Data 318 [Data Mining]
Dr. Gregory Tanner
Students: Drocezesky, Ivana - Ramirez, Paula – Zamora, Bryan
Final Project Report

Final Project: Predicting Casualties in Traffic Crashes of New York City

Introduction

The primary objective of this project is to leverage the principles of data mining to construct a model aimed at predicting whether a traffic incident in New York City will lead to casualties. By harnessing the knowledge and skills acquired through data mining course through the semester, we aspire to develop a model endowed with a high degree of accuracy in its predictions.

The significance of this project lies in its potential to influence road safety measures and inform more responsible driving. Accurate predictions regarding the risk and aftermath of traffic collisions are crucial for implementing targeted preventive strategies, heightening public awareness, and refining emergency response protocols.

The foundation of this project rests upon the Motor Vehicle Collisions crash table, which encapsulates details about each crash event. Sourced from police-reported motor vehicle collisions across New York City, this data is integral for generating predictions. However, it's imperative to acknowledge that the reliability of the data is contingent upon the accuracy of the police reports (MV104-AN), which are subject to potential amendments based on revised crash details.

In constructing our predictive model for traffic accidents in New York City, careful consideration was given to the selection of variables to ensure robustness and efficacy. The variables chosen to encompass a range of factors known to influence the likelihood and severity of accidents. Firstly, variations in weather conditions throughout the year can significantly impact road safety, making 'Season/Time of the Year' a crucial variable to include. Additionally, the temporal aspect of collisions is vital, with certain times of the day posing heightened risks due to factors such as impaired driving and reduced visibility, thus warranting the incorporation of 'Crash Time' into our model. 'Crash Factors' provide essential insights into the root causes of accidents, such as speeding or distracted driving, thereby informing effective preventive measures. Considering the diverse safety features and operational characteristics of different vehicle types, 'Vehicle Types' emerged as a significant variable influencing the severity of collisions.

Finally, 'Location (Latitude, Longitude)' serves as a proxy for various factors including road infrastructure, traffic density, and local driving habits, aiding in the identification of high-risk zones. By integrating these variables into our model, we aim to enhance its predictive capabilities and contribute to the advancement of road safety measures in New York City.

The developed model holds the potential to be employed as a proactive tool in forecasting the outcomes of traffic incidents. By analyzing the selected predictor variables, the model can furnish stakeholders with actionable insights to preemptively address potential hazards on the road.

Cleaning Data

In the data preprocessing phase, several steps were undertaken to refine the original dataset for analysis. Initially, unnecessary variables were removed to reduce the size of the data. Similarly, a sample of 1000 observations was extracted for detailed analysis, ensuring manageable data volume. An examination of the variables included in the original dataset was conducted, along with an assessment of their respective data types to inform subsequent preprocessing steps. Special attention was devoted to the target variable, namely the total number of injuries and fatalities resulting from traffic accidents, to gain insights into the primary focus of the analysis.

Following data selection, predictor variables were identified based on their potential predictive value. Key variables such as the number of people injured and killed, crash date, crash time, vehicle type, and contributing factors to crashes were chosen. The selection of 'Vehicle type 1' was selected over other vehicle type columns due to its high data availability across observations and its attribution as the primary cause of the crash in many instances.

In the refinement of predictor variables, efforts were made to enhance their interpretability. The total number of injuries and deaths was consolidated into a single variable, termed 'casualties,' and transformed into a boolean format to facilitate binary classification. Moreover, temporal variables such as 'CRASH.DATE' and 'CRASH.TIME' were transformed into categorical variables representing seasons and time of day, respectively. Categorization of 'VEHICLE.TYPE.CODE.1' into four distinct types aimed to simplify vehicle classification for analysis, categorizing vehicles as Passenger, Commercial, Special Purpose, or Unknown/Other. Contributing factors to crashes were grouped into five overarching categories, including reckless driving, driver behavior, environmental factors, poor driving, and other miscellaneous factors.

To ensure data integrity, observations with missing values were removed from the dataset. Finally, a training and test set split was performed, with 80% of the data allocated for training the predictive model and 20% reserved for evaluating model performance, ensuring robustness and reliability in model assessment.

Exploratory Analysis

Central to the success of our predictive model is the relationship that will be established among the predictors and the target variable, traffic incident casualties. For our choice of predictors, a multidimensional consideration of the underlying dynamics of traffic incidents in NYC is adopted. Therefore, through comprehensive data examination and analysis, we can understand the relationships among the predictors and how they might be associated with crashes and the fatalities that result from the crashes.

Plots will provide a succinct way to do this by comparing the distribution of casualties to the predictors. These plots offer visual hints to potential links and dependencies between the predictors and the target variable and are necessary for the success of our model.

For instance, figure 1 provides the distribution of casualties resulting from traffic incidents in our dataset. It reveals that from our observations, 720 instances recorded no casualties, while 219 instances resulted in fatalities. This visualization aids us in gaining insights into the distribution of outcomes between instances of casualties and those fortunate enough to avoid harm.

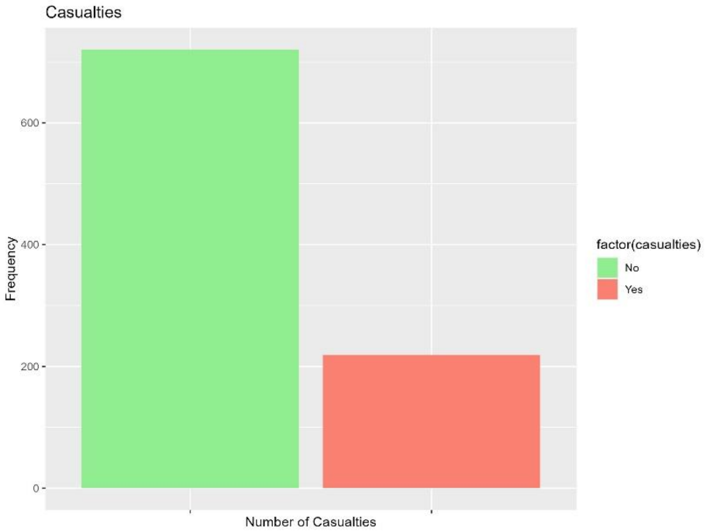

Figure 1. Number of Casualties

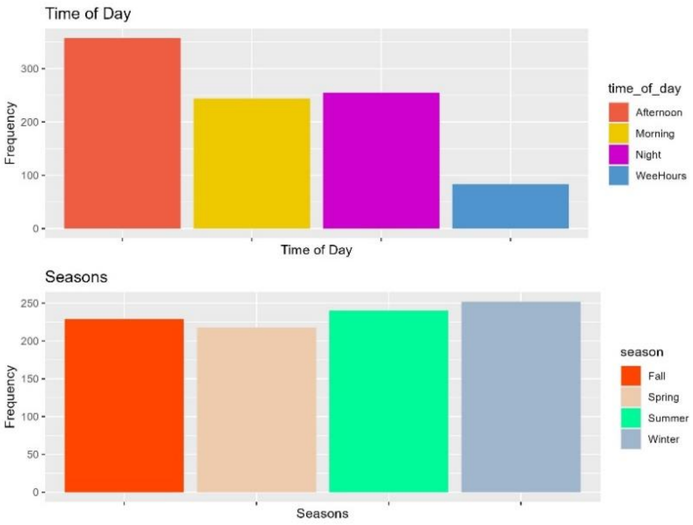

Figure 2. Crashes per Time of Day and Season

Figure 2 illustrates the temporal distribution of crashes categorized by time of day and season. Afternoon exhibits the highest frequency of crashes, with 357 observations recorded during this period. Morning and night periods demonstrate comparable crash frequencies, comprising 244 and 255 observations, respectively. Early hours of the day ("Wee Hours") exhibits a lower incidence of crashes with a number of 83. Regarding seasonal trends, the distribution of crash observations remains relatively consistent across all seasons, with winter recording the highest count at 252 crashes.

Figure 3 delves into the traffic crashes in New York by vehicles involved and contributing factors. Passenger vehicles emerge as the predominant vehicle type associated with crashes with a frequency of 479, followed by the unspecified category with a total of 397 observations. Conversely, commercial and special-purpose vehicles exhibit comparatively lower frequencies, each recording fewer than 50 observations.

From Figure 3, the contributing factors, the category labeled "other" emerges as the most prevalent contributing factor to crashes with a number of 307. Environmental factors, on the other hand, are identified as the least significant contributors to crashes. Driver’s behavior, poor and reckless driving exhibit moderate frequencies, with observations ranging from 150 to 260.

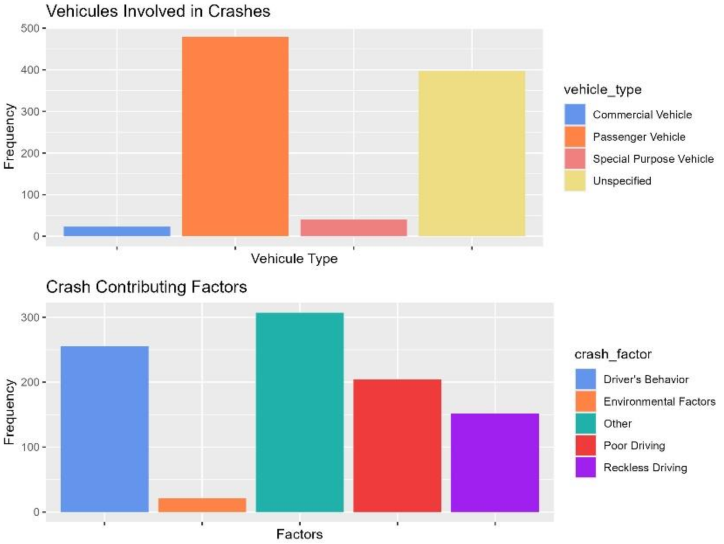

Figure 3. Crashes per Vehicle Involved and Contributing Factor

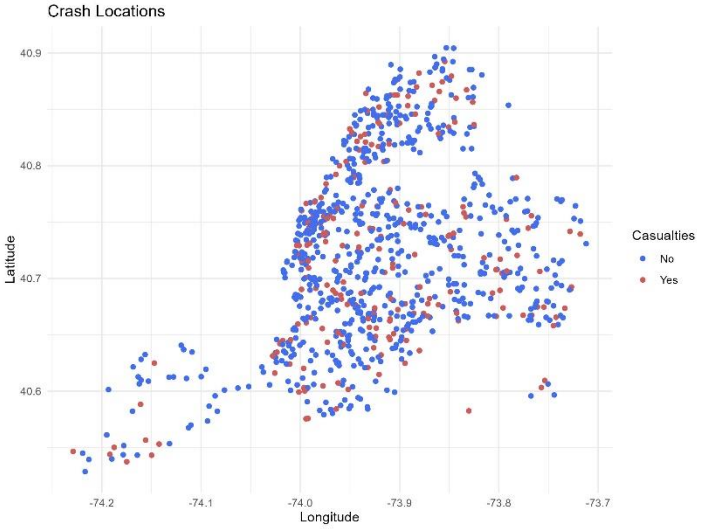

Figure 4. Crashes per Longitude and Latitude

Figure 4 shows a geospatial distribution in latitude and longitude of crashes resulting in casualties in New York City. Since the red dots are the casualties, then it is possible to see that there is no correlation and that the data points are not following any trend.

Prediction Models and their results

Model 1: Logistic Regression

Table 1. Confusion Matrix

	Reference	
Prediction	Yes	No
Yes	24	49
No	20	95

True Positives (TP): 24 | False Positives (FP): 49 |

False Negatives (FN): 20 | True Negatives (TN): 95 |

The model correctly identified 24 instances of "Yes" and 95 instances of "No", while it misclassified 49 instances of "Yes" as "No" and 20 instances of "No" as "Yes".

Table 2. Key Metrics of Model’s Performance

Accuracy	Kappa	Sensitivity	Specificity	Area Under the Curve
0.633	0.167	0.545	0.659	0.613

Using a threshold probability of 0.26 to classify observations into the "Yes" or "No" categories for the target variable "casualties" generated an accuracy of 0.633. Indicating that the model correctly predicts casualties 63.3% of the time, while a Kappa of 0.167 is quite low. Sensitivity of 54.5% and specificity of 65.9% suggest that the model has slightly better performance in identifying non-casualties. An Area Under the Curve (AUC) value of 0.613 suggests that there is some separation between the model's true positive rate and false positive rate, but the degree of separation is not very high.

Model 2: Naive Bayes

Table 2. Confusion Matrix

	Reference	
Prediction	Yes	No
Yes	0	0
No	44	144

True Positives (TP): 0 | False Positives (FP): 0 |

False Negatives (FN): 44 | True Negatives (TN): 144 |

Table 2. Key Metrics of Model’s Performance

Accuracy	Kappa	Sensitivity	Specificity	Area Under the Curve
0.766	0	0.000	1.000	0.636

A Laplace smoothing with a parameter of 0.5 was applied to handle any zero probabilities, which generated an accuracy of 0.766. Indicating that approximately 76.6% of predictions were correct, while a Kappa of 0 suggests no agreement beyond chance. Sensitivity of 0% and specificity of 1% suggest that the model has better performance in identifying non-casualties while underperforming in predicting casualties. An Area Under the Curve (AUC) value of 0.636 suggests that the model's ability to discriminate between positive and negative instances is slightly better than random chance.

Model 3: K Nearest Neighbor

Table 2. Confusion Matrix

	Reference	
Prediction	Yes	No
Yes	13	58
No	31	86

True Positives (TP): 13 | False Positives (FP): 58 |

False Negatives (FN): 31 | True Negatives (TN): 86 |

Table 3. Key Metrics of Model’s Performance

Accuracy	Kappa	Sensitivity	Specificity	Area Under the Curve
0.526	-0.088	0.295	0.597	0.567

Using a threshold probability of 0.26 to classify observations into the "Yes" or "No" categories for the target variable "casualties" generated an accuracy of 0.526. Indicating that approximately 52.6% of predictions were correct, while a negative Kappa of -0.088 suggests poor agreement between observed and predicted classifications beyond chance. Sensitivity of 29.5% and specificity of 59.7% reveal that the classifier struggles to correctly identify both positive and negative instances. An Area Under the Curve (AUC) value of 0.567 suggests that the model's ability to discriminate between positive and negative instances is slightly better than random chance.

Comparing the Models

Comparing the three predictive models developed for forecasting casualties in traffic crashes in New York City reveals varying levels of performance and effectiveness. The logistic regression model achieved an accuracy of 63.3%, with a sensitivity of 54.5% and a specificity of 65.9%. While its accuracy is above random guessing, its performance in correctly identifying casualties is modest, as indicated by the low sensitivity. The Naive Bayes model outperformed the logistic regression with an accuracy of 76.6%, yet it exhibited a sensitivity of 0% and a specificity of 100%. While excelling in predicting non-casualty instances, it failed to identify any true positive instances of casualties, revealing significant limitations in its predictive power. The k-nearest neighbor (KNN) model, on the other hand, displayed the lowest accuracy of 52.6%, with a sensitivity of 29.5% and a specificity of 59.7%. Despite its better balance between sensitivity and specificity compared to Naive Bayes, its overall performance remains suboptimal, suggesting challenges in correctly classifying both positive and negative instances.

The results and performance of the predictive models underscore several shortcomings and areas for improvement. Firstly, all models exhibit limitations in accurately predicting casualties, with varying degrees of sensitivity and specificity deficiencies. While logistic regression and KNN models struggle with balancing sensitivity and specificity, Naive Bayes excels in one aspect while severely underperforming in the other. This highlights the need for more nuanced modeling approaches capable of capturing the complexity of factors influencing traffic crash outcomes. Additionally, the relatively low values of Kappa across all models suggest poor agreement between observed and predicted classifications beyond chance, indicating room for refinement in model calibration and feature selection. Furthermore, the moderate AUC values indicate that while the models possess some discriminative ability, there is still significant room for improvement in distinguishing between positive and negative instances more effectively.

Moving forward, several strategies can be employed to enhance the predictive accuracy and robustness of the models. Firstly, incorporating additional relevant variables, such as driver demographics, road conditions, and vehicle characteristics, could enrich the predictive models' feature space and capture more nuanced relationships. Furthermore, exploring more advanced modeling techniques, including ensemble methods and deep learning algorithms, may help uncover intricate patterns within the data and improve predictive performance. Moreover, conducting rigorous model validation and refinement processes using real-world data can ensure the models' adaptability and generalizability across diverse traffic scenarios. Additionally, ongoing monitoring and updating of the models based on evolving traffic patterns and regulatory changes can ensure their relevance and effectiveness in enhancing road safety measures in New York City.

Variables analysis

The analysis conducted in the report and Rmarkdown provides valuable insights into the variables utilized for predicting casualties in traffic crashes. The variables considered include factors such as time of day, season, vehicle type, location, and crash factors. These variables offer significant predictive value, as they capture various dimensions of traffic incidents that can influence the likelihood of casualties. However, it's essential to acknowledge that these variables may not fully encapsulate all relevant factors contributing to traffic crashes.

For example, while the model incorporates the time of day, it may not consider specific environmental conditions or road infrastructure elements that could impact crash severity. Similarly, although vehicle type is accounted for, nuances such as vehicle condition or driver behavior may not be fully captured, potentially limiting the model's predictive accuracy.

Models’ shortcomings

Despite the comprehensive consideration of variables, the models showcased in the report exhibit notable shortcomings. The Logistic Regression model, while demonstrating moderate predictive ability, suffers from relatively low accuracy and AUC values. Moreover, the Naive Bayes model's inability to correctly identify any true positive instances of casualties highlights significant limitations in its sensitivity.

These shortcomings suggest that the models may not adequately capture the complexity of factors influencing traffic crash outcomes, leading to suboptimal performance. Additionally, the k-NN classification model, while offering potential insights, may also face challenges such as overfitting or sensitivity to the choice of k value, which could affect its generalizability and predictive robustness.

Improvements to be made

To improve the project's effectiveness and predictive accuracy, several future enhancements can be considered. Firstly, the incorporation of additional relevant variables, such as weather conditions, road maintenance, or driver demographics, could enrich the models' predictive capabilities. Additionally, leveraging more advanced modeling techniques, such as ensemble methods or deep learning algorithms, may help capture intricate patterns and nonlinear relationships within the data.

Furthermore, conducting thorough feature engineering and selection processes to identify the most informative predictors and mitigate noise could enhance model performance. Lastly, ongoing validation and refinement of the models using real-world data could ensure their adaptability to evolving traffic conditions and improve their reliability in predicting casualties in traffic crashes.

Conclusion

In conclusion, our project aimed to predict casualties in New York City traffic crashes using data mining techniques. While our models showed promise, they also revealed limitations in accurately predicting outcomes. We identified variables like weather, time of day, and vehicle types as significant predictors but recognized the need for more sophisticated modeling approaches. Moving forward, we suggest incorporating additional variables, exploring advanced modeling techniques, and conducting thorough validation processes to enhance predictive accuracy. Continuous monitoring and updates based on real-world data will ensure the models' effectiveness in improving road safety measures in New York City. Despite challenges, our project offers valuable insights and lays the groundwork for future enhancements in traffic crash prediction.

Data 318 [Data Mining]
Dr. Gregory Tanner
Students: Drocezesky, Ivana - Ramirez, Paula – Zamora, Bryan
Final Project Report

List of References

"Motor Vehicle Collisions - Crashes." Data.gov, https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes. Accessed: May 1, 2024.