toc165493229)

[**11 Appendix** 37](#_Toc165493230)

[11.1 Data File Structure and Field Input Specifications as provided by the Department of Transportation [38] 37](#_Toc165493231)

# **List of Figures**

[Figure 1- Major fields used in study. Complete list of field names found in List of Tables - Data File Structure and Field Input Specifications. 8](#_Toc165491691)

[Figure 2 - Number of incidents based on different weather conditions. 9](#_Toc165491692)

[Figure 3 - Histogram showing the number of incidents with 1 to 25 cars. 10](#_Toc165491693)

[Figure 4 - Histogram showing the number of incidents with 26 to 75 cars. 11](#_Toc165491694)

[Figure 5 - Histogram showing the number of incidents with 76-125 cars. 11](#_Toc165491695)

[Figure 6 - Descriptive statistics of Car 12](#_Toc165491696)

[Figure 7 - Bar chart showing the number of incidents by railroad company. 12](#_Toc165491697)

[Figure 8 - Value count for categorical variables. 13](#_Toc165491698)

[Figure 9 - Distribution of incidents by day of month. 14](#_Toc165491699)

[Figure 10 - Distribution of incidents by weather conditions. 14](#_Toc165491700)

[Figure 11 - Standardized distribution of incidents by train speed 15](#_Toc165491701)

[Figure 12 - Distribution of incidents by type. 15](#_Toc165491702)

[Figure 13 - Standardized distribution of incidents by type. 16](#_Toc165491703)

[Figure 14 - Standardized distribution of incidents by temperature before outliers are removed. 16](#_Toc165491704)

[Figure 15 - Standardized distribution of incidents by temperature after outliers are removed. 17](#_Toc165491705)

[Figure 16 - Distribution of incidents by visibility range. 17](#_Toc165491706)

[Figure 17 - Distribution of incidents for Weather Conditions vs. Incident Type. 18](#_Toc165491707)

[Figure 18 - Distribution of Incident Type by Train Speed Buckets. 19](#_Toc165491708)

[Figure 19 - Subset of numerical variables in shown via boxplot. 20](#_Toc165491709)

[Figure 20 - Box Plot of Weather 20](#_Toc165491710)

[Figure 21 - A Simple Neural Network 23](#_Toc165491711)

[Figure 22 - Confusion matrix for logistic regression model before feature selection. 24](#_Toc165491712)

[Figure 23 - Classification Report for logistic regression before feature selection. 25](#_Toc165491713)

[Figure 24 - Confusion matrix for logistic regression after feature selection. 26](#_Toc165491714)

[Figure 25 - Classification report for logistic regression after feature selection. 27](#_Toc165491715)

[Figure 26 - Confusion matrix for neural network 29](#_Toc165491716)

[Figure 27 - Classification report of the testing set for neural network model 30](#_Toc165491717)

[Figure 28 - Comparison of model accuracies. 31](#_Toc165491718)

# **Acknowledgements**

We would like to thank the University of North Texas for facilitating the course in which this paper was produced. For the research itself, we would like to thank the Federal Railroad Administration for providing such a robust dataset to be used for this study among others. Finally, we would like to thank our professor, Dr. Javier Rubio-Herrero, for providing us with direction on how to improve the study.

# **Abstract**

Railroads move most of the freight within the United States along with a fair number of passengers. However, while they are the most efficient method of transportation, railroads can also be extremely hazardous. Since 1975, the Federal Railroad Administration has received reports of over 318,000 railroad incidents of varying severity. In this study we utilize that dataset to find correlations between attributes of the train, humans operating it, the safety systems they are equipped with, and which type of incident is most likely to occur. Utilizing feature selection, we chose the most relevant variables among the nearly 250 available in the dataset. Logistic regression and neural networks were used to produce models that can predict the type of incident using the attributes stated previously. This logistic regression and neural network models provided classification accuracies on validation data of 83.2% and 84.3% respectively.

**The Search for Safety within Railroad Derailments**

# **Introduction**

A broken rail car is hurtling off the tracks as fiery smoke engulfs the horizon. Unfortunately, these terrifying pictures are bleak reminders of the constant danger associated with railway transit. As important as safety progress has been, improvement is still a major concern because any incident can be life-altering. Accidents are often minor, with damage being limited to track or equipment only. But incidents can also become far more severe in an instant. Freight railroads play a crucial role in the national economy, handling more than 40% of intercity freight ton-miles [1]. However, train accidents have the potential to harm infrastructure and rolling stock, disrupt operations, and pose risks to both individuals and the environment.

Tracking of incident data began in 1975 by the Federal Railroad Administration. Their database has shown an average of 3,667 incidents have led to 947 injuries and $333.7 million in damages annually [2]. These values may seem high to someone unfamiliar with the rail industry, but the peak values reached as high as 10,531 incidents and 3,149 injuries in 1978 [3]. However, in 2023 a relatively low 1,795 incidents and 266 injuries occurred despite a $336.2 million damage cost [4]. The top accident damage value came in 2013 to the tune of $494.6 million with overall costs since 1975 totalling $16.3 billion [3]. For example, a major derailment in East Palestine, Ohio in February of 2023 caused a major environmental disaster and has mounting damages over $800 million [5]. This intricacy results from the careful interaction of several railway companies' safety protocols with environmental elements such as hazardous weather and varying track conditions. Although specific facets of railway safety have been studied in the past, previous studies have highlighted the need for a risk system to be developed [6].

Large budgets are allocated annually to developing and maintaining the operations of railroad transportation systems worldwide. Large portions of these budgets often go toward operations and safety improvement. However, numerous people have continued to lose their lives due to different types of railroad accidents. Human error significantly contributes to the incidents that occurred. In fact, a British study showed that 70% of railway accidents occurred on the four main lines between 1990 and 1997 that were results of human error [7]. In 2004, 53% of railway switching yard accidents in the United States, excluding railroad crossing incidents, involves human factor causes [8]. Safety has been improved significantly among railroads over the last two decades, but the incident rate in railroad yards remains above industry average [9]. While this leaves much more room for improvement, the current momentum has come about thanks to the railroad industry’s adoption of safety culture.

Safety culture on the railroad means the exhaustive attempt to keep railroad employees and the greater public safe while moving goods throughout the country. Especially among railroads, unsafe behavior is strongly discouraged and can be met with harsh penalties. Safety culture refers to how safety is placed as a priority within an organization. This outcome results in policies and actions. There is no single definition for the term safety culture, and this is considered as a drawback and leads to uncertain research. However, we can say that the safety culture of an organization is the product of group and individual behavior and thought process [10]. For example, Conductors working extra shifts can lead to operational mistakes due to human error or lack of awareness causing incidents that can result in loss of life for the worker or others around them. Another main factor of railroad safety is adverse weather conditions such as heavy rainfall or extreme temperatures. These conditions can exacerbate the situation and increase the risk of accidents [11]. Railroad crashes can be brought about by various elements, including inadequate tracks, human error, mechanical failures, and harsh weather conditions. A comprehensive examination of the factors that lead to crashes found that mechanical failures account for approximately 40% of all accidents, while human error and track defects each account for 20% of accident rates on average [12].

The goal of our research is to assist in improving railroad safety. Incidents involving rail equipment must be reported to the Federal Railroad Administration through FRA F 6180.54 or Form 54. We take advantage of the abundant, thorough data found in the Railroad Equipment Accident/Incident Source Data (Form FRA F 6180.54) collection, exploring a wealth of knowledge about events, the surrounding environment, and business activities. Unlike other studies that focus on the Safety Risk Model (SRM) [13], we plan to pursue the following research paths:

1. It is obvious that cold temperatures and precipitation would have a negative impact on any form of transportation, and yet despite causing almost 20% of railway infrastructure damage little research has been done to calculate the risk factor environmental factors apply to rail operations [14]. This thorough analysis will highlight weak points in corporate safety procedures and highlight the distinct environmental risk environments that every business must contend with.
2. Uncover the inner workings of certain businesses by evaluating previous incidents and closely examining the ways in which crew training, track maintenance, and incident response affect each company's overall risk profile. Through the identification of optimal methodologies and prospective avenues for enhancement within enterprises, our aim is to facilitate the industry-wide implementation of safer operating approaches.
3. Create a route for anticipatory prevention: By applying dimensionality reduction methods such as Principal Component Analysis (PCA), we want to create a new framework for risk assessment that can produce accuracy comparable to that of Bayesian interference, decision tree, and petri-net models [15]. This framework will serve as an effective instrument to predict the type of incident along with the severity of said incident, should it occur.

Our research methodology employs Feature Selection along with Principal Component Analysis (PCA), a neural network, and stepwise logistic regressionto analyze the interplay of various factors, such as weather conditions, train attributes, track characteristics, safety protocols, and human factors in railroad incidents. By decomposing the dataset, we aim to identify patterns and relationships that underlie the occurrence and severity of incidents in the railroad industry. Railroad safety is a complex issue influenced by numerous factors. We hope our study will shed light on the key drivers of incidents and provide actionable recommendations for improving safety practices by leveraging advanced analytical techniques and comprehensive data analysis within the US rail environment. With this knowledge, we hope to contribute to building a more robust and secure train system for the future.

In essence, this project represents a pioneering endeavor to leverage the capabilities of machine learning techniques in analyzing railroad safety and mitigating the risks associated with derailments. Our primary goal is to identify actionable insights that can serve as a foundation for targeted interventions and advancements in safety protocols and infrastructure management strategies.

This paper aims to give a background about railroad incidents and their impact in the first section of the introduction followed by an elaboration about previous research done on them in the second section with the literature review. The research methodology is discussed in the third section of the paper and then the analysis and results of the proposed model is presented in the fourth section and finally, the fifth section concludes the paper.

# **2 Literature Review**

The research focuses on the causes of freight-train derailments in the US which include derailments due to weather, low visibility, and extreme temperatures [16]. Need for advancements in infrastructure and technology in the enhancement of safety measures throughout the railroad operations is very crucial [17]. To understand the risks associated with railroad operations throughout the years, countless studies have provided valuable insights into how much of a key factor track maintenance is in avoiding incidents on the railroad.

Railway tracks are the most critical part of the railway system, accidents/incidents related to the track maintenance constituted to 30-40% of total incidents in the United States, over the last decade. They need to be maintained regularly to avoid the risk of accidents and need to be monitored continuously to check if there are any degradations. In the 21st century, data driven models such as deep learning methods are used in the predictive maintenance of railway track. For example, a 2014 study used a multitude of analytical models on a vast amount of track data to predict track failures and save between 200k USD and 5MM USD annually [18]. A similar study done in 2020 used newer technologies like vibration monitors and electric sensors to create a solution that could possibly even be expanded to other areas within the railroad [19]. An earlier study in 2011 used an iterative heuristic approach to eliminate constraint violations and reduce travel costs [20]. In 2021 there was a hierarchical classification, or taxonomy, built for railway track maintenance planning and scheduling of rail in the European Union. A second 2021 study utilized Building Information Modelling (BIM) and the digital twin method to enhance the operations of the Taipei Metro [21]. The most efficient strategies can significantly reduce track failures and maintenance costs and increase life expectancy of track components while also allowing operators to plan maintenance ahead of time [22].

Extreme weather conditions can have impact of railway infrastructure. High temperatures can cause buckling of rails and structure expansion. Similarly, extreme low temperatures along with snow and ice can cause damage to overhead lines, cracking, or breaking of rails [23]. Because of these extremes, the need for a risk assessment and an understanding of different weather impact relationships is clear.

Risk assessment techniques have been developed by employing machine learning methods. These techniques consume operational data and provide insight into the degree of risk or ties factors to increased risk. This is important because the fields in which they are used often come with dire consequences for even the simplest of accidents. Examples of these historically include a variety of ML methods, but recently the use of artificial intelligence has begun to play a more prominent role. For example, a 2020 study exploring deep learning used a convolutional neural network in the pursuit of passenger falls [24]. Another more recent study completed in 2023 utilized a multitude of models to predict the need for evacuation following a hazardous materials incident [25]. Further risk assessment techniques have been developed using Bayesian fusion methods [26] using the data from grade crossings provide valuable insights into the strategies that can employed for enhancing railroad safety measures. Furthermore, Fuzzy Bayesian Networks (FBN) [27] have been used to analyse the risk associated with railway derailments due to the impact of various types of weather and environmental factors. Statistical methodologies like robust multivariate analysis [28] and multivariate adaptive regression splines [29] that are used for a large number of inputs provide a basis for the methodology that can be applied in analysing the data acquired.

Alongside these techniques, the Federal Railroad Administration [30] has formulated a set of guidelines and principles that provide a better understanding of the safety procedures employed and regulated in the railroad sector. For example, the Federal Railroad Safety Act (FRSA) acts as a base for railroad safety regulations in US. It helps the FRA develop and maintain safety standards for railroad operations [31]. The Code of Federal Regulations (CFR) covers other aspects of the railroad such as train track standards, type of materials transportation, and operating procedures [32]. FRA Bulletins provide information on all rule changes and safety issues [33]. FRA Circulars document best practices and provide recommendations for safety regulations [33].

Also, the Federal Hours of Service Act of 1907 regulated hours worked and minimum rest periods but failed to do so on a weekly or monthly basis. This led to a study in 2000 by the Office of Research and Development at the Federal Railroad Administration to improve fatigue levels among railroad workers [34]. To draw insights from our data we are integrating the core machine learning techniques with a variety of different approaches for railway safety analysis. We are considering Form FRA F 6180.54 which helps us in preprocessing data by having a standardized input methodology that allows for increased data integrity. Exploratory data analysis followed by ML models will also be utilized.

Because of our extensive dataset, we plan to utilize Principal Component Analysis (PCA) to analyze data and as a form of dimension reduction. This is completed through transformation of a larger dataset into a smaller one while losing a minimal amount of the original information. Other studies have used PCA. One study used a Principal Component Analysis introduction to a Support Vector Machine method to achieve 75.5% accuracy of track geometry defects, lowering the number of incidents due to this type of track failure [35].

We will utilize multiple methodologies in our pursuit of an accurate way to quantify the risk of railway incidents, utilizing multiple statistical methods like decision trees and logistic regression for variable factors in terms of environment and derailment conditions. Using these, we can predict the key factors causing derailments. This multiple-model analysis gives us accuracy, precision, and recall which helps us to decide the best fit. Not only are performance and reliability considered, but legislative and operations-based decisions could also be made based on the robustness of the results [36].

# **3 Data**

## 3.1 Description of the dataset

Our dataset is comprised of a collection of data gathered by the United States Department of Transportation since 1975. It amounts to approximately 218,000 rows and 145 columns and is constantly gathered by the entity through Form FRA F 6180.54 that is mandatory based on the annual reporting threshold [37]. When an incident occurs for any railroad within the United States and damages raise above the threshold for that year, it is required to be reported directly to the Federal Railroad Administration via this form.

This form provides information that is highly structured and rich in detail. Not only does it provide basic information like the date and time that the event happened, but it also provides information key to investigations of the incident and for studies of events similar to this exercise. While there are many columns that are sparsely populated, the dataset lends itself to a diverse set of study possibilities. While ours focuses on the relationship of factors like weather and visibility having an impact on incident risk, there are also possibilities for text mining due to description fields that contain comprehensive details on how the incident came about. Further knowledge of the form would also allow for further analysis of the causes of the incidents by utilizing the underlying cause codes.

Our project utilizes a dataset that consists of a total of 145 columns, with definitions of variables that were used provided in the table below [38]. Because this table consists of many variables, some columns may be excluded. The dataset is comprised of year, month, railroad code, incident number, car damage, speed, equipment and track damage, etc, which will prove useful in our research. To effectively analyze the dataset and predict incidents, various approaches can be employed, including examining the time of incident, incident frequency, identifying contributing factors, assessing severity, and conducting risk assessments. Each approach necessitates different features and target variables aligned with specific goals. Key variables for analysis could include incident type for categorization, geographic indicators such as state and milepost, weather conditions, rail properties like tons and typical speed, and incident causes. By integrating these variables into analytical models, it becomes possible to discern patterns, mitigate risks, and enhance incident prevention strategies within the railway system.

The Dataset has many columns, where they provide detailed information about accidents. There are many variables like IYR2, IMO2, RR2, INCDTNO2, DUMMY1, GXID, TRKDNSTY, RRCAR2, CARNBR2, POSITON2, LOADED2, CAUSE2, DUMMY2, DUMMY3, ENGHR, ENGMIN, CDTRHR, CDTRMIN, DUMMY4, DUMMY,DUMMY6, ALCOHOL, DRUG, DUMMY7, SSB1, SSB2, NARR1, NARR2, NARR3, NARR4, NARR5, NARR6, NARR7, NARR8, NARR9, NARR10, NARR11, NARR12, NARR13, NARR14, NARR15, RCL, Latitude, Longitude, SIGNAL, MOPERA, ADJUNCT1, ADJUNCT2, ADJUNCT3, SUBDIV, which are having missing values, which doesn’t add any value for our analysis.

The dataset provides rich information about the incidents. A vast amount of data has been collected since 1975, which needs to be cleaned. In future analysis, based on the selected parameters, this will help provide a proper understanding of this dataset can identify many patterns and trends contributing to railroad accidents. All fields from the dataset are available in the List of Fields section. Below we have listed the fields used within our study.

|  |  |
| --- | --- |
| **FIELD NAME** | **DEFINITION** |
| amtrak | Amtrak involvement |
| iyr | year of incident |
| imo | month of incident |
| railroad | railroad code (Reporting RR) |
| incdtno | railroad assigned number |
| iyr2 | year of incident |
| imo2 | month of incident |
| rr2 | railroad code (Other RR involved) |
| incdtno2 | other railroad assigned number |
| iyr3 | year of incident |
| imo3 | month of incident |
| rr3 | railroad code (RR responsible for track maintenance) |
| incdtno3 | railroad assigned number |
| dummy1 | blank data expansion field |
| gxid | grade crossing id number |
| year | year of accident / incident |
| month | month of incident |
| day | day of incident |
| timehr | hour of incident |
| timemin | minute of incident |
| ampm | am or pm |

Figure 1- Major fields used in study. Complete list of field names found in List of Tables - Data File Structure and Field Input Specifications.

## 3.2 Descriptive Statistics of interest

The basic descriptive statistics we observed were that the total number of incidents gathered since 1975 surpassed 218,000. Fifteen percent of these incidents, 32,969 in total, resulted in temperatures below 32 degrees Fahrenheit or 0 degrees Celsius. Heat had a very minor impact on incidents, with only 1,303 incidents or 0.6% of total incidents, occurring at temperatures at or above 100 degrees Fahrenheit. Considering incidents involving passenger trains were only 3 percent of all incidents, it is worth mentioning that the other 97 percent occurred with freight trains that could be carrying chemicals, explosives, or radioactive waste. It was also surprising to find that, as shown in the chart below, 64 percent of all incidents have taken place when the weather is clear, as compared to foggy, rainy, or other safety-impending weather conditions. Finally, it brings hope that in 2023 the fewest number of incidents and injuries occurred in industry history with 1,795 incidents and 266 injuries.

A pie chart with different colored numbers

Description automatically generated with medium confidence

Figure 2 - Number of incidents based on different weather conditions.

The dataset contains both numerical and categorical variables with a split of 51.7% and 47.6% respectively.

Descriptive statistics of numerical variables

The below histogram shows the number of cars involved in the incidents. They range from 0-175 approximately and the x-axis is broken down into intervals of 25 cars each. According to the histogram, there is a high concentration of incidents involving one or few cars and the incidents reduce as the number of cars increases. The dataset for the variable "CARS" comprises a total of 218,470 observations. On average, there are approximately 1.31 cars involved in incidents, with a considerable standard deviation of around 6.24, indicating significant variability around this mean value. The minimum number of cars involved in an incident is 0, suggesting that a substantial portion of incidents do not involve any cars at all. Similarly, both the 25th and 75th percentiles indicate that a notable proportion of incidents do not involve any cars, with 0 representing the value below which 25% and 75% of observations fall, respectively. The median, at 0, signifies that half of the incidents in the dataset involve no cars. However, there is also an extreme outlier, with the maximum observation showing an incident involving a staggering 182 cars. This wide range of values underscores the diverse nature of incidents captured in the dataset.

The dataset for the variable "TRAIN\_WEIGHT" is comprised of a total of 2,101 observations. On average, the weight of trains involved in incidents is approximately 3,383 tons, with a substantial standard deviation of around 5,061.25, indicating significant variability in train weights. The minimum weight recorded in the dataset is 0 tons, suggesting that a notable portion of incidents do not involve any trains. The 25th percentile, at 0 tons, underscores this observation. The median weight, at 540 tons, highlights that many incidents involve relatively light trains. Additionally, the 75th percentile indicates that 75% of incidents involve trains weighing 5,215 tons or less. However, there are also extreme outliers, with the maximum train weight recorded at a staggering 42,900 tons, depicting incidents involving exceptionally heavy trains. This wide range of weights showcases the diverse nature of train incidents captured in the dataset.

The below plots show the distribution of the number of cars involved in incidents. The X-axis represents the number of cars involved and Y-axis represents the number of incidents. Here, the histogram is divided into bins, where each bin represents the number of car counts. The height of the bar represents the number of incidents that fall within that bin range. We can observe that the distribution is skewed towards a smaller number of cars, which means that there are more accidents with a smaller number of cars involved. We included 3 visualizations to show greater detail within the bins.

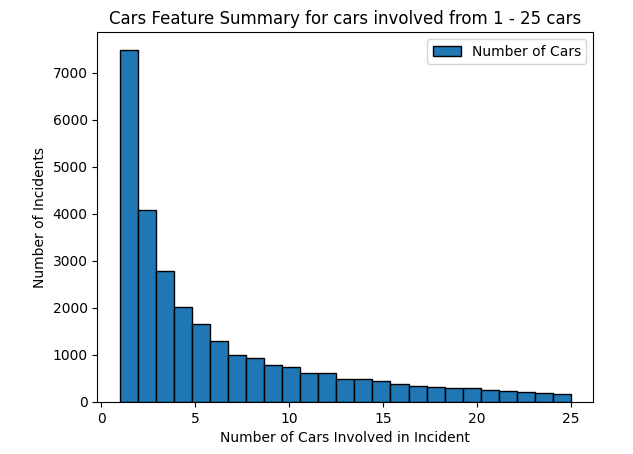


Figure 3 - Histogram showing the number of incidents with 1 to 25 cars.

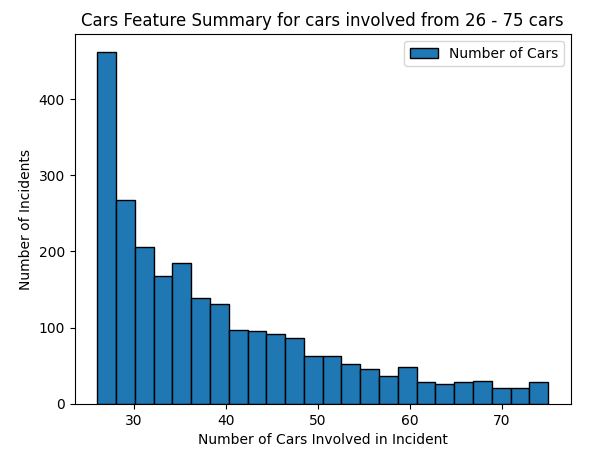


Figure 4 - Histogram showing the number of incidents with 26 to 75 cars.

A graph of blue bars

Description automatically generated

Figure 5 - Histogram showing the number of incidents with 76-125 cars.

|  |  |
| --- | --- |
| **Statistic** | **Value** |
| count | 218,470 |
| mean | 1.316684 |
| std | 6.247457 |
| min | 0 |
| 25% | 0 |
| 50% | 0 |
| 75% | 0 |
| max | 182 |

Figure 6 - Descriptive statistics of Car

Below we can see the distribution of incidents across different railroad companies, highlighting the top 10 companies involved in the most incidents. All other companies are included in Others. This plot gives a clear picture of which companies need to work on improving their processes to reduce or avoid incidents.

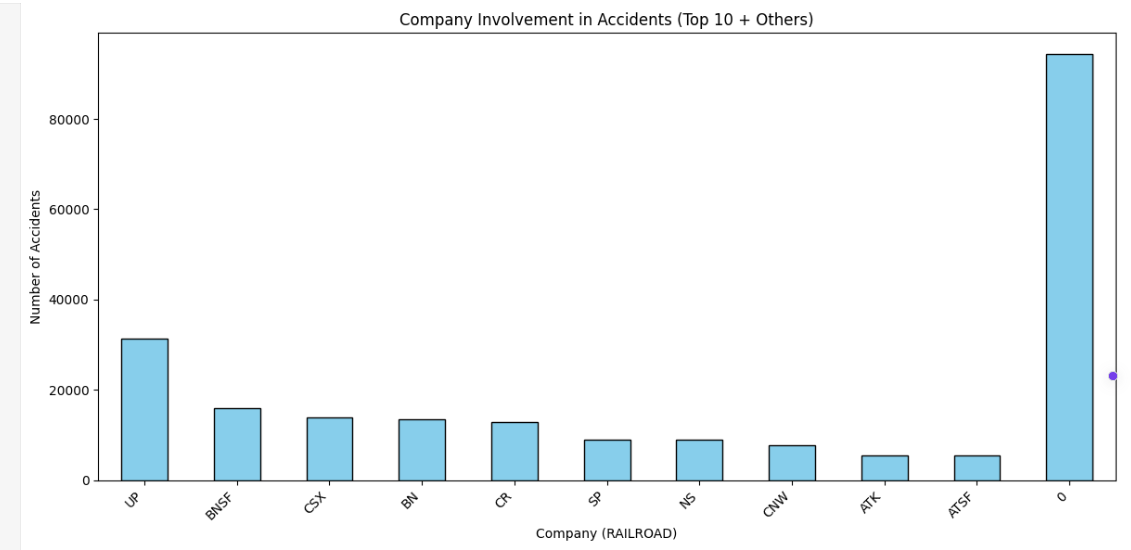


Figure 7 - Bar chart showing the number of incidents by railroad company.

Descriptive statistics of categorical variables

Here we are observing the RAILROAD feature that refers to the name of the company in charge of the train involved in the incident.

|  |  |
| --- | --- |
| Value Counts for Categorical Variables | |
| RAILROAD |  |
| UP | 530 |
| BNSF | 368 |
| NS | 328 |
| CSX | 286 |
| ATK | 70 |
| ..... | |
| KCT | 1 |
| ITHR | 1 |
| ISRR | 1 |
| IR | 1 |
| YVRR | 1 |
| Name: count, Length: 154, dtype : int64 | |

Figure 8 - Value count for categorical variables.

Value count for categorical variables

Above we can see the distribution of incidents across different railroad companies. From this analysis, we can observe that Union Pacific is involved in 530 incidents, which is higher compared to other railroad companies, followed by BNSF with 368 involved incidents. This provides insight into which companies need to improve their safety processes.

## 3.3 Exploratory Data Analysis

Below we have performed analysis on variables which are useful for our analysis. Exploratory data analysis is seen as an essential step in data analysis process, aiming to understand the patterns and relationships within the dataset [39]. It involves many steps like visual exploration, analytical exploration, identifying anomalies and patterns, data cleaning, and pre-processing. Visual exploration involves producing graphical representations of data to identify patterns, outliers, trends, and any other insights. Common visualizations used are histograms, box plots, and heatmaps [40]. We have tried Univariate analysis and Bivariate analysis which helped us to understand the dataset. Univariate analysis is used to gain insights from individual characteristics of variables and Bivariate analysis provides insights of a variable’s relation with another variable. The below plots help us to understand the individual variable. For example, incidents that happen on a day of the month. Here we can observe, on average, the number of incidents are similar throughout the month.

A graph with lines and numbers

Description automatically generated

Figure 9 - Distribution of incidents by day of month.

The below plot shows the distribution of incidents depending on weather conditions. Each bar on the X-axis represents a specific weather condition (e.g., rainy, snow, sunny), and the Y-axis represents the number of incidents. We can observe most of the incidents have taken place when the weather is clear, followed by cloudy and rain. Cloudy and rain makes sense, as it would impede visibility and make surface textures more slippery. The high number of incidents with clear weather is a bit alarming, as that would not add safety obstacles via natural factors. With further weather data to standardize the natural impacts, it is unclear whether this could be due to a far greater number of instances where the weather is clear for incidences to occur.

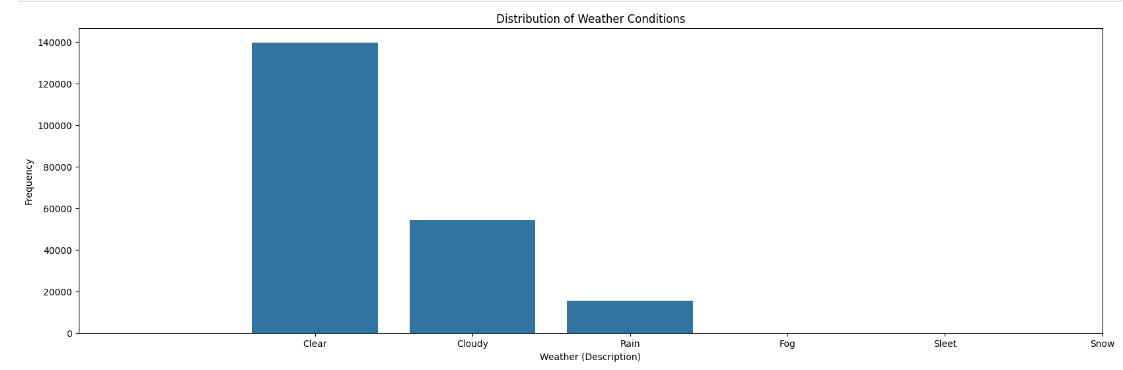


Figure 10 - Distribution of incidents by weather conditions.

The below histogram shows the relationship between the train speed and the number of incidents. The X-axis shows the specific range of train speed. The Y-axis represents the frequency of incidents. So, we can observe that trains of lower speed are involved in more incidents compared to fast moving trains.

A graph of blue bars

Description automatically generated

Figure 11 - Standardized distribution of incidents by train speed

The below histogram shows the frequency of specific types of incidents. It can be a great tool to identify the most common incidents. We can concentrate on those specific incidents and investigate the causes to take necessary preventive measures. It is clear from the below histograms that derailments are the most common railroad accidents and should be a point of focus. Collisions and crossings are less common, but often draw a high level of attention from the local public. These would make for further areas of focus for process improvement.

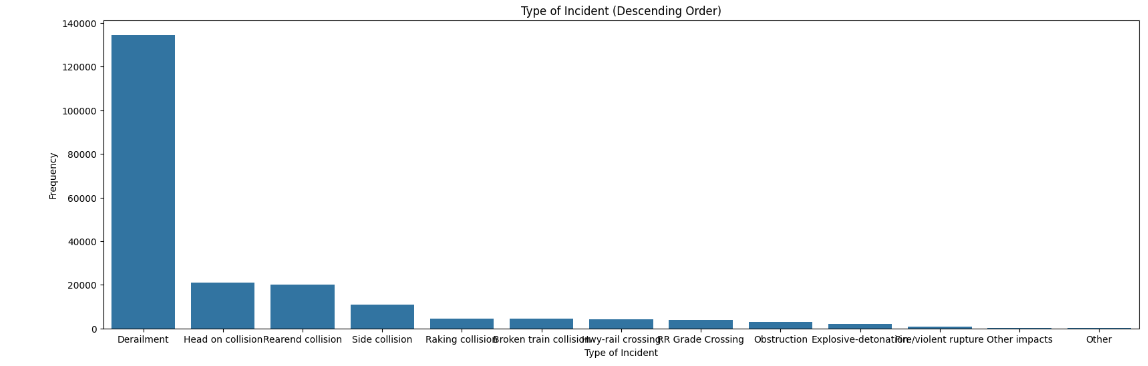


Figure 12 - Distribution of incidents by type.

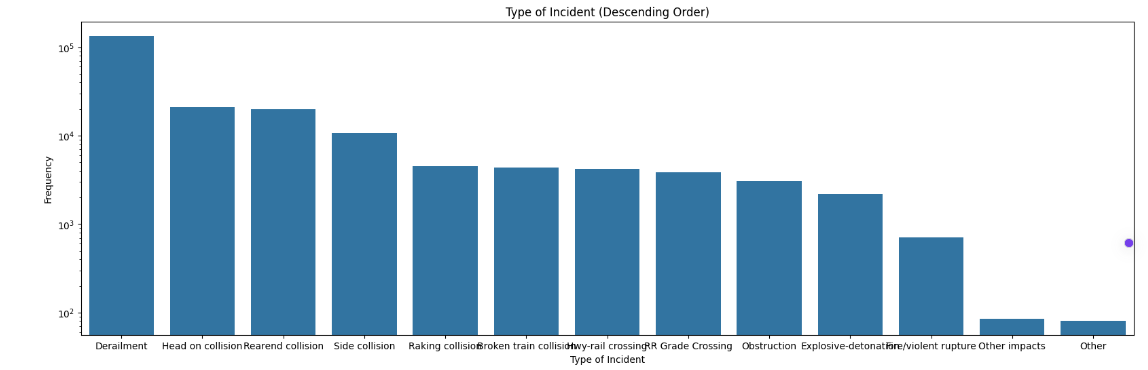


Figure 13 - Standardized distribution of incidents by type.

This histogram can be a valuable tool to explore the relationship between the temperature and the number of incidents. The tall bars correspond to the number of incidents that occurred. We can observe a few outliers where temperatures are at or above 200 degrees Fahrenheit that need to be excluded or replaced by mean or median.

A graph with blue squares

Description automatically generated

Figure 14 - Standardized distribution of incidents by temperature before outliers are removed.

This histogram is created after outliers are excluded. The tallest bar signifies the largest number of incidents occurring at that temperature, showing to be 75 degrees Fahrenheit.

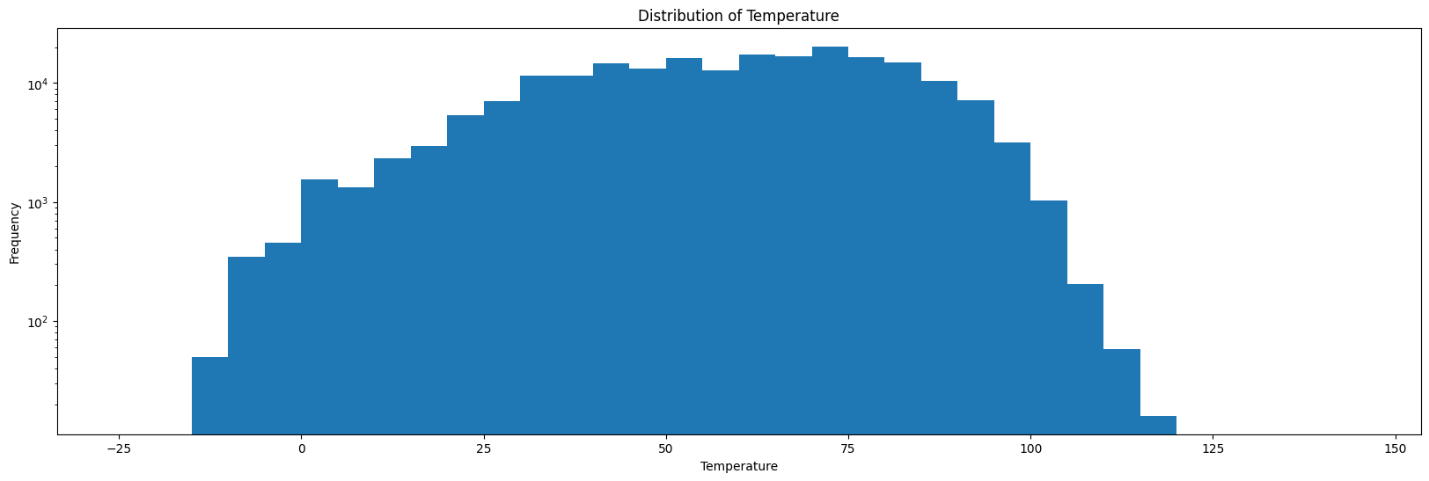


Figure 15 - Standardized distribution of incidents by temperature after outliers are removed.

From the above chart we can observe that most incidences occurred between -15 to 120 degrees Fahrenheit. A larger number of incidents occurred during warmer temperatures and the rate of incidents are highest in the range of 40 to 80 degrees Fahrenheit.

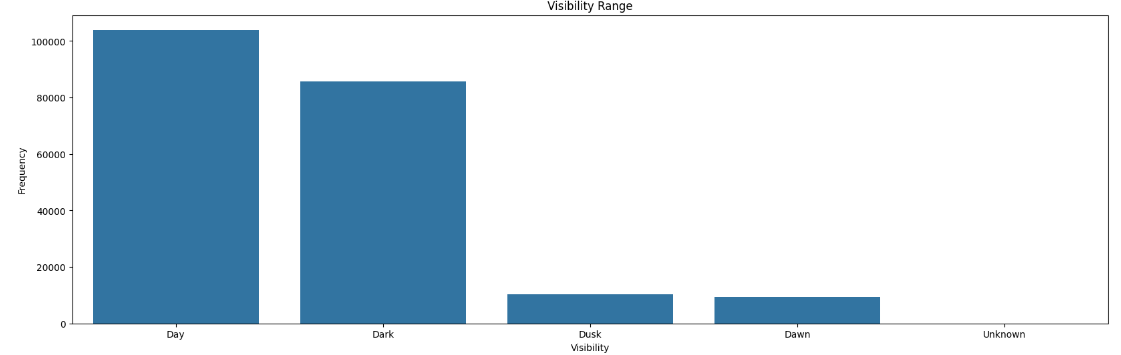


Figure 16 - Distribution of incidents by visibility range.

X-axis represents the visibility types such as day, dark, dusk, and dawn. The Y-axis represents the number of incidents that occurred under a specific visibility type. This shows that there are far more incidents during the Day and after Dark than during Dusk or Dawn. This could be due to the amount of time available for each situation, or there could be slower traffic during Dusk and Dawn.

Bivariate analysis

Here is the seaborn count plot where the distribution of weather against the incidents type is shown. With this plot we can visualize the relationship between incident types and weather conditions. The bar represents the frequency of incidents for a particular combination of incident type and weather conditions.

A graph of different colored lines

Description automatically generated

Figure 17 - Distribution of incidents for Weather Conditions vs. Incident Type.

In the above plot, it is clearly visible that most of the incidents occur in clear weather conditions and most of them are because of derailment. As previously discussed, the weather condition could be due to the availability of clear conditions.

In the same way, below is one more plot to check the relation between the train speed and incident type. Here we have a clear view of how the frequency of incidents are distributed across different train speed ranges. It also helps us to identify which kind of incidents are more prevalent in each speed bucket. It can also help identify speed buckets with incident patterns. For example, it is clear to see that Obstruction incidents occur frequently at very low speeds.

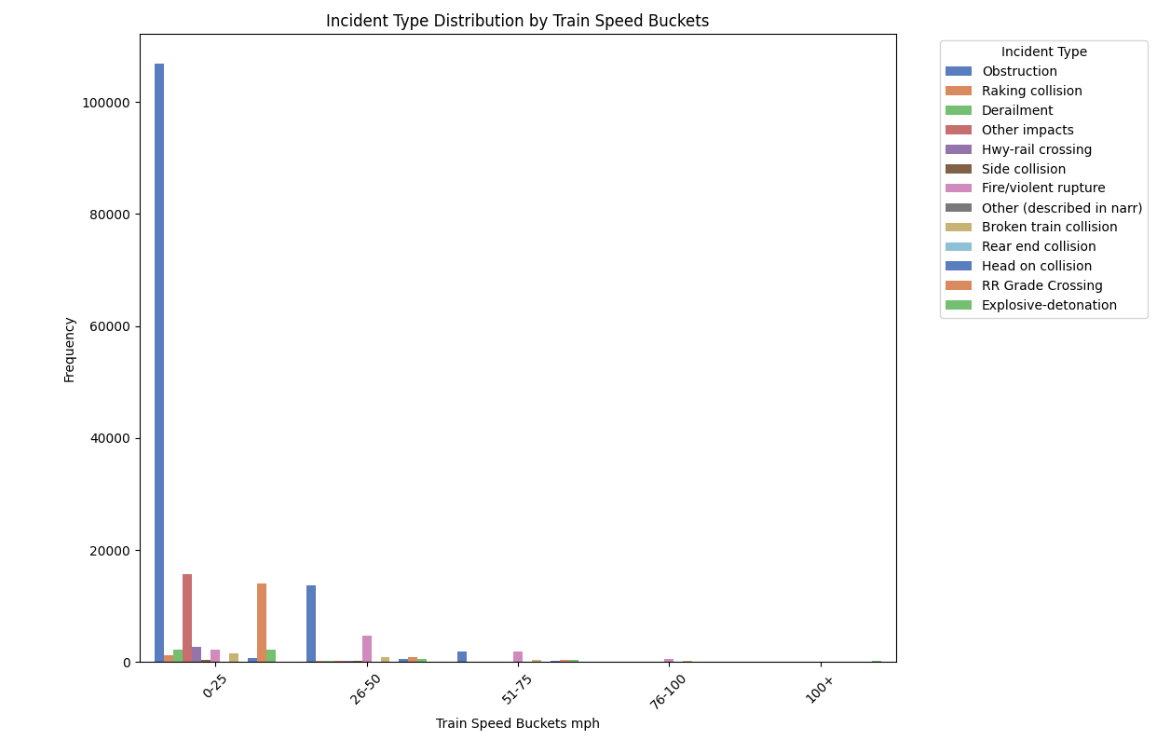


Figure 18 - Distribution of Incident Type by Train Speed Buckets.

## 3.4 Outliers and missing values

There are many reasons for missing data points, understanding that mechanism would help us to decide the best way to find the imputation method. For example, there could be missing values at random and in other cases not missing at random. There are many ways missing values can be handled, but in this project, we used mode and median method. Mode is very suitable for imputing categorical variables and median is a good choice for numerical variables as it is less sensitive to outliers compared to mean. Initially we checked the percentage of missing values for all the variables and tried to impute the variables which have more than 20 percent of missing values. Here it calculates the median value for the variable and fills the missing values with median value. The mode method was used for categorical variables, where the most frequent value is selected and used to fill the missing values.

A diagram with a blue rectangular bar

Description automatically generated with medium confidence

Figure 19 - Subset of numerical variables in shown via boxplot.

Here, we have deleted outliers for all variables. This removed fog, sleet, and snow due to their lack of frequency.

A diagram of weather and weather

Description automatically generated

Figure 20 - Box Plot of Weather

# **4 Methodology**

Here we are employing the SelectFromModel algorithm for feature selection with numerical variables. It utilizes the Random Forest Classifier internally and is a feature selection technique from scikit-learn, acting as a wrapper around other estimators to identify important features.

Initially, we provide the dataset and target variable to SelectFromModel. Internally, it fits the estimator on the data to learn the relationship between features and the target variable. Based on the learned feature importances, it selects a subset of features that the estimator deems most relevant for prediction. Finally, SelectFromModel provides the features that we can use for further analysis [41].

In the case of categorical variables, we must encode them and transform them to numerical variables, by which we can capture the relationships between the variables. We have used the Count Encoding technique for the encoding. It calculates the frequency of each unique category within a feature, then iterates through each column and creates a new column name (column\_name\_count).

Next, we use the Random Forest algorithm that ensembles multiple decision trees for our feature selection. Here the algorithm runs a process on the ensembled data and utilizes SelectFromModel class, which leverages the important features learned by Random Forest model.

We combine feature encoding (count encoding) to transform categorical variables, and for feature selection we used Random Forest to identify important features for our final model. Also, there are a few features for which the missing data is more than 40 percent, so we had dropped those features. We can impute those values if needed, but we felt that filling up a high percentage of data points may lead to biased data, loss of information, false patterns. Overall, categorical variables were reduced from 69 variables to 19 variables.

The internal model used here is the Random Forest Classifier. A Random Forest classifier is an ensemble learning model that combines predictions from multiple decision trees to improve accuracy. When we train a Random Forest model, it creates several training sets by randomly sampling from our original data, which is called boot strapping. For each tree, a different subset of features is selected from the total list of features. This in turn helps to prevent overfitting. When we present a new data point, all the trees in the forest throw a prediction and the most voted of them will be the final predicted class. Random Forest model mainly helps to reduce overfitting, high accuracy and handles missing values by itself, without any imputing [42].

For our analysis, the Random Forest classifier will consist of 100 decision trees and each tree will consider a subset of features at each split. The algorithm creates multiple decision trees, each trained on a random subset of features and data points (with replacement), then each tree makes a prediction. The final prediction is the majority vote from all the trees.

Feature Selection calculates feature importance during training. The importance of each feature reflects how much each feature contributes to splitting the data at decision nodes in the tree, indicating their relevance for classification [42].

SelectFromModel employs the Random Forest Classifier to learn feature importances on numerical data and the target variable. Based on these importances, it selects relevant features. This can lead to improved model performance, dimensionality reduction, and a better understanding of relationships between features.

Initially, we had a total of 75 numerical features, out of which 28 features were extracted.

Since we don’t have a target variable in the dataset, we are considering the type of incident as the target variable. If we can predict what type of incident may occur, there would be many opportunities for us to prevent it from happening or take necessary action to reduce the impact.

Once we have all the data, we can impute the missing values. And then identify the target variable which is ‘TYPE’ (type of accident) in our case, and we define the target variable as ‘y’. And we define the rest of all the features as ‘X’ by dropping the target variable from X.

Now we use train\_test\_split to split our data into training and testing data.

Here we are splitting the data into 80:20 ratio, where 80 percent of the data will be used to train the model and other 20 percent of data, we will use for testing. Now X\_train, y\_train are used to train our model and X\_test, y\_test is used to evaluate our model’s performance.

Here, we utilize ColumnTransformer to construct a pipeline for processing numerical and categorical features separately [43]. StandardScaler is applied to scale numerical features, which standardizes the features by subtracting the mean and dividing by the standard deviation, thereby enhancing the model's stability [44]. Categorical features are encoded using OneHotEncoder, which generates new binary features for each unique category [45]. We employ the logistic regression model, which is a linear classification algorithm that predicts the data point belonging to a specific class. In our scenario, it predicts the type of incident when all other features are provided to the model [46].

We are utilizing the logistics regression model, which is a linear classification algorithm that predicts the data point belonging to specific class. There are 2 types of logistics regression models. The first one is standard Logistics regression, which is used for binary classification where we have only 2 classes in our target variables. The other type is Multinomial Logistics Regression which can be used when we have more than 2 classes in our target variable. In our case, we must use the second one, as we have multiple classes. In our case, it predicts the type of incident, when we provide all other features to the model. Its working scenario is like random forest because it involves training a separate logistics regression model for each of the class from our variable. Each model distinguishes its class from all the rest combined. Then, during prediction for a new data point, each of the models predicts the probability of the data point to its assigned class. The class with the highest predicted probability of all the models is the final prediction, which can be called the confidence score. The model with the highest confidence score wins and assigns its class label as predicted class.

After working on the Logistics classifier model, we tried using neural networks with the keras function which have enhanced abilities. As logistics regression is a linear model, it can learn only linear relationships between features and the target variable. On other hand, neural networks are nonlinear models, so they can learn different types of complex relationships between the variables.

The logistics classifier requires manually engineering features to learn complex relationships, which can be time consuming. However, neural networks can learn complex relationships during the training process. This can be a major benefit if we do not have prior understanding of the relationships in our data. For this reason, neural networks can perform better compared to logistics regression when we have more features in the dataset. They can handle more parameters and can adapt better to complex decision boundaries.

Using Neural networks with Keras have a history of impressive performances in various machine learning tasks and have the capability to capture complex patterns and relationships that may not be effectively modelled by logistics classifiers. A few concepts that we will be using are:

In simple words, a neural network can be described as a machine learning model inspired by the human brain structure. They consist of interconnected layers of artificial neurons that process input data to make predictions. Each neuron performs a calculation and passes the result to next layer [47] [48] . The outline of a simple neural network can be seen in the figure below.

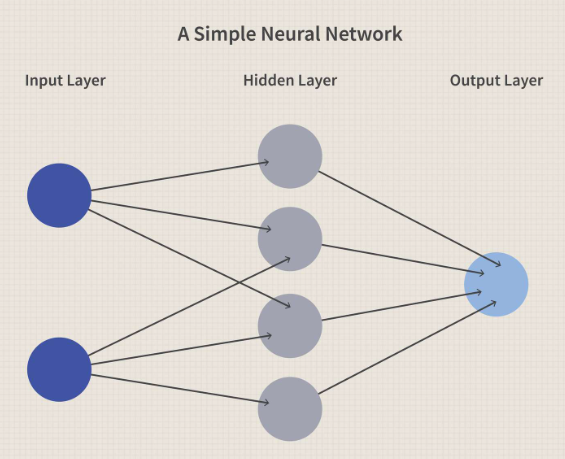


Figure 21 - A Simple Neural Network

Neural Networks are organized into layers, where information flows from input layer to output layer through hidden layers. The input layer is the point through which features enter the network. Hidden layers are considered as the engine of the model, as they perform all the calculations on data. They also learn complex patterns and relationships between the features. Finally, the output layer is where the final prediction is produced [49].

Neurons are building blocks of the model. They receive signals from other neurons, then a special function is applied to it, which generates output signal. This output signal will be an input for the next neuron in the following layer [49]. The activation function introduces non-linearity into the network, enabling the model to learn complex relationships in the data [50].

In Multiclass classification problems, like ours, the output layer needs to predict which category an input belongs to. So, it converts the raw output scores of the network into probabilities, where each value represents the likelihood of belonging to a particular class. This is known as the SoftMax Function [49] [51].

Our model consists of 2 hidden layers, each with 32 neurons, followed by an output layer using SoftMax activation for multi-class classification. The number of layers and neurons can be adjusted as per the complexity of the network. More layers and neurons can capture greater complexity but can also increase the risk of overfitting. Here the neural network model is trained to predict the type of incident that would occur based on the input features. Our model also utilized Keras functionality. Keras can be defined as an API written in python which can be used to simplify Deep learning. It uses libraries like tensor flow, providing a simple and user-friendly interface for building and training neural networks [52].

# **5 Results**

We evaluate the model by using many metrics. A couple of them can be found here:

Training accuracy: 0.8597

Testing accuracy: 0.8246

We can observe that our model achieved a training accuracy of 85.97% and testing accuracy of 82.46%, which indicates that the model performed well on the data it was trained and performs well on any unseen data. The difference between the training and testing accuracy is around 3.5% which is a generalization gap. This suggests that the model would have learned some specific patterns which are not necessary in testing data. It’s a good sign that the gap is not too large. The model is capturing underlying patterns from the data to a favourable extent. However, there may still be improvements to be made. We could attempt to reduce the generalization gap by using techniques like hyperparameter tuning or collecting more data. Also, we could check for overfitting of the model on the training data. In that case, we could use techniques like regularization to mitigate the overfitting.

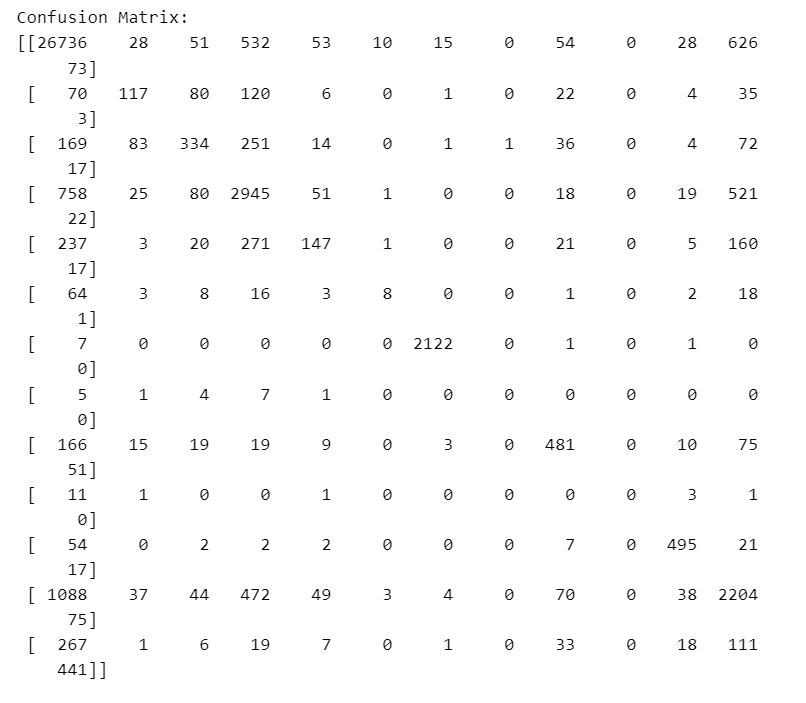


Figure 22 - Confusion matrix for logistic regression model before feature selection.

The confusion matrix is a table primarily used to describe the performance of a classification model on a dataset where the correct results are known. The values on the diagonal represent the number of correctly classified instances for each class. Therefore, higher values on the diagonal indicate better performance. Values above the diagonal represent false positives, indicating incorrect predictions by the model. High values here suggest poor performance for those classes. Values below the diagonal represent false negatives, where the model predicted incorrectly.

Our model appears to perform well for class 0, with 26,736 correct predictions on the diagonal and relatively few false positives and negatives. However, the model struggles with class 1. There are a fair number of correct predictions, but also some incorrect predictions. Additionally, there are false positives for classes 2, 3, and 4, indicating that the model confuses class 1 with other classes. Overall, the model seems to perform well for a few classes. There are ways to improve performance, such as cost-sensitive learning and feature engineering.

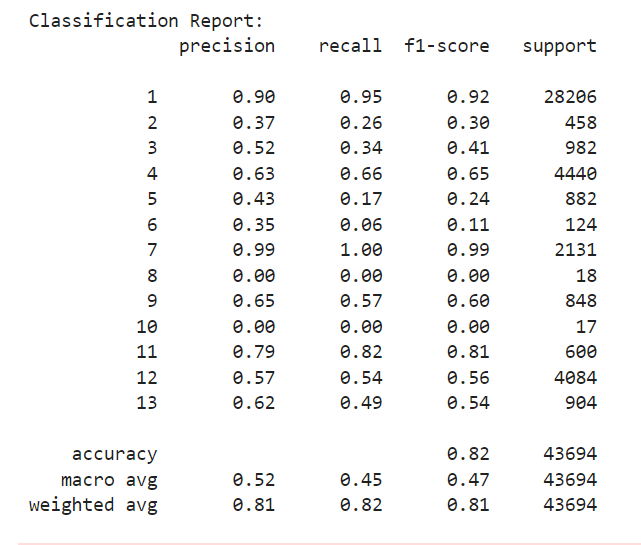


Figure 23 - Classification Report for logistic regression before feature selection.

This breakdown provides a comprehensive analysis of the model's performance.

Structure of the report:

* Support: This column displays the total number of true instances for each class in the test data.
* Precision: This metric represents the proportion of predicted positives that were correct. High precision indicates that the model performs well for that class and is adept at avoiding false positives.
* Recall: This metric represents the proportion of actual positives that were correctly identified by the model. Higher recall indicates the model's good performance for that class.
* F1-Score: This score combines precision and recall. A good F1-Score indicates a balanced performance between precision and recall [53] [54].

Class 1 exhibits high precision (0.90), high recall (0.95), and an F1-score of 0.92, indicating excellent performance. The model accurately predicts most instances of class 1. Classes 2-6 have lower F1-scores, indicating suboptimal performance for these classes. Similar to class 1, the model performs well for class 7. These classes demonstrate low values in all metrics, suggesting issues with the test data impacting performance. Classes 9, 11, 12, and 13 show moderate performance, with scores ranging from 0.4 to 0.8. This report allows us to analyze which classes the model excels in and where improvements are needed. It also leverages the model to predict potential accident types before they occur. By analyzing historical data, the model identifies patterns and conducts analyses for prediction.

There are numerous benefits to this approach:

* Predicting types of incidents in advance enables railroad companies and stakeholders to focus on a smaller number of variables while moving passengers and freight. This would allow them to focus on fewer variables, therefore being able to cover each more thoroughly.
* Companies can allocate resources appropriately and communicate with relevant departments based on accident types and prioritize safety standards.

## 5.1 Model Evaluation after Feature Selection

Here are the metrics we used to check on the performance of our model:

Training accuracy: 0.9451

Testing accuracy: 0.8317

Though the model is performing well on the training data, its performance on testing data is not great. So, there could be overfitting here. Techniques like regularization can be used to improve testing accuracy.

Confusion Matrix:

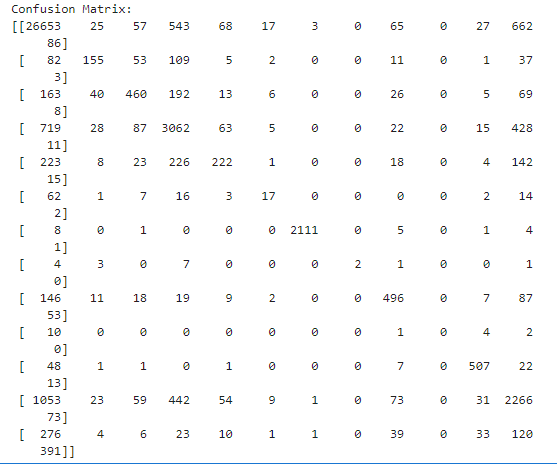


Figure 24 - Confusion matrix for logistic regression after feature selection.

Looking at row1-Class1, there are 26,653 correctly classified instances and relatively low misclassified ones. In row2, 155 are correctly classified, and there are significant number of misclassified points belonging to Class4. This indicates that the model is confusing Class2 with Class4 in some cases. Class7 is performing very well, as it is generally classified correctly.

Classification Report:

Below we can observe that Class1 and Class7 are performing very well. The majority of classes- 2, 3, 4, 5, 9, 11, 12, and 13 have moderate scores. This means at times they may make some wrong predictions, but they still make correct decisions in many cases. Class 6, 8, and 10 have low scores and misclassify in most cases, so these variables need to be worked on at later stages. So, we need to concentrate more on low performing classes and try to use techniques like class imbalance in case these classes are underrepresented in training data.

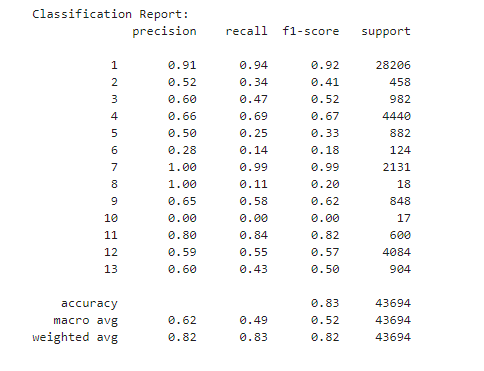


Figure 25 - Classification report for logistic regression after feature selection.

This approach leverages the model to predict potential accident types before they occur. By running through the historical data, the model runs different patterns and analysis to predict.

We can derive a fair amount of information from this model. By predicting the most likely type of incidents in advance, railroad companies and stake holders can take necessary preventive measures to reduce the affect or avoid the accidents. Also, companies can adjust resource allocation and communicate with concerned departments based on accident types. Early accident prediction can help companies to prioritize safety standards.

To improve this study, we need to address imbalanced classes. Techniques like Oversampling, Undersampling, or SMOTE can be employed to balance the classes, leading to improved predictions and better model accuracy.

Find below results of our neural networks model:

Epoch 1/10

4370/4370 ━━━━━━━━━━━━━━━━━━━━ 923s 211ms/step - accuracy: 0.7883 - loss: 0.6955 - val\_accuracy: 0.8459 - val\_loss: 0.4713

Epoch 2/10

4370/4370 ━━━━━━━━━━━━━━━━━━━━ 928s 212ms/step - accuracy: 0.9601 - loss: 0.1376 - val\_accuracy: 0.8523 - val\_loss: 0.5268

Epoch 3/10

4370/4370 ━━━━━━━━━━━━━━━━━━━━ 899s 206ms/step - accuracy: 0.9967 - loss: 0.0131 - val\_accuracy: 0.8374 - val\_loss: 0.6891

Epoch 4/10

4370/4370 ━━━━━━━━━━━━━━━━━━━━ 18273s 4s/step - accuracy: 0.9992 - loss: 0.0030 - val\_accuracy: 0.8474 - val\_loss: 0.8521

Epoch 6/10

4370/4370 ━━━━━━━━━━━━━━━━━━━━ 878s 201ms/step - accuracy: 0.9993 - loss: 0.0021 - val\_accuracy: 0.8502 - val\_loss: 0.8459

Epoch 7/10

4370/4370 ━━━━━━━━━━━━━━━━━━━━ 959s 220ms/step - accuracy: 0.9998 - loss: 8.2234e-04 - val\_accuracy: 0.8538 - val\_loss: 0.9172

Epoch 10/10

4370/4370 ━━━━━━━━━━━━━━━━━━━━ 878s 201ms/step - accuracy: 0.9999 - loss: 4.4620e-04 - val\_accuracy: 0.8460 - val\_loss: 0.9466

Training accuracy: 0.9691

Testing accuracy: 0.8432

Here the result is output from our neural network model using keras over 10 epochs. Let’s understand it better:

Epoch 1/10: Indicates the current number of Epoch out of total 10 epochs. The number of steps completed out of the total steps was 4370/4370. The time taken for each epoch was 932s or 211ms/step.

Accuracy: Accuracy of the model on training data at current epoch. Indicates the proportion of correctly classified samples.

Loss: This shows the error of the model’s predictions.

val\_accuracy: The accuracy of the model on validation data at the current epoch. Indicates how well the model works on unseen data.

val\_loss: The loss of the model on validation data at the current epoch. Indicates the error of the model on unseen data.

Confusion Matrix:

The number of correctly classified instances at row1-Class1 are 27113 and relatively low misclassified ones. The neural networks model has better accuracy compared to the logistic regression model.

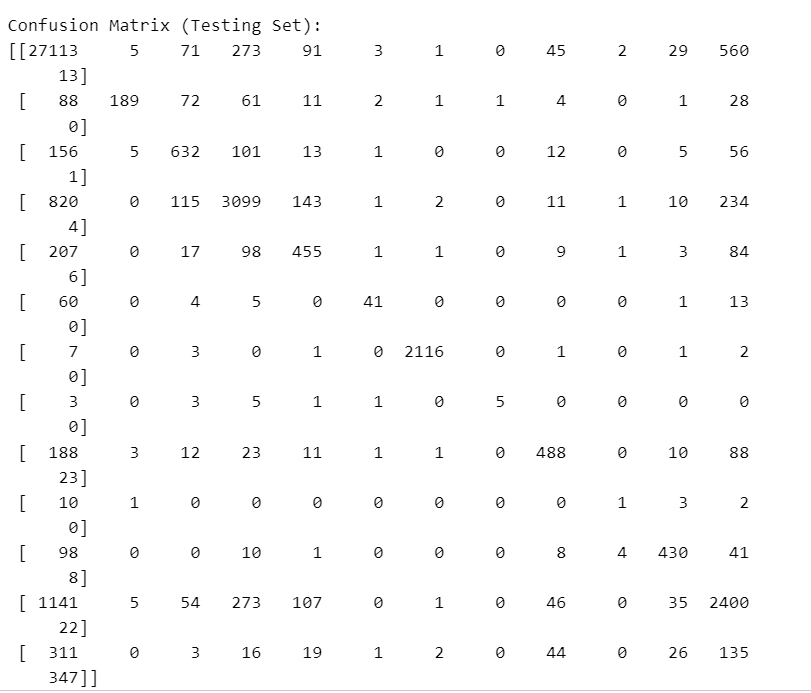


Figure 26 - Confusion matrix for neural network

Classification Report:

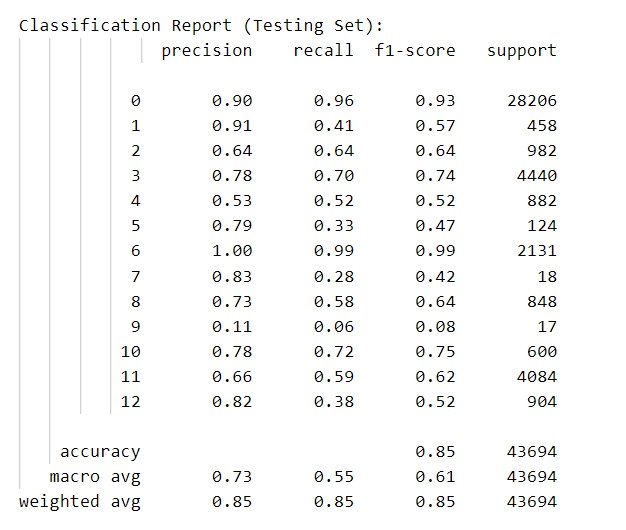


Figure 27 - Classification report of the testing set for neural network model

The classification report shows the accuracies for Classes with higher number of datapoints have a moderate f1 score with the final accuracy of this model being 85%.

The training and validation numbers indicate that the model is learning well. However, the increase in validation loss towards the end of training suggests potential overfitting. This can be addressed later by using some techniques like regularization. We can work on hyperparameter tuning to try a different number of epochs to increase the performance of the model. Though the training accuracy is good, validation numbers of projects are a different story. So, this may not be a good model for our existing dataset.

## 5.2 Stepwise Regression Technique

One more technique to reduce our features is Stepwise regression, where it selects the best subset of variables that best predicts the outcome variable. Also, it would remove all other variables from the model based on their statistical significance using criteria such as p-value, Bayesian information criterion, and a few other measures of model fit.

The stepwise algorithm usually searches for all the combinations of features or subsets of features. As the number of features are high within our dataset, the number of possible combinations grows exponentially and leads to a significant increase in computation time. For each subset of features, the model needs to train and predict using cross-validation or other validation techniques. This is also highly time-consuming. The runtime also depends on the computational resources available.

The process starts with an empty model and iteratively adds one predictor at a time. At each step, the predictor variable that provides the best improvement in the model is added, based on predefined criterion. This is called forward selection. Once all variables are added in the model, the algorithm begins removing one predictor variable at a time. At each step, the variable with the least impact on the model fit is removed, based on predefined criterion. This process repeats until no further variables can be removed and is called backward selection.

Now stepwise regression combines forward selection and backward selection. It starts with an empty model and alternates between adding and removing variables until no further improvements in model fit can be achieved. This is mostly used for feature selection in linear regression models when we are dealing with more features. So, this model helps to identify the most important features from the data set, while reducing the overfitting at the same time. However, there are some limitations with this model.

Stepwise models can lead to overfitting and the process of iteratively testing multiple variables can inflate the type 1 error rate, leading to unreliable results. The final model may depend more on the dataset and the specific selection criteria used, resulting in instability and lack of generalizability.

Despite all the limitations, this model can be useful tool for exploratory analysis and initial model building [55] [56].

However, this model consumes a lot of time when the number of features is high. In our case, the run time was too long to complete the model, so we could not find the final optimal subset of features and decided not to move forward with this model.

# **6 Conclusions**

## 6.1 Logistics Regression Model

The Logistics Regression model is achieving a training accuracy of 85.97% and validation accuracy of 82.45%. It shows a generalization gap of around 3.5%, indicating that it may have learned some specific patterns in training data that are not present in validation data. Interpretation of the confusion matrix and classification report suggests that the model performs well for few classes but struggles with others. Overall, the Logistics Regression model provides a good tool to predict incident types but there is some room for improvement, mainly for the classes with low performance.

## 6.2 Neural Network Model

This model achieved a training accuracy of 96.91% and validation accuracy of 84.32%. Though the training accuracy is high, there are signs of overfitting as indicated by the increase in validation loss towards the end of training. The model’s performance varies across different classes, with some showing excellent performance and others showing moderate to poor performance. Further hyperparameter tuning and regularization techniques may improve the model’s performance.

The plot below shows the comparison of accuracy of the three different machine learning models for classification task. As you can see, the Neural Network models appear to have the highest accuracy among the three. This suggests the neural network model is most affective for the dataset we have.

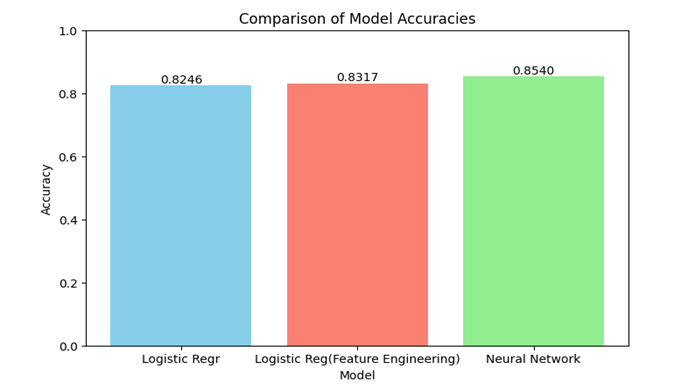


Figure 28 - Comparison of model accuracies.

Both models are showing promising results. The Logistics Regression model provides a simple and interpretable solution while the neural network model offers the potential to capture complex patterns. Further improvements can be made by fine tuning, work on over-fitting, and exploring a few techniques to balance classes with low performance.

## 6.3 New Questions and Future Exploration:

Class Imbalance: Addressing class imbalance in the dataset could improve the performance of the model, mainly for classes with less instances.

Feature Engineering: Engineering all the features could provide more information, which would make a better model.

Time-Series Analysis: Analyzing trends over time could provide valuable insights into incident patterns and help in forecasting future incidents.

External Factors: Exploring the effect of external factors like weather, maintenance schedules and human factors on incidents occurrences could enhance the model’s predictive possibilities.

We could improve our models, especially Neural Network model, by utilizing techniques like regularization and hyperparameter tuning. Other techniques like Oversampling, Undersampling or SMOTE can be explored to address issues with data imbalance within classes. Would other machine learning models like Random Forest or Support Vector Machine be more effective for this task?

If more time permits, further exploration can include:

• Implementing the techniques mentioned above to address data imbalance and overfitting.

• Work on other machine learning algorithms.

• Visualize the relationships between the variables and the target variable, to gain deeper insights.

Furthermore, there has recently been a new agreement within the railroad industry to gather data on “Close Calls,” or instances where an incident nearly occurred but was avoided [57]. We believe that utilizing that data alongside ours would allow the model to be trained metrics for non-incidents as well as situations where incidents occurred. That would, in theory, allow for a model that could classify the probability of an incident occurring rather than the type or severity of a potential incident should it occur.

# **10 References**

|  |  |
| --- | --- |
| [1] | X. Liu, "Statistical Causal Analysis of Freight-Train Derailments in the United States," *Journal of Transportation Engineering, Part A: Systems,* vol. 143, no. 2, 2016. |
| [2] | Kesler, Kevin, Bing and Alan, Railroad R\&D 1970-2017, vol. 50978, 2018, p. V001T11A001. |
| [3] | Aldrich and Mark, Back on Track: American Railroad Accidents and Safety, JHU Press, 2018. |
| [4] | Bridgelall, Raj, P. Lu, J. Szmerekovsky and B. Bhardwaj, Ranking risk factors in financial losses from railroad incidents: a machine learning approach, vol. 2677, SAGE Publications Sage CA: Los Angeles, CA, 2023, pp. 299--309. |
| [5] | S. Detrow, E. Ryan and T. Ermyasfeb, "NCPR," 5 February 2024. [Online]. Available: https://www.northcountrypublicradio.org/news/npr/1228772709/one-year-after-the-toxic-train-derailment-east-palestine-is-a-town-divided. [Accessed 10 February 2024]. |
| [6] | M. Bulakh, A. Okorokov and D. Baranovskyi, "Risk system and railway safety," in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing, 2021, p. 042074. |
| [7] | D. B. S., Human Reliability and Error in Transportation Systems, New York: Springer Science & Business Media, 2007. |
| [8] | S. &. V. A. Reinach, "Application of a human error framework to conduct train accident/incident investigations. Accident Analysis & Prevention," vol. 38, no. 2, pp. 396-406, 2006. |
| [9] | S. J. &. G. J. B. Reinach, "Railroad yard safety," *Perspectives from labor and management. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* vol. 46, no. 13, pp. 1201-1204, 2002. |
| [10] | T. Farrington-Darby, L. Pickup and J. R. Wilson, Safety culture in railway maintenance, vol. 43, Elsevier, 2005, pp. 39--60. |
| [11] | . F. Nemry and H. Demirel, Impacts of Climate Change on Transport: A focus on road and rail transport infrastructures, European Commission Luxembourg, 2012. |
| [12] | B. Strauch, Investigating human error: Incidents, accidents, and complex systems, CRC Press, 2017. |
| [13] | N. Esmaeeli, F. Sattari, L. Lefsrud and R. Macciotta, "Assessing the Risks Associated with the Canadian Railway System Using a Safety Risk Model Approach," *Transportation Research Record,* vol. 0, no. 0, p. 03611981231176549, 2023. |
| [14] | A. Lotfi and M. S. Virk, "Railway operations in icing conditions: a review of issues and mitigation methods," *Public Transport,* vol. 15, no. 3, pp. 747--765, 2023. |
| [15] | P. Rungskunroch, J. Anson and S. Kaewunruen, "Benchmarking on railway safety performance using Bayesian inference, decision tree and petri-net techniques based on long-term accidental data sets," *Reliability Engineering & System Safety,* vol. 213, p. 107684, 2021. |
| [16] | M. J. Koetse and P. Rietveld, The impact of climate change and weather on transport: An overview of empirical findings, vol. 14, Transportation Research Part D: Transport and Environment, 2009, pp. 205--221. |
| [17] | G. Kefalidou, Identifying rail asset maintenance processes: a human-centric and sensemaking approach, vol. 20, Springer, 2018, pp. 73--92. |
| [18] | H. Li, D. Parikh, Q. He, B. Qian, Z. Li, D. Fang and A. Hampapur, "Improving rail network velocity: A machine learning approach to predictive maintenance," *Transportation Research Part C: Emerging Technologies,* vol. 45, no. 0968-090X, pp. 17-26, 2014. |
| [19] | M. Binder, V. Mezhuyev and M. Tschandl, "Predictive Maintenance for Railway Domain: A Systematic Literature Review," *IEEE Engineering Management Review,* vol. 51, no. 2, pp. 120-140, 2023. |
| [20] | F. Peng, S. Kang, X. Li, Y. Ouyang, K. Somani and D. Acharya, "A Heuristic Approach to the Railroad Track," *Computer-Aided Civil and Infrastructure Engineering,* vol. 26, no. 2, pp. 129-145, 2011. |
| [21] | S. Kaewunruen, J. Sresakoolchai and Y. Lin, "Digital twins for managing railway maintenance and resilience," *Open Res Eur.,* vol. 2, pp. 1-91, 2021. |
| [22] | J. H. J. Z. C. J. S. H. &. P. N. Xie, "Systematic literature review on data-driven models for predictive maintenance of railway track," *Implications in geotechnical engineering. Geosciences,* vol. 10, no. 11, p. 425, 2020. |
| [23] | E. J. S. O. I. G. K. &. Q. A. Palin, " Implications of climate change for railway infrastructure.," *Wiley Interdisciplinary Reviews: Climate Change,* vol. 12, no. 5, 2021. |
| [24] | H. Alawad, S. Kaewunruen and M. An, "A Deep Learning Approach Towards Railway Safety Risk Assessment," *IEEE Access,* vol. 8, pp. 102811-102832, 2020. |
| [25] | H. Ebrahimi, F. Sattari, L. Lefsrud and R. Macciotta, "A machine learning and data analytics approach for predicting evacuation and identifying contributing factors during hazardous materials incidents on railways," *Safety Science,* vol. 164, p. 106180, 2023. |
| [26] | F. F. P. P. Y. J. &. F. L. Saccomanno, "Estimating countermeasure effects for reducing collisions at highway–railway grade crossings. Accident Analysis & Prevention," vol. 39(2), 2007. |
| [27] | S. K. S. A. M. &. S. J. M. Dindar, "Bayesian Network-based probability analysis of train derailments caused by various extreme weather patterns on railway turnouts. Safety science," pp. 20-30, 2018. |
| [28] | S. F. v. F. J. &. B. R. Møller, "Robust methods for multivariate data analysis. Journal of Chemometrics: A Journal of the Chemometrics Society," pp. 549-563, 2005. |
| [29] | T. T. R. F. J. H. &. F. J. H. Hastie, "The elements of statistical learning: data mining, inference, and prediction," vol. 2, pp. 1-758, 2009. |
| [30] | H. R. &. T. S. I. Crossing, "Federal Railroad Administration.," 2005. |
| [31] | "Federal Railroad Safety Act of 1970," [Online]. Available: https://www.congress.gov/91/statute/STATUTE-84/STATUTE-84-Pg971.pdf. |
| [32] | "Code of Federal Regulations Title 49 - Transportation," [Online]. Available: https://www.ecfr.gov/current/title-49. |
| [33] | "FRA Advisory Circulars and Bulletins," [Online]. Available: https://railroads.dot.gov/railroad-safety. |
| [34] | M. Coplen and D. Sussman, "Fatigue and alertness in the United States railroad industry part II: fatigue research in the Office of Research and Development at the Federal Railroad Administration," *Transportation Research Part F: Traffic Psychology and Behaviour,* vol. 3, no. 4, pp. 221-228, 2000. |
| [35] | Y. a. H. L. a. W. H. a. Z. H. Liao, "Prediction models for railway track geometry degradation using machine learning methods: a review," *Sensors,* vol. 22, p. 7275, 2022. |
| [36] | F. A. a. G. D. J. a. S. R. a. A. L. Awad, "Predicting urban rail transit safety via artificial neural networks," *Safety science,* vol. 167, p. 106282, 2023. |
| [37] | "Monetary Threshold Notice | FRA," [Online]. Available: https://railroads.dot.gov/safety-data/forms-guides-publications/guides/monetary-threshold-notice. [Accessed 24 January 2024]. |
| [38] | "Rail Equipment Accident/Incident Data (Form 54) | Department of Transportation - Data Portal," [Online]. Available: https://data.transportation.gov/Railroads/Rail-Equipment-Accident-Incident-Data-Form-54-/85tf-25kj/about\_data. [Accessed 21 March 2024]. |
| [39] | J. W. Tukey, "Exploratory data analysis," *Addison-Wesley,* 1977. |
| [40] | W. N. Venables and B. D. Ripley, "Modern applied statistics with S," *Springer Science & Business Media,* 2002. |
| [41] | "SciKit Learn - SelectFromModel," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature\_selection.SelectFromModel.html. |
| [42] | L. Breiman, "Random forests," *Machine learning,* vol. 45, no. 3, pp. 5-32. |
| [43] | "SciKit Learn - ColumnTransformer," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.compose.ColumnTransformer.html. |
| [44] | F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research 12,* pp. 2825-2830, 2011. |
| [45] | "Scikit learn - OneHotEncoder," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html. |
| [46] | G. James, D. Witten, T. Hastie and R. Tibshirani, "An introduction to statistical learning," *Springer,* vol. 112, no. 1, p. Chapter 4, 2013. |
| [47] | Y. LeCun, Y. Begio and G. Hinton, Deep learning. Nature, 2015, pp. 436-444. |
| [48] | C. Bishop, Neural networking for pattern recognition, Oxford university press, 1995. |
| [49] | s. Hochreiter and J. Schmidhuber, "Long Short-term memory," *Neural computation,* 1997. |
| [50] | s. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "A simple way to prevent network from overfitting," *Journal of Machine Learning Research,* p. 15, 2014. |
| [51] | D. Rumelhart, G. Hinton and R. Williams, Learning representation by back-propagating errors, vol. 323(6088), Nature, 1986, pp. 533-536. |
| [52] | A. Geron, "Hands-On Machine Learning With Scikit-Learn, Karas & TensorFlow," 2017. |
| [53] | D. M. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & co.," *Journal of Machine Learning Technologies,* vol. 2, no. 1, pp. 3-11. |
| [54] | "Scikit Learn," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification\_report.html. |
| [55] | D. H. Jr., S. Lemehow and R. Sturdivant, "Applied Logistic Regression," 2013. |
| [56] | F.E. Harrell Jr., "Regression Modeling Strategies: With Application to Linear Models, Logistic Regression, and Survival Analysis," 2015. |
| [57] | [Online]. Available: https://bnsf.com/news-media/news-releases/newsrelease.page?relId=bnsf-railway-atda-and-fra-reach-c3rs-safety-agreement. |
| [58] | Z. Zhang, X. Liu and K. Holt, "Positive Train Control (PTC) for railway safety in the United States: Policy developments and critical issues," *Utilities Policy,* vol. 51, pp. 33-40, 2018. |
| [59] | E. a. B. A. a. {. A. a. S. P. Koz{\l}owski, "Classification Trees in the Assessment of the Road--Railway Accidents Mortality," *Energies,* vol. 14, 2021. |
| [60] | H. a. H. S.-Y. a. P. K. a. L. H. a. K. T. Lee, "Real-time hybrid deep learning-based train running safety prediction framework of railway vehicle," *Machines,* vol. 9, p. 130, 2021. |
| [61] | P. S. Mosteller and J. W. Tukey, "Data analysis and regression: A second course in statistics," *Addison-Wesley,* 1977. |
| [62] | D. C. Montgomery and E. A. Peck, "Introduction to linear regression analysis," *John Wiley & Sons,* 2018. |
| [63] | S. Vijayarani and S. Srivatsa, "An Effective Approach for Preprocessing Data in Data Mining," *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS),* vol. 1, no. 2, pp. 34-38, 2012. |

# **11 Appendix**

## 11.1 Data File Structure and Field Input Specifications as provided by the Department of Transportation [38]

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| FIELD NAME | FIELD POSITION | FIELD TYPE | FIELD LENGTH | DEFINITION | BLOCK # ON FORM 6180.54 | CONVERSION |
| amtrak | 1 | A/N | 1 | Amtrak involvement |  |  |
| iyr | 2 - 3 | A/N | 2 | year of incident | 5 |  |
| imo | 4 - 5 | A/N | 2 | month of incident | 5 |  |
| railroad | 6 - 9 | A | 4 | railroad code (Reporting RR) | 1a |  |
| incdtno | 10 - 19 | A/N | 10 | railroad assigned number | 1b |  |
| iyr2 | 20 - 21 | A/N | 2 | year of incident | 5 |  |
| imo2 | 22 - 23 | A/N | 2 | month of incident | 5 |  |
| rr2 | 24 - 27 | A | 4 | railroad code (Other RR involved) | 2a |  |
| incdtno2 | 28 - 37 | A/N | 10 | other railroad assigned number | 2b |  |
| iyr3 | 38 - 39 | A/N | 2 | year of incident | 5 |  |
| imo3 | 40 - 41 | A/N | 2 | month of incident | 5 |  |
| rr3 | 42 - 45 | A | 4 | railroad code (RR responsible for track maintenance) | 3a |  |
| incdtno3 | 46 - 55 | A/N | 10 | railroad assigned number | 3b |  |
| dummy1 | 56 - 59 | A/N | 4 | blank data expansion field |  |  |
| gxid | 60 - 66 | A/N | 7 | grade crossing id number | 4 |  |
| year | 67 - 68 | A/N | 2 | year of accident / incident | 5 |  |
| month | 69 - 70 | A/N | 2 | month of incident | 5 |  |
| day | 71 - 72 | A/N | 2 | day of incident | 5 |  |
| timehr | 73 - 74 | N | 2 | hour of incident | 6 |  |
| timemin | 75 - 76 | N | 2 | minute of incident | 6 |  |
| ampm | 77 - 78 | A/N | 2 | am or pm | 6 |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| FIELD NAME | FIELD POSITION | FIELD TYPE | FIELD LENGTH | DEFINITION | BLOCK # ON FORM 6180.54 | CONVERSION |
| type | 79 - 80 | A/N | 2 | Type of accident: 01=Derailment 02=Head on collision 03=Rearend collision 04=Side collision 05=Raking collision 06=Broken train collision 07=Hwy-rail crossing 08=RR Grade Crossing 09=Obstruction 10=Explosive-detonation 11=fire/violent rupture 12=Other impacts 13=Other (described in narr) | 7 |  |
| cars | 81 – 83 | N | 3 | # of cars carrying hazmat | 8 |  |
| carsdmg | 84 – 86 | N | 3 | # of hazmat cars damaged or derailed | 9 |  |
| carshzd | 87 – 89 | N | 3 | # of cars that released hazmat | 10 |  |
| evacuate | 90 – 95 | N | 6 | # of persons evacuated | 11 |  |
| division | 96 – 115 | A/N | 20 | railroad division | 12 | For accidents prior to June 1, 2011. For accidents occurring June 1, 2011, or later this field will be blank. |
| station | 116–135 | A/N | 20 | Nearest city and town | 13 |  |
| milepost | 136-141 | A/N | 6 | Milepost # | 14 |  |
| state | 142-143 | A/N | 2 | FIPS state code | 15 |  |
| temp | 144-146 | N | 3 | Temperature in degrees Fahrenheit | 17 |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| FIELD NAME | FIELD POSITION | FIELD TYPE | FIELD LENGTH | DEFINITION | BLOCK # ON FORM 6180.54 | CONVERSION |
| visiblty | 147 | A/N | 1 | Daylight period:  1=dawn 2=day 3=dusk 4=dark | 18 |  |
| weather | 148 | A/N | 1 | Weather conditions: 1=clear 2=cloudy 3=rain 4=fog 5=sleet 6=snow | 19 |  |
| trnspd | 149-151 | A/N | 3 | Speed of train in miles per hour: Blank=unknown | 28 |  |
| typspd | 152 | A/N | 1 | Train speed type: E=estimated R=recorded Blank=unknown | 28 |  |
| trnnbr | 153-156 | A/N | 4 | Train id number | 27 |  |
| trndir | 157 | A/N | 1 | Train direction: 1=north 2=south 3=east 4=west | 24 |  |
| tons | 158-162 | N | 5 | Gross tonnage, excluding power units | 29 |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| FIELD NAME | FIELD POSITION | FIELD TYPE | FIELD LENGTH | DEFINITION | BLOCK # ON FORM 6180.54 | CONVERSION |
| typeq | 163 | A/N | 1 | Type of consist: 1=freight train 2=passenger train-pulling\* 3=commuter train-pulling\* 4=work train 5=single car 6=cut of cars 7=yard/switching 8=light loco(s) 9=main./inspec. Car A=spec. MoW eq B=passenger train-pushing\* C=commuter train-pushing\* D=EMU\*\* E=DMU\*\* | 25 | \*As of June 1, 2011 - Name Changes  \*\*As of June 1, 2011 - New selections not available before June 1, 2011 |
| eqatt | 164 | A/N | 1 | Equipment attended 1=yes 2=no | 26 |  |
| trkname | 165-184 | A/N | 20 | Track identification | 21 |  |
| trkclas | 185 | A/N | 1 | FRA track class: 1-9, X | 22 |  |
| trkdnsty | 186-191 | A/N | 6 | Annual track density-gross tonnage in millions | 23 |  |
| typtrk | 192 | A/N | 1 | Type of track: 1=main 2=yard 3=siding 4=industry | 20 |  |
| rrcar1 | 193-196 | A/N | 4 | Car initials (first involved) | 31a(1) |  |
| carnbr1 | 197-202 | A/N | 6 | Car number (first involved) | 31a(1) |  |
| positon1 | 203-205 | A/N | 3 | Car position in train (first involved) | 31b(1) |  |
| loaded1 | 206 | A/N | 1 | Car loaded or not (first involved); Y=Yes N=No Blank=Not Applicable | 31c(1) |  |
| rrcar2 | 207-210 | A/N | 4 | Car initials (causing) | 31a(2) |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| FIELD NAME | FIELD POSITION | FIELD TYPE | FIELD LENGTH | DEFINITION | BLOCK # ON FORM 6180.54 | CONVERSION |
| carnbr2 | 211-216 | A/N | 6 | Car number (causing) | 31a(2) |  |
| positon2 | 217-219 | A/N | 3 | Car position in train (causing) | 31b(2) |  |
| loaded2 | 220 | A/N | 1 | Car loaded or not (causing): Y=Yes N=No Blank=Not applicable | 31c(2) |  |
| headend1 | 221 | N | 1 | # of head end locomotives | 34a(1) |  |
| midman1 | 222 | N | 1 | # of mid train locomotives, manual | 34b(1) |  |
| midrem1 | 223 | N | 1 | # of mid train locomotives, remote | 34c(1) |  |
| rman1 | 224 | N | 1 | # of rear end locomotives, manual | 34d(1) |  |
| rrem1 | 225 | N | 1 | # of rear end locomotives, remote | 34e(1) |  |
| headend2 | 226 | N | 1 | # of head end locomotives, derailed | 34a(2) |  |
| midman2 | 227 | N | 1 | # of mid train locomotives, manual, derailed | 34b(2) |  |
| midrem2 | 228 | N | 1 | # of mid train locomotives, remote, derailed | 34c(2) |  |
| rman2 | 229 | N | 1 | # of rear end locomotives, manual, derailed | 34d(2) |  |
| rrem2 | 230 | N | 1 | # of rear end locomotives, remote, derailed | 34e(2) |  |
| loadf1 | 231-233 | N | 3 | # of loaded freight cars | 35a(1) |  |
| loadp1 | 234-236 | N | 3 | # of loaded passenger cars | 35b(1) |  |
| emptyf1 | 237-239 | N | 3 | # of empty freight cars | 35c(1) |  |
| emptyp1 | 240-242 | N | 3 | # of empty passenger cars | 35d(1) |  |
| caboose1 | 243-245 | N | 3 | # of cabooses | 35e(1) |  |
| loadf2 | 246-248 | N | 3 | # of derailed loaded freight cars | 35a(2) |  |
| loadp2 | 249-251 | N | 3 | # of derailed loaded passenger cars | 35b(2) |  |
| emptyf2 | 252-254 | N | 3 | # of derailed empty freight cars | 35c(2) |  |
| emptyp2 | 255-257 | N | 3 | # of derailed empty passenger cars | 35d(2) |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| FIELD NAME | FIELD POSITION | FIELD TYPE | FIELD LENGTH | DEFINITION | BLOCK # ON FORM 6180.54 | CONVERSION |
| caboose2 | 258-260 | N | 3 | # of derailed cabooses | 35e(2) |  |
| eqpdmg | 261-267 | N | 7 | Reportable equipment damage in $ | 36 |  |
| trkdmg | 268-274 | N | 7 | Track, signal, way & structure damage in $ | 37 |  |
| cause | 275-278 | A/N | 4 | Primary cause of incident (refer to Appendix C) | 38 |  |
| cause2 | 279-282 | A/N | 4 | Contributing cause of incident (refer to Appendix C) | 39 |  |
| caskldrr | 283-285 | N | 3 | # killed for reporting RR-calculated from Form F6180.55’s submitted |  |  |
| casinjrr | 286-289 | N | 4 | # injured for reporting RR-calculated from Form F6180.55a’s submitted |  |  |
| caskld | 290-292 | N | 3 | Total killed for all RR’s involved-calculated from Form F6180.55a’s submitted |  |  |
| casinj | 293-296 | N | 4 | Total injured for all RR’s involved-calculated from Form F6180-55a’s submitted |  |  |
| accause | 297-300 | A/N | 4 | Accident cause code from jointcd 1 record for this incident (refer to Appendix C) |  |  |
| acctrk | 301 | A/N | 1 | Type track code from jointcd 1 record for this incident |  |  |
| acctrkcl | 302 | A/N | 1 | FRA track class from jointcd 1 record for this incident (FRA track class: 1-9,X) |  |  |
| highspd | 303-305 | A/N | 3 | Maximum speed reported for equipment involved: Blank=unknown |  |  |
| accdmg | 306-313 | N | 8 | Total reportable damage on all reports in $ |  |  |
| dummy2 | 314-316 | A/N | 3 | Blank data expansion field |  |  |
| stcnty | 317-322 | A/N | 6 | FIPS State & County code |  |  |
| totinj | 323-326 | N | 4 | Total injured for railroad as reported on Form F6180.54 |  |  |
| FIELD NAME | FIELD POSITION | FIELD TYPE | FIELD LENGTH | DEFINITION | BLOCK # ON FORM 6180.54 | CONVERSION |
| dummy3 | 327-332 | A/N | 6 | Blank data expansion field |  |  |
| totkld | 333-336 | N | 4 | Total killed for railroad as reported on Form F6180.54 |  |  |
| engrs | 337 | A/N | 1 | # of engineers on duty: blank=not applicable | 40 |  |
| firemen | 338 | A/N | 1 | # of firemen on duty: blank=not applicable | 41 |  |
| conductr | 339 | A/N | 1 | # of conductors on duty: blank=not applicable | 42 |  |
| brakemen | 340 | A/N | 1 | # of brakemen on duty: blank=not applicable | 43 |  |
| enghr | 341-342 | A/N | 2 | # of hours engineers on duty: blank=not applicable | 44 |  |
| engmin | 343-344 | A/N | 2 | # of minutes engineers on duty (+enghr): Blank=not applicable | 44 |  |
| cdtrhr | 345-346 | A/N | 2 | # of hours conductors on duty: blank=not applicable | 45 |  |
| cdtrmin | 347-348 | A/N | 2 | # of minutes conductors on duty (+cdtrhr): Blank=not applicable | 45 |  |
| jointcd | 349 | A/N | 1 | Indicates railroad reporting |  |  |
| region | 350 | A/N | 1 | FRA designated region |  |  |
| dummy4 | 351 | A/N | 1 | Blank data expansion field |  |  |
| typrr | 352-353 | A/N | 2 | Type railroad – ICC categories: 1^{st} position indicates class 1, 2, or 3 RR |  |  |
| dummy5 | 354-356 | A/N | 3 | Blank data expansion field |  |  |
| rrdiv | 357-362 | A/N | 6 | RR division code |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| FIELD NAME | FIELD POSITION | FIELD TYPE | FIELD LENGTH | DEFINITION | BLOCK # ON FORM 6180.54 | CONVERSION |
| method | 363-382 | A/N | 20 | Method of operation (series of 1 position codes) A=ATCS  B=Auto train control C=Auto train stop D=Cab signals E=Traffic control F=Interlocking G=Automatic block rules H=Current of traffic I=Time/table/train orders J=Track warrant control K=Direct traffic control L=Yard limits M=Special instructions N=Other than main track O=Other (specify in a Narrative) P=Positive train control | 30 | As of June 1, 2011 – this field (item box) will be replaced by the “Type of Territory” field on the Form F6180.54.  For accidents occurring June 1, 2011, or later this field will be blank. |
| narrlen | 383-386 | N | 4 | Length of narrative |  |  |
| dummy6 | 387-390 | A/N | 4 | Blank data expansion field |  |  |
| year4 | 391-394 | A/N | 4 | Four character year identification |  |  |
| rrempkld | 395-397 | N | 3 | # of RR employees killed as reported on Form F6180.54 | 46 |  |
| rrempinj | 398-400 | N | 3 | # of RR employees injured as reported on Form f6180.54 | 46 |  |
| passkld | 401-403 | N | 3 | # of passengers killed as reported on Form F6180.54 | 47 |  |
| passinj | 404-406 | N | 3 | # of passengers injured as reported on Form F6180.54 | 47 |  |
| otherkld | 407-409 | N | 3 | # of others killed as reported on Form F6180.54 | 48 |  |
| otherinj | 410-412 | N | 3 | # of others injured as reported on Form F6180.54 | 48 |  |
| county | 413-432 | A/N | 20 | County name (see FIPS codes for associated codes) | 16 |  |
| cntycd | 433-435 | A/N | 3 | FIPS county code |  |  |
| alcohol | 436-437 | A/N | 2 | # of positive alcohol tests | 32 |  |
| drug | 438-439 | A/N | 2 | # of positive drug tests | 32 |  |
| dummy7 | 440-451 | A/N | 12 | Blank data expansion field |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| FIELD NAME | FIELD POSITION | FIELD TYPE | FIELD LENGTH | DEFINITION | BLOCK # ON FORM 6180.54 | CONVERSION |
| passtrn | 452 | A/N | 1 | Were there passengers being transported: Y=Yes N=No  Blank=Not applicable | 33 |  |
| ssb1 | 453-472 | A/N | 20 | Special study block 1 Type of Track an accident/incident occurred on CWR = Continuous welded rail track OTH = Other | 49 | For accidents occurring June 1, 2011, or later. |
| ssb2 | 473-492 | A/N | 20 | Special study block 2 | 49 |  |
| narr1 | 493-592 | A/N | 100 | Narrative | 52 |  |
| narr2 | 593-692 | A/N | 100 | Narrative | 52 |  |
| narr3 | 693-792 | A/N | 100 | Narrative | 52 |  |
| narr4 | 793-892 | A/N | 100 | Narrative | 52 |  |
| narr5 | 893-992 | A/N | 100 | Narrative | 52 |  |
| narr6 | 993-1092 | A/N | 100 | Narrative | 52 |  |
| narr7 | 1093-1192 | A/N | 100 | Narrative | 52 |  |
| narr8 | 1193-1292 | A/N | 100 | Narrative | 52 |  |
| narr9 | 1293-1392 | A/N | 100 | Narrative | 52 |  |
| narr10 | 1393-1492 | A/N | 100 | Narrative | 52 |  |
| narr11 | 1493-1592 | A/N | 100 | Narrative | 52 |  |
| narr12 | 1593-1692 | A/N | 100 | Narrative | 52 |  |
| narr13 | 1693-1792 | A/N | 100 | Narrative | 52 |  |
| narr14 | 1793-1892 | A/N | 100 | Narrative | 52 |  |
| narr15 | 1893-1992 | A/N | 100 | Narrative | 52 |  |
| rcl | 1993 | A/N | 1 | Remote control locomotive = 0,1,2, or 3 0 = not a remotely controlled operation 1 = remote control portable transmitter 2 = remote control tower operation 3 = remote control portable transmitter (more than one remote control) | 30a |  |
| latitude | 1994-2003 | N | 10 | Latitude in decimal degrees, explicit decimal, explicit +/- (WGS84) | 50 |  |
| FIELD NAME | FIELD POSITION | FIELD TYPE | FIELD LENGTH | DEFINITION | BLOCK # ON FORM 6180.54 | CONVERSION |
| longitud | 2004-2014 | N | 11 | Longitude in decimal degrees, explicit decimal, explicit +/- (WGS84) | 51 |  |
| signal | 2015 | A/N | 1 | Type of territory – signalization (mandatory) 1 = signaled 2 = not signaled | 30 | New data for accidents occurring June 1, 2011, or later. |
| mopera | 2016 | A/N | 1 | Method of operation/Authority for movement (mandatory) 1 = signal indication 2 = direct train control 3 = yard/restricted limits 4 = block register territory 5 = other than main track | 30 | New data for accidents occurring June 1, 2011, or later. |
| adjunct1 | 2017 | A/N | 1 | Supplemental/adjacent codes (mandatory) (\*mandatory to the extent that all applicable codes are entered) See FRA Guide for Preparing Accident/Incident Reporting for codes | 30 | New data for accidents occurring June 1, 2011, or later. |
| adjunct2 | 2018 | A/N | 1 | Supplemental/adjacent codes (mandatory) (\*mandatory to the extent that all applicable codes are entered) See FRA Guide for Preparing Accident/Incident Reporting for codes | 30 | New data for accidents occurring June 1, 2011, or later. |
| adjunct3 | 2019 | A/N | 1 | Supplemental/adjacent codes (mandatory) (\*mandatory to the extent that all applicable codes are entered) See FRA Guide for Preparing Accident/Incident Reporting for codes | 30 | New data for accidents occurring June 1, 2011, or later. |
| subdiv | 2020 - 2039 | A/N | 20 | Railroad subdivision | 12 | New data for accidents occurring June 1, 2011, or later. |