**Vehicle Collision Severity Prediction: Final Report**

**Stakeholder**

The primary stakeholder for this project is the New York City Department of Transportation (NYC DOT) and the New York Police Department (NYPD). These organizations aim to reduce the number of serious motor vehicle collisions by identifying high-risk areas and potential factors contributing to severe crashes, enabling proactive measures to improve road safety.

**Problem to Solve**

The problem at hand is to predict the severity of motor vehicle collisions in New York City. Specifically, the goal is to predict serious collisions, defined as those resulting in injuries or fatalities. By accurately identifying these collisions in advance, NYC DOT and NYPD can allocate resources more effectively, implement targeted safety measures, and ultimately reduce the number of serious crashes and improve public safety.

**Dataset Source**

The dataset used for this project is the “Motor Vehicle Collisions - Crashes” dataset provided by NYC Open Data. The dataset contains crash records from 2012 to 2025, offering detailed information on over 2 million collisions in New York City. It includes attributes such as the location of the crash, time, contributing factors, vehicle types, and more.

* Dataset link: NYC Open Data - Motor Vehicle Collisions

**Models Tried and Reason for Selection**

Several machine learning models were explored to predict serious collisions:

1. **Logistic Regression (Baseline Model)**
   * **Reason for selection**: A simple and interpretable model to serve as a baseline for comparison. Logistic regression is often used for binary classification problems and provides initial insights into the dataset.
   * **Pros**: Easy to implement and interpret.
   * **Cons**: Struggled with class imbalance and had poor performance in predicting serious crashes.
2. **Random Forest Classifier**
   * **Reason for selection**: An ensemble model that aggregates decisions from multiple trees, Random Forest handles class imbalance better and captures non-linear relationships between features.
   * **Pros**: Robust to overfitting, better than logistic regression on unbalanced datasets.
   * **Cons**: Less interpretable and computationally expensive.
3. **LightGBM**
   * **Reason for selection**: A gradient boosting framework that excels at handling large datasets, especially when there are imbalances in class distribution. It is fast, scalable, and has high performance.
   * **Pros**: High accuracy handles imbalanced classes well, faster training time compared to other gradient boosting models.
   * **Cons**: Requires careful hyperparameter tuning.
4. **XGBoost (Best Performing Model)**
   * **Reason for selection**: Known for its effectiveness in classification tasks, especially in structured/tabular data with imbalanced classes. XGBoost typically yields high performance across different domains.
   * **Pros**: Very strong predictive power handles large datasets and imbalanced classes well, good for feature importance analysis.
   * **Cons**: More complex to tune, takes longer to train compared to simpler models like logistic regression.

**Features Selected and Engineered**

Feature selection and engineering were crucial in improving the model's performance. The following features were selected and engineered:

1. **Time-related Features**:
   * IS\_LATE\_NIGHT: Indicator for crashes occurring between 12:00 AM - 5:00 AM. Late-night crashes tend to be more severe.
   * IS\_WEEKEND: Indicator for crashes occurring on weekends (Saturdays and Sundays). Weekend crashes tend to involve more serious incidents due to factors like higher alcohol consumption.
2. **Borough and Location Features**:
   * BOROUGH: Identified boroughs with higher crash severity. For example, Brooklyn and Manhattan tend to have more crashes but also more serious incidents.
   * BOROUGH\_STREET\_RISK: A composite feature accounting for crash frequency per street and borough.
3. **Vehicle and Contributing Factor Features**:
   * IS\_BIKE\_INVOLVED: Indicator for crashes involving bicycles. These incidents tend to have a higher severity due to vulnerable road users.
   * CONTRIBUTING\_FACTOR\_\*: Multiple binary features were created based on known contributing factors such as “Driver Inattention”, “Unsafe Speed”, and “Alcohol Involvement”.
4. **Day of the Week**:
   * DAY\_OF\_WEEK: Categorized into days of the week to identify patterns in collision severity across different days.

The features were selected based on domain knowledge (e.g., time and location) and EDA insights, such as the higher likelihood of severe crashes during late-night hours or weekends.

**Model Evaluation and Metrics**

Models were evaluated based on precision, recall, F1 score, accuracy, and AUC-ROC. Since the dataset was imbalanced, metrics like precision and recall were more critical than accuracy. Here's why:

* **Precision**: Measures how many of the predicted serious crashes are serious. High precision ensures the model isn't flagging too many false positives.
* **Recall**: Measures how many actual serious crashes are correctly identified. High recall ensures that serious crashes are being detected.
* **F1 Score**: A balance between precision and recall, crucial in scenarios where both false positives and false negatives are costly.

The XGBoost model provided the best balance between precision and recall, with an F1 score of 54% and recall of 75%.

**Future Work and Improvements**

Given more time or additional data, the following could be explored:

1. **Hyperparameter Tuning**: Further optimization of the hyperparameters could lead to improved model performance.
2. **Additional Features**: Integrating weather data, traffic data, or real-time data from traffic cameras might enhance model accuracy.
3. **Resampling Techniques**: Use of SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance more effectively.
4. **Model Ensembles**: Combining multiple models into an ensemble to further boost performance and robustness.

**Recommendation for Client Use**

While the precision and recall of the current model (especially the LightGBM model) are not perfect, they are adequate for identifying a significant number of serious crashes (recall of 77%) and are much better than random guessing. The model is well-suited for providing early warnings for potential serious crashes, which can be used by NYC DOT and NYPD to allocate resources to high-risk areas.

Given the high recall and reasonable precision, this model can be deployed in a real-time monitoring system to flag crashes for immediate response, particularly in areas identified as high-risk locations.

**Deployment Strategy**

The model will be deployed as part of a real-time traffic monitoring system. Key aspects of deployment include:

1. **Integration with Existing Traffic Systems**: The model can be integrated into the NYC emergency response system or other traffic monitoring dashboards to provide alerts for serious crashes as they are reported.
2. **Real-Time Data Ingestion**: New crash data would be ingested and processed regularly (daily or weekly) to keep the model updated.
3. **API for Predictions**: The model would be deployed as an API that can be queried for real-time predictions, enabling rapid alerts to be sent to the relevant authorities.

**Project Overview**

The goal of this project is to develop a predictive machine learning model to identify serious motor vehicle collisions in New York City. A serious collision is defined as any crash that results in one or more injuries or fatalities. Utilizing the publicly available “Motor Vehicle Collisions - Crashes” dataset from NYC Open Data, the project analyzes over 2.1 million crash records from 2012 to 2025. The end objective is to support traffic safety efforts by helping predict which types of collisions are more likely to be severe, thereby enabling better-targeted interventions.

**Data Source and Preprocessing**

The dataset originally contained over 2.1 million rows and more than 50 columns. After removing columns with more than 90% missing values (e.g., VEHICLE DAMAGE, LOCATION ZIP), and rows with null values in critical fields like BOROUGH, LATITUDE, LONGITUDE, and vehicle-related features, the cleaned dataset comprised 1,377,073 rows.

A new target column SERIOUS\_CRASH was engineered. It is binary:

* 1 if any of the following were greater than 0: NUMBER OF PERSONS INJURED, NUMBER OF PERSONS KILLED, NUMBER OF PEDESTRIANS INJURED, NUMBER OF CYCLIST KILLED, etc.
* 0 otherwise.

Resulting label distribution:

* Non-serious crashes (0): 1,216,760 (88.35%)
* Serious crashes (1): 160,313 (11.65%)

Feature engineering added columns like:

* IS\_LATE\_NIGHT: 1 if crash time between 12:00 AM and 5:00 AM
* IS\_WEEKEND: 1 for Saturday or Sunday
* IS\_BIKE\_INVOLVED: 1 if vehicle types included "Bike", "Bicycle", or "E-Bike"
* BOROUGH\_STREET\_RISK: A composite metric based on crash frequency per street and borough

Categorical features were encoded using one-hot encoding. The dataset was split into 80% training (1,101,658 rows) and 20% testing (275,415 rows) using stratified sampling to maintain class balance.

**Exploratory Data Analysis (EDA)**

Key observations from EDA include:

* Weekend crashes accounted for 27% of total crashes, but 36% of serious crashes
* Late night crashes (12:00–5:00 AM) made up only 8% of total crashes but 14.7% of serious crashes
* Bicycle-involved collisions had a serious crash rate of 21.4%, compared to 11.2% in the general dataset
* Top contributing factors in serious crashes:
  + Driver Inattention/Distraction: **19.2%**
  + Unsafe Speed: **8.7%**
  + Alcohol Involvement: **3.6%** (compared to **1.1%** in non-serious crashes)

A heatmap of crash density revealed significant clustering in boroughs like Brooklyn, Manhattan, and the Bronx, with specific corridors repeatedly showing high crash severity.

**Visualizations:**

**A graph of blue bars

AI-generated content may be incorrect.**

**Avg. Seriousness by Contributing Factor**

****"Failure to Yield Right-of-Way" is the most serious contributing factor, with the highest average crash seriousness (~0.42), indicating it's a leading cause of severe accidents.

"Passing Too Closely" and "Backing Unsafely" are among the least serious factors, suggesting these behaviors tend to result in less severe crashes.

**A graph of a crash

AI-generated content may be incorrect.**

**Serious vs Non-Serious Crashes by Hour of Day**

**** While the **absolute number** of crashes is low (during 12AM – 5 AM), the **ratio of serious to non-serious crashes is higher** compared to other hours.

This suggests that **crashes occurring during late-night hours tend to be more severe**.

 Non-serious crashes are more frequent overall, but serious ones spike more sharply during these hours.

**A graph of a number of years

AI-generated content may be incorrect.**

**Serious Crashes Over the Years**

** 2013–2018** had the **highest number of crashes**, peaking around 2017-2018.

 **Non-serious crashes ("No")** always significantly outnumber **serious crashes ("Yes")**.

 There is a **noticeable decline** in both categories starting in **2019**, with a **sharp drop in 2020**, possibly due to the COVID-19 pandemic reducing traffic volume.

 From **2020 onwards**, both types continue to decline gradually, with **2025 showing the lowest counts** (likely incomplete data or prediction).

**A graph with red and blue squares

AI-generated content may be incorrect.**

**Serious vs Non-Serious Crashes by Day of Week**

**** Insight: Weekends (Saturday, Sunday) show higher proportions of serious crashes, despite fewer total incidents.

 May reflect drunk driving, speeding, or lower seatbelt use during leisure trips.

**A graph of a crash

AI-generated content may be incorrect.**

**Serious Crashes by Vehicle Type**

**** Sedans are involved in the highest number of serious crashes, suggesting their prevalence and possibly higher vulnerability or exposure on the roads.

 Passenger Vehicles and Station Wagons/SUVs also show a significant count of serious crashes, indicating these common vehicle types contribute notably to severe incidents.

**A graph of a bar graph

AI-generated content may be incorrect.**

**Top 20 Borough-Street Locations by Crash Count**

**** QUEENS\_NORTHERN BOULEVARD and QUEENS\_QUEENS BOULEVARD have the highest crash counts among all street-borough combinations, indicating critical areas for targeted traffic safety interventions.

 BROOKLYN\_ATLANTIC AVENUE and MANHATTAN\_BROADWAY also show notably high crash frequencies, suggesting dense traffic or complex road conditions in those locations.

**A graph of a number of injuries

AI-generated content may be incorrect.**

 **Most Dangerous Location**:

* **Brooklyn\_Atlantic Avenue** has the highest serious injury count, just under 1200.
* Followed closely by **Queens\_Northern Boulevard**.

 **High-Risk Boroughs**:

* **Brooklyn** dominates the list with **7 out of 20** streets.
* **Queens** follows with **5 streets**.
* **Manhattan** contributes **4**, **Bronx** with **2**, and **Staten Island** with **1** (Hylan Boulevard).

**A graph with dots and numbers

AI-generated content may be incorrect.**

**Clusters of Serious Crashes (K-Means on Map)**

Map of New York City showing 10 clusters of serious crashes using K-Means clustering.

* Spatial Trends: Serious crashes cluster around major avenues like Flatbush Ave, Grand Concourse, and 125th Street.
* Urban Safety Planning: Helps direct city resources to danger zones for proactive crash mitigation.

**Model Development and Evaluation**

**Baseline Logistic Regression**

* Accuracy: 73%
* Precision (serious crashes): 53%
* Recall (serious crashes): 6%
* F1 Score (serious crashes)(1): 11%

Despite high accuracy, the model failed to detect serious crashes due to class imbalance.

**A number of numbers on a white background

AI-generated content may be incorrect.**

**Random Forest Classifier**

* Accuracy: 66%
* Precision: 35%
* Recall: 29%
* F1 Score: 32%

Improved detection of serious crashes, but still limited sensitivity.

A screenshot of a computer screen

AI-generated content may be incorrect.

**A graph of a vehicle

AI-generated content may be incorrect.**

**LightGBM (with tuned scale\_pos\_weight=8)**

* Accuracy: 55%
* Precision: 35%
* Recall: 77%
* F1 Score: 48.0%
* LightGBM significantly improved recall, making it better for practical use where catching serious crashes is critical.

A number of numbers in a row

AI-generated content may be incorrect.

A graph with text and numbers

AI-generated content may be incorrect.

**XGBoost**

* Hyperparameters:
  + max\_depth=10, n\_estimators=300, scale\_pos\_weight=3
* Accuracy: 60.1%
* Precision: 36.4%
* Recall: 64.6%
* F1 Score: 46.5%

A number of numbers on a white background

AI-generated content may be incorrect.

A graph with text and numbers

AI-generated content may be incorrect.

This confirms the significant influence of time, location, and vehicle type on crash severity.

**A graph with a line

AI-generated content may be incorrect.**

**Precision-Recall Curve**

High average precision suggests the model is particularly effective at identifying true positives.

 Curves are close to top-right, which is ideal.

 For imbalanced datasets, this confirms robustness of the model.

**Spatial Clustering of Serious Crashes**

Using K-Means clustering, serious crash locations were grouped into 10 clusters. These represent geographic crash hotspots, with clusters concentrated along:

* Flatbush Avenue and Atlantic Avenue (Brooklyn)
* 125th Street (Harlem)
* Grand Concourse (Bronx)
* Lower Manhattan and Midtown intersections

The cluster centroids were plotted on an interactive map using Folium, visually highlighting areas requiring urgent road safety interventions.

**Conclusion and Recommendations**

This project shows that XGBoost and LightGBM models can robustly predict serious vehicle collisions using NYC crash data, even in the presence of class imbalance. With F1-scores around 48% and recall rates above 75%, these models can be used in real-time traffic monitoring systems to alert authorities of potentially serious incidents.

**Recommendations**:

* Deploy model with LightGBM in dashboard environments for DOT, NYPD, or 311 operators.
* Integrate weather and traffic speed data to further improve model accuracy.
* Apply SMOTE or cost-sensitive learning for deeper imbalance handling.
* Use clustering output to guide city planning and crash prevention programs (e.g., speed bumps, bike lane protections).
* Conduct monthly model retraining to keep the system updated with the most recent trends in urban traffic safety.