Analysis of Traffic Incidents in Boston

Srijith Gomattam, Rohin Patel, Shrey Patel, Parth Shah
(DS Project Group 50)

Abstract

As students in Boston, we're investigating traffic crash patterns to improve safety, focusing on intersections where most accidents occur. Our city's unique layout significantly influences these issues. Our main goal is to analyze crash data to understand and prevent accidents. We're creating a predictive model to determine the mode of transportation in crashes and identify the most influential factors. This information will help us issue targeted warnings, such as pinpointing dangerous streets for pedestrians or times with higher accident risks. Our effort aims to make Boston safer for new residents, especially college students, by raising awareness and guiding safer navigation throughout the city.

Introduction

Defining the Problem:

Although they are often seen as unavoidable occurrences, traffic crashes are not merely random events. They follow discernible patterns which are influenced by infrastructure, driver behavior, and a variety of other factors. As students in the northeastern region, particularly in the bustling city of Boston, we recognize the critical importance of understanding these patterns because they pertain to the safety of our peers and the public in general. Our analysis attempted to delve deeper into the crash data from Boston to gain insights and potentially help inform preventive actions. Like any city, intersections in Boston are hotspots for traffic-related accidents because of many factors. These include driver behavior, unclear signage, pedestrian crossings, and more. Over half of the combined total of fatal and injury crashes occur at or near intersections (Intersection Safety). Boston is one of the oldest cities in the United States and has a unique road layout that isn't always conducive to the modern traffic patterns and the dense population. While the city is walkable, it is also a major traffic hub. Congested streets create a more frenzied atmosphere, increasing the likelihood of accidents involving various modes of transportation. The mode of transport involved in crashes varies, with motor vehicles, pedestrians, and bikes being the most common types. Traffic crashes can result in not only property damage but also serious injuries or fatalities. It is essential to understand where, when, and why these crashes occur to develop strategies for prevention. In 2015, out of the 200 largest US cities, Boston was last in having the safest drivers (Hofherr, J.). This ranking emphasizes that traffic accidents are a very significant local problem.

Motivation:

For prospective Northeastern students, it may be intimidating for many to come from their small suburban town to the bustling city of Boston. This was the case for many of our group members, as none of us were accustomed to the chaos of rush hour traffic. While it is exciting moving into the big city, we were worried about the potential risks that came with living in a metropolitan area. More specifically, vehicular collisions. Some streets in Boston are busier than others, and by identifying this, we can help Boston newcomers proceed to some areas with an air of caution. We hope to especially help college students like ourselves navigate Boston in a safer and more informed manner. Through this analysis, we hoped to create a sort of “PSA” so that we can make an impact and perhaps prevent an accident or two which would have otherwise happened.

Objectives:

Our overarching objective for this project is to find patterns in the data about traffic-related incidents and draw conclusions from the data to increase awareness about accident prevention in the city.

We aimed to develop a predictive model to ascertain the mode of transportation involved in road accidents by analyzing various contributing factors. Additionally, we identify and quantify the factors that exert the most influence on each type of crash. The ultimate goal is to leverage these insights to enhance public safety measures by issuing targeted warnings. For instance, if the analysis reveals that specific streets pose a higher risk for pedestrians, the findings can be utilized to alert individuals to exercise increased caution in those areas. Conversely, if time emerges as a more significant factor, tailored warnings can be disseminated accordingly to improve overall road safety. Our work so far helps us take significant steps toward achieving this goal.

Related Work

Our project is similar to some other projects that delve into traffic-related accidents using data analysis. The first one that we found was a project by Northeastern students, available on a GitHub repository by user shriramkarthikeyan which offers a detailed analysis of crash accidents. Their approach segmented data based on various criteria including driving under the influence of alcohol and drugs, seat belt usage by the driver, and traffic device malfunctions, among others. This project's depth and criteria selection provide a useful reference point for our analysis, especially in understanding how different variables can influence accident rates and severity. Similarly, another GitHub project by user ray310 examined motor vehicle collisions in New York City. This group's analysis focused on assessing the severity of accidents, identifying their timing and locations, pinpointing areas with severe incidents, and proposing potential causes and solutions. Additionally, they highlighted less evident high-risk zones, a crucial aspect for comprehensive understanding. Although their study is based on NYC data, the methodologies and insights offer a valuable comparative perspective for our Boston-focused analysis. By examining these related works, we refined our approach, learned from their methodologies, and possibly uncovered new aspects to explore in our own analysis of Boston's traffic-related accidents.

References

Hofherr, J. (2015, September 3). Bostonians crash more than twice as often as the average driver. Boston.Com. https://www.boston.com/cars/news-and-reviews/2015/09/03/bostonians-crash-more-than-twice-as-often-as-the-average-driver/

Intersection safety. (n.d.). FHWA. Retrieved October 15, 2023, from https://highways.dot.gov/research/research-programs/safety/intersection-safety

ray310(2022)NYC-Vehicle-Collisions[Source code].https://github.com/ray310/NYC-Vehicle-Collisions

shriramkarthikeyan(2019)Motor-Vehicle-Crash[Source code].https://github.com/shriramkarthikeyan/Motor-Vehicle-Crash

Methodology

1. Data Acquisition:

The primary dataset for our analysis is the "Vision Zero Crash Records" dataset from the City of Boston's public data portal. This comprehensive dataset includes details of traffic-related incidents in Boston, offering rich information on various aspects like the time of the incident, location, type of transportation involved, and more. The dataset is publicly accessible at Boston's data portal, ensuring transparency and reproducibility of our analysis.

2. Data Preparation:

First we had to clean the data by handling missing values, erroneous entries, and inconsistencies. For instance, missing values in street-related columns are filled with 'Unknown', and the 'dispatch_ts' column is converted to datetime format for ease of analysis. We also had to implement the creation of new variables that could be insightful for the analysis.

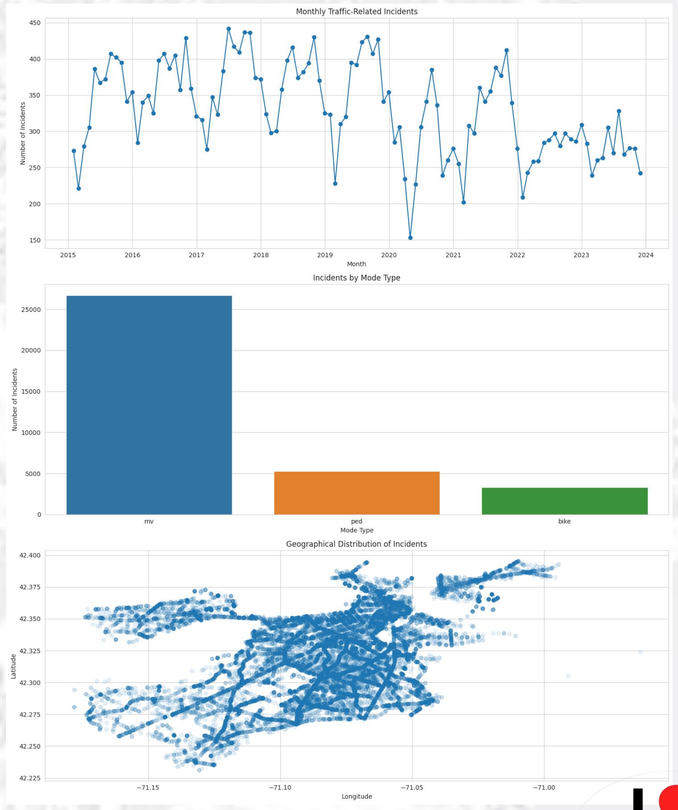
For example, we proposed to create a 'num_incidents' variable during the Exploratory Data Analysis phase. This variable will aggregate incidents to analyze patterns across different dimensions. An EDA is conducted to understand the dataset's structure and uncover initial patterns. This includes generating visualizations like line charts for monthly incident trends, bar charts for incidents by mode type, and scatter plots for geographical distribution. EDA helps in formulating hypotheses and guiding subsequent analysis. To finish cleaning up the data, we had to normalize data. This involved handling categorical variables, like converting them into dummy variables, and preparing the dataset for machine learning models.

3. Model Selection:

For predicting the mode of transportation involved in road accidents, we have selected the Random Forest Tree Classifier algorithm as a an ensemble learning method for classification due to its efficiency in handling complex datasets. The decision is based on the fact that the algorithm shows high accuracy given a well-defined feature, as the Random Forest algorithm is great at handling various types of data, managing unbalanced datasets, and avoiding miscues from overfitting. The data was first cut down into a smaller dataset to improve computational efficiency. This was done using the train_test_split to keep stratification. This smaller dataset was then partitioned into a reproducible training set and testing set, where training is 80% of the data and testing is 20% of the data.

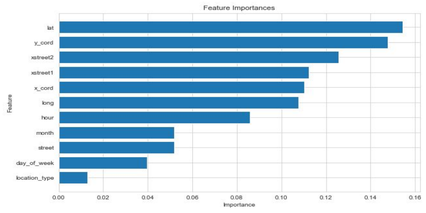
We integrated Randomized Search CV for hyperparameter tuning, focusing on parameters such as 'n_estimators' (number of trees), 'max_depth' (tree depth), and 'min_samples_split' (minimum samples for node splitting). This approach, preferred over Grid Search, efficiently samples from a range of parameter settings, reducing computational demand, particularly beneficial given the large size of our dataset. Randomized Search CV's ability to specify distributions for continuous hyperparameters adds a layer of flexibility and control to our model tuning. This was especially useful in a very large dataset like the Vision Zero Crash Records" dataset.

The model's performance is evaluated using cross-validation and metrics such as accuracy, precision, recall, and the F1-score. This comprehensive evaluation not only benchmarks our model against other alternatives but also verifies its generalizability beyond our specific case study. Additionally, a sensitivity analysis is conducted post-tuning to understand the impact of varying hyperparameters on model performance. This ensures that the model is not only optimized for our current dataset but also adaptable to similar datasets in different urban contexts.

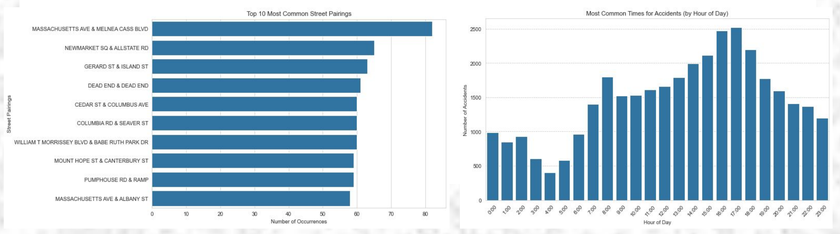


Results and Evaluation

Our analysis of traffic-related accidents in Boston, employing Random Forest, GBM, and SVC models, demonstrated promising results with each model achieving approximately 75% accuracy. However, the GBM model was not able to predict pedestrian and bike accidents as well as the other two, so it was immediately eliminated from contention for best model. After hyperparameter tuning, the random forest model showed a higher accuracy so, it was chosen as the model for this project. After conducting feature importance on the RF model, we found that location (features such as lat, long, x_coord, y_coord, and the streets) contributed the most to the accident designation. After location comes hour in terms of importance, followed by the month and then the day of the week. Despite the general success of our models, they exhibited limitations in accurately predicting incidents involving bikes and pedestrians, due to the challenges posed by unbalanced data. This underscores the need for more refined modeling approaches and enhanced data collection methods to improve predictions in these areas. Overall, given the size of the dataset and the number of factors at play, we believe that our model’s accuracy is formidable and can be used in future analyses.



After conducting further analysis, we identified that certain intersections were particularly prone to accidents, especially those involving pedestrians and cyclists. Some of these busiest intersections include Mass Ave and Melna Cass Blvd, Newmarket Sq and Allstate Rd, and Gerard St and Island St. This information is crucial for targeted interventions in these high-risk areas, aiding city planning and public safety strategies. In our analysis of time, we found that the most accidents happen between 4:00 and 5:00 pm, which align closely with Boston's traffic-hours. Using this data, we would advise Boston newcomers to try to travel less on the busier roads and plan their travelling earlier in the day or late at night.



Impact

Our solution can benefit multiple stakeholders and mitigate negative effects on various groups. Some beneficiaries include residents, students, and visitors in Boston. By identifying high-risk areas and understanding the factors contributing to accidents, our analysis can lead to the implementation of targeted safety measures. This can make the bustling city safer for everyone. It can especially help newcomers and students navigate the city more safely and increase overall awareness. City planners and traffic management authorities can potentially be informed about targeted preventive strategies which could increase awareness campaigns or adjustments to traffic regulations. A safer urban environment can positively influence the perception of Boston as a whole. Overall, we hope to make a positive impact through better traffic flow and reduced accident rates.

Conclusion

Our project, centered on analyzing traffic crash patterns in Boston, particularly at intersections, has achieved several key milestones towards enhancing urban safety. By leveraging the comprehensive "Vision Zero Crash Records" dataset, we have successfully identified and analyzed patterns in traffic-related incidents. Our methodology, encompassing rigorous data preparation and the implementation of a Random Forest Tree Classifier algorithm, was further strengthened by the incorporation of Randomized Search CV for hyperparameter tuning. This approach allowed us to navigate the complexities of a large dataset efficiently. Our models, including the Random Forest, GBM, and Non-Linear SVC, demonstrated good performance.

Our findings are particularly beneficial for new residents and students in Boston, assisting them in safer navigation through the city. If we were to continue, we aim to further refine our model to improve its accuracy, especially for incidents involving bicycles and pedestrians. This may involve integrating more detailed traffic and pedestrian flow data, as well as continuous model updates to adapt to evolving urban dynamics. Our efforts contribute to the broader goal of creating safer, more navigable urban environments, highlighting the potential of data-driven insights in fostering safer city living conditions.