**NYC Vehicle Collision Analysis**

**IST-718: Big Data Analytics**

**Group No: 2-2**

**Group Members:**

Austin Rodrigues

Abhishek Shinde

Neha Sharma

Prakruti Kumar

**Project Overview:**

The NYC Vehicle Collision Analysis project leverages advanced data analytics and machine learning techniques to enhance urban safety and efficiency. The objective is to develop a forecasting model using historical collision and weather data to predict crash occurrences and identify contributing factors. By employing a GBT algorithm, the project uncovers temporal trends, such as collision frequency variations across different times, seasons, and weather conditions, and reveals borough-specific insights, with Brooklyn experiencing the highest number of accidents.

**Prediction:**

The prediction method uses machine learning—more especially, a Gradient Boost Tree model or commonly known as GBT model — to calculate the probability of car crashes using past data. The goal is to anticipate the likelihood of collisions under various circumstances, including weather, time, and location, and to detect periods of high accident rates by examining historical trends.

Preprocessing was done on the data to do this, which included resolving missing and unbalanced data and extracting pertinent elements such locational, meteorological, and temporal attributes.

The GBT model was optimized for best results since it can handle complex, non-linear interactions. Its efficacy was demonstrated by evaluation metrics including Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), which provided borough-specific insights. These insights offer a strong support for focused safety measures and resource efficiency.

**Inference:**

Finding important factors affecting collision rates and establishing relationships between weather, time patterns, and accident frequency are the main goals of the method. The GBT model's feature importance revealed important factors including the time of day, the weather, and the nature of the roads, indicating that focused interventions like better road signage or lighting in busy locations might greatly lower crash rates.

**Goal:**

**Detect potential collisions**: Use a predictive model to determine when vehicle collisions are likely, factoring in weather, time of day, the location, etc.

**Find Causative Elements**: Examine historical records to discover critical variables and trends that play significant roles in collision rates, such as time-based and environmental elements.

**Guide Safety Strategies**: Actionable insights to help to prepare proactive mitigation strategy as example improved traffic management, resource allocation, findings help to enhance road safety.

**Analyze Temporal and Spatial Trends**: Chart collision frequency by time and season, and by borough, to identify periods and locations of high risk.

**Enable Data-Driven Decision Making**: Leverage insights for urban planning, policy development, and real-time response tactics to reduce collision likelihoods and enhance traffic safety.

**Data Exploration**:

A graph showing the number of percents

Description automatically generated with medium confidence

* **Findings**: Mid-year seasonal peaks indicate more summertime traffic, highlighting the necessity of increased road safety protocols. Although the downward trend points to encouraging advancements, it also highlights the necessity of maintaining and modifying interventions.

A graph showing the number of accidents by trimming

Description automatically generated

* **Findings**: Brooklyn and Queens may have greater collision rates because of their denser populations and more active traffic. This suggests that specific traffic safety measures are required in these municipalities.

**Interesting/surprising results:**

Conventionally assumed, the collisions appear mostly during harsh weather conditions like snow and rain. Surprisingly, through analysis, highest number of collisions occurred during clear and dry days.

A graph showing weather conditions

Description automatically generated

**Findings**: Despite predictions, clear weather causes more incidents, either because of increased traffic or complacent drivers. This implies that dangerous driving practices should be addressed in clear weather as well as in inclement weather during safety programs.

**Summary of methods used to solve the problem:**

**Data Preparation and Collection**:

The data was thoroughly cleaned and made to provide accurate information.  
The data went through preprocessing, which included using class balancing and imputation techniques to handle missing and unbalanced data.  
To improve the model's predictive ability, feature engineering was used to extract pertinent variables such locational elements, weather patterns, and temporal trends.

**Exploratory Data Analysis (EDA):**

To examine collision trends across boroughs, months, and weather conditions, visualizations including bar charts and line graphs were made. Borough-specific accident rates, and the unexpected discovery that clear weather had the greatest collision rates were among the important themes that EDA identified.

**Machine Learning Model**:

A lot of models were trialed and tested on the data to find the perfect fit for the project. Models included Linear Regression, Random Forests, and Gradient Boosted Trees (GBT), Naives Bayes and finally the GBT model was selected due to its ability to handle intricate, non-linear interactions with resilience. The processed dataset was used to train the model, and hyperparameter optimization was used to guarantee peak performance. The accuracy of the model was assessed across boroughs using performance indicators such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

**Model Evaluation and Refinement:**

To evaluate accuracy, the model's predictions were contrasted with real collision data. An analysis of borough-specific variability revealed that Staten Island had lower variability, which improved accuracy, whereas Brooklyn and Manhattan had more variability, necessitating further model modifications.

**Results summary:**

This project tested several machine learning models to forecast daily and weekly motor vehicle accidents across the five boroughs of New York City, utilizing weather, temporal, and lagged features. Models included Linear Regression, Random Forests, and Gradient Boosted Trees (GBT), Naives Bayes, with varying feature transformations and evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R².  
  
Linear Regression models, both weekly and daily, provided quick baselines for the task but had limited performance. The weekly model had high MAE values (e.g., 120.81 for QUEENS) and struggled with significant borough-specific variation. The daily model improved significantly, particularly in boroughs like QUEENS (MAE: 38.81), but it failed to generalize well in others, such as the Bronx (MAE: 2223.96), indicating challenges with capturing daily variations and seasonality.

The Random Forest models performed better than linear regression. Even though the weekly Random Forest model reduced errors slightly, it had limitations in capturing temporal trends, particularly in QUEENS and BROOKLYN (MAE: 182.71 and 190.04, respectively). On the other hand, the daily Random Forest model achieved significant error reductions (e.g., QUEENS: MAE: 6.83, RMSE: 8.63) across all boroughs, demonstrating the advantages of using detailed temporal and scaled features. However, it lagged behind GBT models in terms of variance explanation.

Gradient Boosted Trees (GBT) models emerged as the most effective approach. The GBT model with daily features consistently had the lowest MAE and RMSE across most boroughs. For example, QUEENS had an MAE of 4.56 and RMSE of 7.87, while STATEN ISLAND had a R² of 0.715, indicating a strong ability to explain variance in the data. integrating lagged features significantly improved temporal predictions by capturing trends and dependencies from prior days. BROOKLYN and BRONX had lower R² values (0.026 and -1.008, respectively), indicating potential for feature engineering to improve prediction accuracy.

The GBT model with log-transformed features also produced competitive results. Log transformations improved the handling of skewed data, reducing errors in specific boroughs. The MANHATTAN model had the lowest MAE (3.37) and RMSE (6.38), as well as a higher R² (0.243) than the standard GBT model. However, for boroughs such as QUEENS and BRONX, it performed slightly worse in variance explanation than the standard GBT model.

In conclusion, GBT model performed best in STATEN ISLAND and QUEENS, showing the best balance between variance explanation and error reduction across boroughs. The performance of the model was greatly enhanced with the addition of lagged features and appropriate scaling. However, differences in R2 values by borough, particularly in BRONX and BROOKLYN, indicate that additional feature engineering or ensemble techniques are required to capture underlying complexities. Because of its better error metrics, the GBT with log-transformed features is advised for MANHATTAN. All things considered, the findings highlight how crucial it is to use sophisticated models and feature transformations to forecast auto accidents and direct data-driven public safety decision-making.

**Problems encountered:**

**Missing and Incomplete Data**: Imputation techniques were needed to fill in the dataset's missing information in crucial fields like weather and precise collision positions.

**Model Performance Variability:** The difficulty of modeling data with broad ranges of outcomes was highlighted by the Random Forest model's higher error rates in boroughs with more collision variability (such as Brooklyn and Manhattan).

**Efficiency and Scalability:** To handle scalability without sacrificing computing efficiency, the pipeline had to be optimized for processing huge datasets with numerous variables.

**Citations:**

<https://anushkasandesara.medium.com/predictive-analytics-on-nyc-collision-data-9f06c94140f2>

<https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/about_data>