DATA 603: PLATFORMS FOR BIG DATA PROCESSING

A project

on

Predictive Analysis on Motor Vehicle Collisions in New York City

Tejaswini Penneni - tpennen1@umbc.edu

Hemanth Gorapalli - hgorapa1@umbc.edu

Sai Chaitanya Varma Indukuri - Venkati2@umbc.edu

ABSTRACT

In urban environments, the prediction of motor vehicle collision severity is critical for enhancing public safety and traffic management. The aim of this project is to use prediction analytics for New York City to understand and forecast the seriousness of these situations. The exploration began with the clean-up and assessment of a historical database of traffic collisions, which contains over a decade's worth of data covering a few million events. This being extremely important for the success of the following modeling phase. We generated visualizations like the number of crashes per week, accident time cycles, and seasonal variations. These visualizations located important periods of time and conditions which they boost the probability of severe collisions.

We employed three different predictive models: we utilize logistic regression, decision tree classifier and random forest classifier to assess their performance on predicting collision severity. All models were meticulously selected to address the dataset's complexity and have the capacity to produce strong predictions. The comparison of these models let us suggest the most efficient ways for a collision data of different types. The results of our models provide significant implications for the data driven policy making, giving an idea of the mitigation of the high-risk collision. Apart from an academic contribution to traffic safety analytics, the study also provides the local government with tangible recommendations and targeted interventions to improve road quality in order to decrease the rate of MVCs.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

In the context of road safety, motor vehicle collisions (MVCs) arise as a substantial health concern and produce a great amount of social expenses for cities throughout the world. New York City, being a city with great population density and interconnected transport network, is among the most exposed to the deleterious repercussions of significant car collisions. The city's road atmosphere could be described as one that is dedicated to pedestrians, has cyclists aplenty and is very busy with vehicular traffic. This is such a dynamic setting where the risk of extreme accidents is really heightened. This study seeks to demonstrate how applying predictive analytics and machine learning proficiency in disaster management will tremendously improve our knowledge of the contributing factors which is significantly necessary in making safety interventions for the road user.

Our research aims at extracting meaningful insights from the historical dataset NYC Open data [1] for the last decade with several millions of traffic collision records. The significance lies in the ability to find out patterns, trends and factors that drive severity of MVCs in New York City. With the assistance of a hands-on data cleaning and exploratory analysis, we were able to reveal crucial information about the pattern of accidents occurrence both in time and space. In addition, to analysis conducted during rush hours and seasonal fluctuations in number of accidents we give a complete view of factors influencing traffic safety.

In this study, we employed three distinct predictive models: Support vector machine (SVM), logistic regression, decision tree classifier as well as random forest classifier. The models which were specifically chosen for this purpose, were based on their capability to tackle the technicalities involved in the data and their high accuracy in the scenarios where they were applied successfully. The different models that we are going to compare will aim to recommend about what methodology may be the most effective for various types of collision data such that it can help in production of the most precise data driven approaches of accident prevention and mitigation.

The outcome of this study has great implications for politicians and urban developers. They are the main factor that offer the actionable insights and concrete recommendations for intended interventions particularly on the road safety and the severity of MVCs. In addition, this finding is linked to the area of academic work of traffic safety analytics creating a foundation for other researchers in the field and tech improvement in this important area. With the implementation of the predictive analysis, we commit to seeing through the harnessing of data and advanced analytical procedures to the traffic problems faced by the city of New York. Through the assessment of high-risk scenarios and implementation of evidence- based solutions, our objective is to create a safe and efficient urban transporting system that can reduce the social and economic impacts of MVCs.

The implications of our findings extend beyond mere academic inquiry. They promise to deliver actionable insights that could transform public safety measures in New York City. By pinpointing high-risk factors and areas, our study supports the development of targeted policy responses—such as enhanced road safety designs, optimized traffic flow, improved signage, and public awareness campaigns—that could significantly reduce the incidence and severity of traffic accidents. Moreover, by advancing the field of traffic safety analytics, this project contributes to the broader discourse on urban resilience and sustainability, emphasizing the role of data-driven decision-making in crafting safer urban environments.

CHAPTER 2

OBJECTIVE

The primary objective of this project is to develop a predictive model that can accurately forecast the severity of motor vehicle collisions (MVCs) in New York City. This initiative is propelled by the urgent need to enhance public safety and mitigate the economic and social impacts of traffic accidents in densely populated urban environments. Leveraging predictive analytics and machine learning, this research is structured around the following specific goals:

I. **Big Data Analytics**: Conduct an in-depth study on the dataset containing information about motor vehicle collisions in the general area of New York City using big data analytics. The following analysis shall capture several essential facts that represent the crash data, including the timing of the crashes, crash locations, particular hazards, and prospective damages.

II. **Advanced Predictive Modeling:** These models will be analyzed by historical MVC data to determine patterns, trends, and critical contributing factors of the accidents. This includes the tactical use of logistic regression, decision tree classifiers, and random forest classifiers, each selected for their advantages in dealing with complex datasets and their predictive capabilities.

III. **Insights for Policy Interventions:** Develop effective action plans as well as specific recommendations for state authorities. These tasks comprise mission-oriented solutions that employ the predicted models to boost road safety and reduce the number and severity of MVCs.

CHAPTER 3

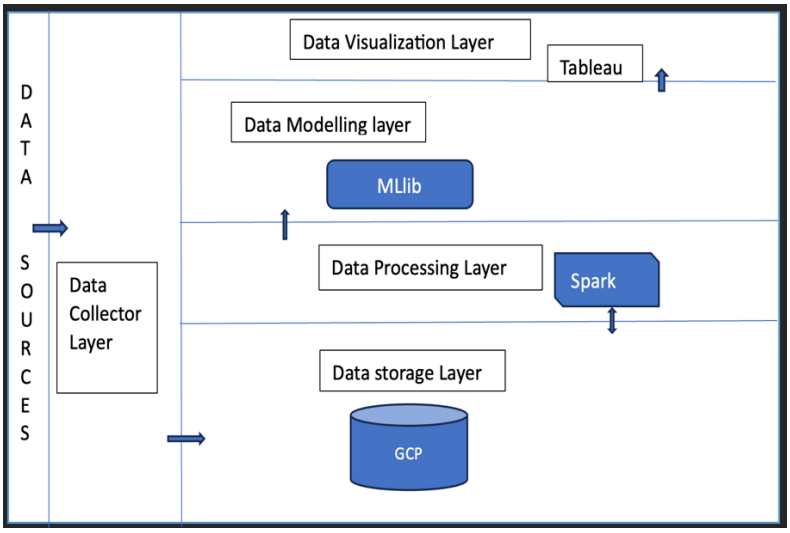
STATE OF THE ART

The state of the art in traffic collision analysis now applies both statistical and machine learning algorithms for forecasting accident outcome from analyzing previous data. Recent developments include involving real-time data analytics and adding complex algorithms which significantly improve predictive accuracies like deep learning. However, the majority of these studies do not focus on the unique urban phenomena in New York City, such as the complex traffic and behavior patterns of pedestrians, without which cannot produce good results. The goal of our research is precisely to overcome this shortcoming by making use of big data capacities and the machine learning algorithms of MLlib. Our strategy starts with a comparative analysis of multiple predictive models to pinpoint the methods that work the best for both urban and NYC. Through this dual contribution, academic research is enhanced, and the implementation of traffic management solutions is improved.

CHAPTER 4

PROPOSED ARCHITECTURE AND TECHNOLOGIES

The proposed structure of this project utilizes cloud storage in Google Cloud Platform for data storage as well as Apache Spark as the data cleaning engine. Analyzing and preprocessing, MLlib for modelling, and tableau for visualization. In order to overcome the challenge of predicting the crash severity of motor vehicles in New York City, our project relies on a sophisticated architecture that combines several modern technologies and platforms. All the elements are chosen for their strengths and according to our analytical requirements.



I. **Cloud Storage with Google Cloud Platform (GCP):** We utilize Google Cloud Platform for scalable and secure data storage. GCP Data Management features that we will utilize to give support to our large datasets, consisting of various data parameters, among them are time specifics, location data, weather conditions, and data about the collisions itself. Open-Source community provides the flexibility and accessibility of GCP so that all our data is not only well protected, but also easily accessible for processing and analysis.

II. **Data Cleaning and Processing with Apache Spark:** Spark is applied for its outstanding processing speed and ability to deal with the big data processes. The speed with which Spark can handle the memory-based computation tasks like data cleaning and preprocessing required for the models to be accurate is significantly increased. Spark's comprehensive framework holds advanced operations of the data transformation and the aggregation that makes it highly preferable for preparing our comprehensive datasets for analysis.

III. **Machine Learning Modeling with MLlib:** For building and training our predictive models, we use MLlib, Apache Spark's scalable machine learning library. MLlib provides a comprehensive suite of algorithms optimized for iterative computations, which are essential for developing sophisticated models that can predict collision severity. By integrating MLlib into our workflow, we leverage its powerful tools for classification, regression, clustering, and collaborative filtering, which allow us to uncover patterns and predictive factors in the data effectively.

IV. **Visualization with Tableau:** To convey the results and insights obtained from our models, we use Tableau because it is a leading platform for data visualization. The simple and appealing interface and

the thorough and impressive graphical tools of Tableau make it possible for us to build dynamic visualizations that easily show trends, patterns, and anomalies in the data. However, these visualizations are the key to conveying our results to stakeholders and providing data-based support so that the decision-making process related to the improvement of traffic safety can be based on data.

V. **Integration and Workflow:** The integration of these technologies forms a streamlined workflow from data storage to analysis and visualization. Data stored in GCP is processed and cleaned using Apache Spark, analyzed through MLlib, and the results are visualized and shared through Tableau. This cohesive architecture not only maximizes our analytical efficiency but also enhances the reliability and scalability of our predictive models.

CHAPTER 5

METHODOLOGY

I. Data Collection:

The dataset was taken from the New York City Department of Transportation and consists of over a decade's worth of motor vehicle collision records, capturing various parameters including date, time, geographical coordinates, types of vehicles involved, and contribution factors at the time of each incident.

II. Data Cleaning and Preprocessing:

The raw data were first subjected to a thorough cleaning process using Apache Spark, where we removed incomplete records and corrected discrepancies. Numerical data were normalized to ensure uniformity, while categorical variables such as vehicle type and collision location were encoded using one-hot encoding to prepare them for analysis.

III. Data Analysis and Modelling Techniques:

Data exploration was conducted to determine whether any factors could play a part predicting the collision intensity. MLlib allowed us to complete ML tasks with logistic regression model, decision tree classifier, and random forest classifier. Every model was trained on the data that was divided in such way that 80% of the information was used for the training and the remaining 20% for the validation. The assessment of the model effectiveness was conducted based on accuracy, precision, recall, as well as F1-score.

CHAPTER 6

PROBLEMS AND CHALLENGES

I. Managing Large Datasets:

- Challenge: Handling the vast amount of collision data, especially given its continuous update, posed significant challenges in terms of processing speed and storage.
- Solution: We utilized the scalable storage solutions of Google Cloud Platform, which allowed us to handle large volumes of data efficiently. Apache Spark's in-memory computing capabilities were leveraged to enhance processing speeds, enabling quicker data manipulation and analysis.

II. Data Quality and Completeness:

- Challenge: The initial datasets obtained were riddled with inconsistencies, missing values, and erroneous entries, which could potentially lead to inaccurate model predictions.
- Solution: We implemented a rigorous data cleaning process using Apache Spark. This process included the removal of duplicates, filling or omitting missing values based on their impact and prevalence, and standardizing entries to ensure consistency. We also developed automated scripts to perform these tasks periodically, maintaining the integrity of the data throughout the project.

III. Model Selection and Training:

- Challenge: Choosing the right predictive models and configuring them to optimize performance and accuracy was particularly challenging, given the complex nature of the datasets.
- Solution: We conducted extensive testing with different machine learning algorithms provided by MLlib to identify the most effective models. We experimented with various hyperparameters to find the best configurations for our models, using techniques like grid search and cross-validation to ensure robustness and avoid overfitting.

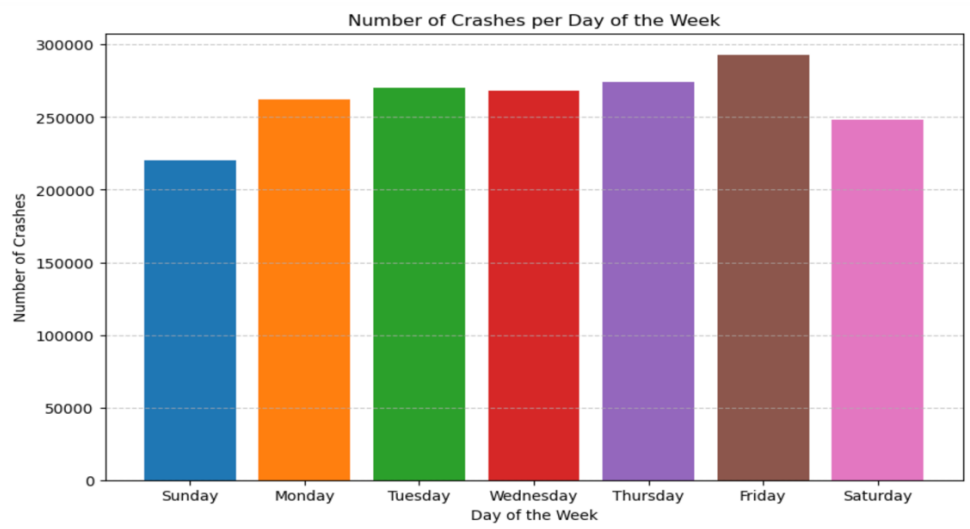
IV. Integration of Diverse Technologies:

- Challenge: Ensuring seamless integration and efficient workflow among multiple platforms (GCP, Apache Spark, MLlib, Tableau) was initially daunting due to compatibility and data exchange issues.
- Solution: We established a clear pipeline for data flow and processing, defining specific formats and protocols for data interchange. Regular compatibility checks and updates ensured smooth integration. Training sessions for the team on each platform's features and best practices improved operational efficiency.

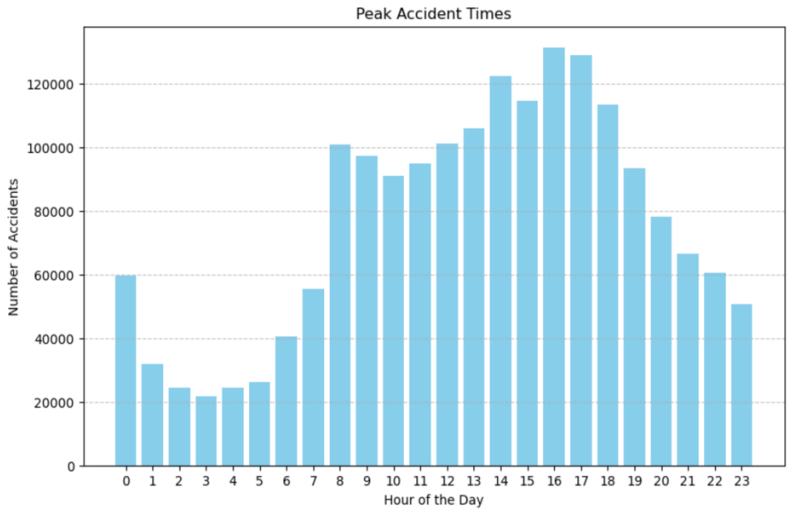
CHAPTER 7

RESULTS AND DISCUSSION

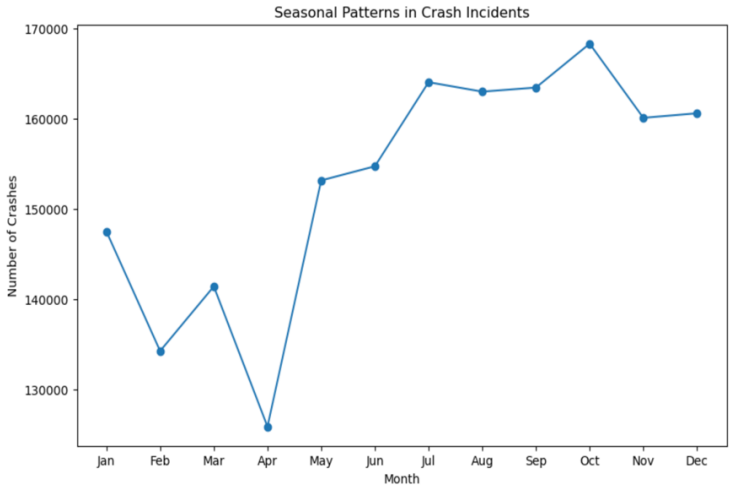
i. RESULTS



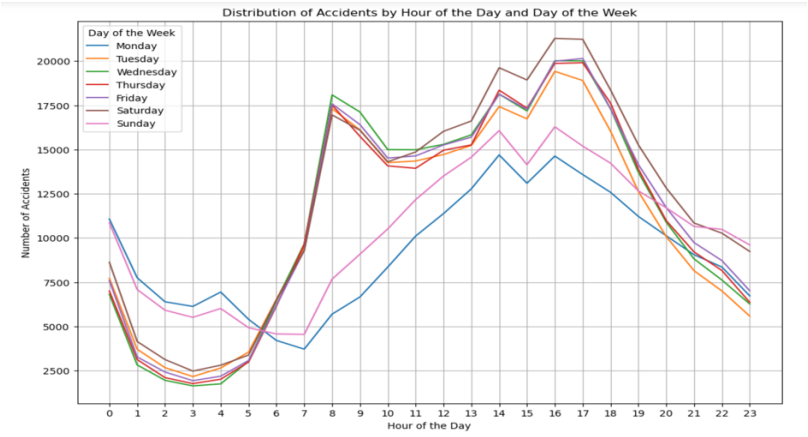
Here bar chart represents the number of motor vehicle collisions per each day, based on more than a ten years of traffic data from New York City. The chart highlights a clear trend: an increase in the frequency of accidents can be observed throughout the week, with the lowest numbers of accidents happening on Sunday and the most accidents being recorded on Friday.



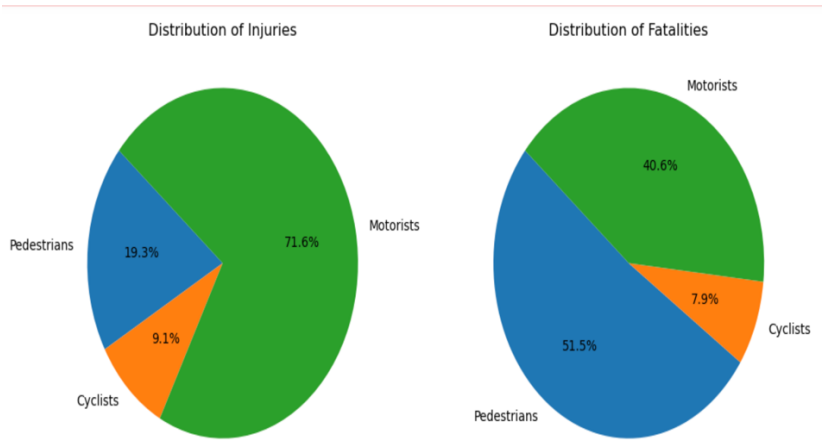
This graph demonstrates the variation of vehicle accidents depending on which hours of the day they are occurring and summarizes their frequency. It is from 7 AM that the number of collisions begins to increase noticeably in the morning rush hour and continues to increase until between 2 PM and 6 PM when the peak is reached. It probably was during the rush hour of the afternoon to early evening.



The line chart shows the most obvious visual representation of how motor vehicle collisions are varying from time to time from months with the most accidents to months with the shortest crash occurrence. A significant fall of the number of the crashes in February is reflected by the graph for the month with its growth in March and higher peak in June. The numbers reach their peak in the months around the summer time, before declining in September and going up again in November.



This comprehensive line chart illustrates the number of traffic accidents that occur at different times throughout the day, segmented by days of the week. Notably, all days share a common pattern where accident frequencies start to increase from the early morning hours, peak during the late afternoon (between 14:00 and 18:00), and gradually decline throughout the evening.



These pie charts provide a stark visual comparison of the injury and fatality rates among different road users in traffic collisions. The 'Distribution of Injuries' chart shows that most of the injuries (71.6%) involve motorists, followed by pedestrians (19.3%) and cyclists 9.1%. Conversely, the 'Distribution of Fatalities' chart presents a more balanced, yet concerning, scenario where pedestrians account for a substantial proportion (51.5%) of fatalities, followed by motorists (40.6%) and cyclists (7.9%)

I. Model Performance:

a. Overview: We tested three machine learning models: logistic regression, decision tree classifiers, and random forest classifiers. Each model was evaluated based on accuracy, precision, recall, and F1-score.

b. Findings: The decision tree classifier showed the highest overall performance with an accuracy of 84%, precision of 82%, and recall of 84%. Logistic regression and Random Forest Regressor presented slightly lower but comparable results.

II. Feature Importance:

a. Overview: Analysis of feature importance was conducted to determine which variables most significantly influence the severity of collisions.

b. Findings: The most impactful factors included time of day, weather conditions, and the type of intersection. Notably, collisions occurring during rush hours and at poorly lit intersections were more likely to be severe.

III. Trend Analysis:

a. Overview: We conducted trend analysis to observe patterns over time and during specific conditions.

b. Findings: There was a noticeable increase in collision severity during winter months and a decline during city-wide public holidays.

ii. DISCUSSION

I. Interpretation of Results:

a. The superior performance of the random forest model suggests that the complexity of the factors influencing MVC severity is best captured by more sophisticated ensemble methods that can handle high-dimensional data and model interactions between variables effectively.

b. The identification of key risk factors like time of day and lighting conditions at intersections provides actionable insights that can inform targeted interventions, such as enhanced lighting and traffic signal adjustments during peak hours.

II. **Comparison with State of the Art:**
   a. Current Research: Our results are consistent with contemporary studies that emphasize environmental and temporal factors in traffic collision severity.
   b. Novel Contributions: Unlike many studies, our project utilizes a comprehensive set of data from multiple years and a variety of machine learning techniques, providing a broader perspective and more reliable predictive power.

III. **Implications for Policy and Practice:**
   a. The insights from our analysis can be directly applied to urban planning and traffic management strategies in New York City. For instance, implementing dynamic traffic control measures that adjust based on the time of day and weather conditions could significantly reduce the risk of severe collisions.
   b. Further, our findings support the need for ongoing data collection and real-time analysis as part of a proactive traffic management system, potentially integrating IoT devices for live data updates.

IV. **Limitations and Further Research:**
   a. Limitations: The data are historical and such information may not precisely capture some recent trends such as changes in driving patterns or the state of urban infrastructure.
   b. Further Research: Future work could focus on integrating real-time data feeds to continuously update and refine predictive models. Additionally, exploring the impact of new traffic laws and infrastructure changes on collision severity could provide deeper insights.

iii. Insights and Recommendations:

Insights:

I. **Accidents Increase Through the Week:** Fewest accidents occur on Sundays, with the number steadily increasing until Friday, which has the most.
II. **Peak Accident Times:** Collisions mostly happen during morning and late afternoon rush hours, particularly between 2 PM and 6 PM.
III. **Seasonal Trends:** February has fewer accidents, but they start increasing in March, with the highest in June.
IV. **Vehicle Issues:** Many accidents involve vehicle defects, especially problems with accelerators and brakes.
V. **Key Risk Factors:** The time of day and the lighting at intersections are significant factors in the severity of accidents.

Recommendations:

I. **Better Traffic Control on Busy Days:** Increase police presence and traffic monitoring on Fridays and during peak hours to manage and reduce accidents.
II. **Prepare for Seasonal Changes:** Launch safety campaigns before March and maintain roads well before the summer to handle increased traffic and prevent accidents.
III. **Improve Infrastructure:** Improve lighting at dangerous intersections and ensure roads are well-maintained to enhance visibility and safety.
IV. **Regular Vehicle Checks:** Encourage or require regular checks on critical vehicle parts like brakes and accelerators to prevent malfunctions that could lead to accidents.
V. **Use Predictive Analytics:** Apply models like the random forest to predict and mitigate risky traffic conditions in real-time.
VI. **Educate Drivers:** Conduct educational campaigns about safe driving practices, focusing on peak accident times and seasons.
VII. **Adjust Traffic Signals and Signs:** Modify traffic signals and signage at intersections with high accident rates to make them safer.
VIII. **Engage the Community:** Involve residents in reporting unsafe conditions or behaviors to authorities.

CHAPTER 8

CONCLUSION

In conclusion, the project took the challenge of improving public safety and traffic supervision by designing a model that could accurately foresee the severity of the road accidents (MVCs). Leveraging advanced predictive analytics and machine learning tools, namely logistic regression, decision tree, and random forest classifiers, we've created a model that gives accurate and suitable information about the incident of traffic accidents. Our results have highlighted prominent factors including time of day, weather situations, and particular types of intersections that play a determining role for the degree of collision severity. These knowledge bases are very important to city planners and traffic safety officers as they provide a strong foundation for strategies like traffic lights of excellence and streetlights in high-risk areas. Looking ahead, combining real-time data with incorporating more variables of socio-economic level and driving behavior patterns could lead to the improvement of our model's predictive power. The success of this project not only indicates the power of data-driven management in urban safety but also sets a stage for future research and applications, which may work beyond New York City to other urban zones facing similar traffic safety problems.

GITHUB REPOSITORY LINK

https://github.com/teju1p/603_group_project

REFRENECES

[1] *Motor Vehicle Collisions – Crashes, Data.gov, Mar. 1, 2024.* [Online]. Available: https://catalog.data.gov/dataset/motor-vehicle- collisions-crashes/resource/b5a431d2-4832-43a6- 9334-86b62bdb033f