EMSE 6574: Programming for Analytics

Predicting Motor Vehicle Crash Severity in New York City

A Machine Learning Approach to Public Safety

Team:

Taekwon Choi, Vidyullatha

Objectives & Dataset Overview

Goal: Build a model to predict High-Severity motor vehicle crashes in NYC.

High Severity = A crash resulting in >= 2 injuries OR >= 1 fatality

Data Source: NYC Open Data API

Dataset Overview:

- 100,000 Motor Vehicle Collision Records (Oct 2024 – Dec 2025)
- Features: Time, Location (Borough), Victim Counts, and Contributing Factors

Key Challenge: Class Imbalance; Only ~ 10% of crashes are high severity. Th is a critical factor for model evaluation

Exploratory Data Analysis

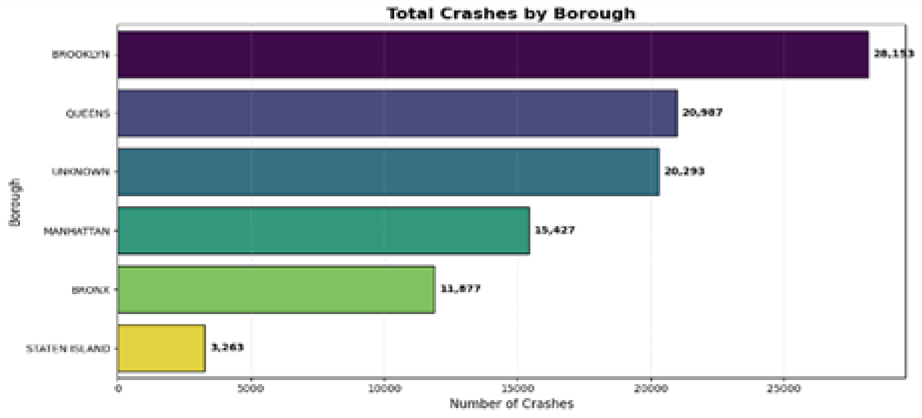
Geographic Distribution: <u>**Brooklyn**</u> has the highest raw volume of crashes, but the <u>**"UNKNOWN" borough**</u> showed the highest severity rate(likely due to major highways/bridges on the city's edge)

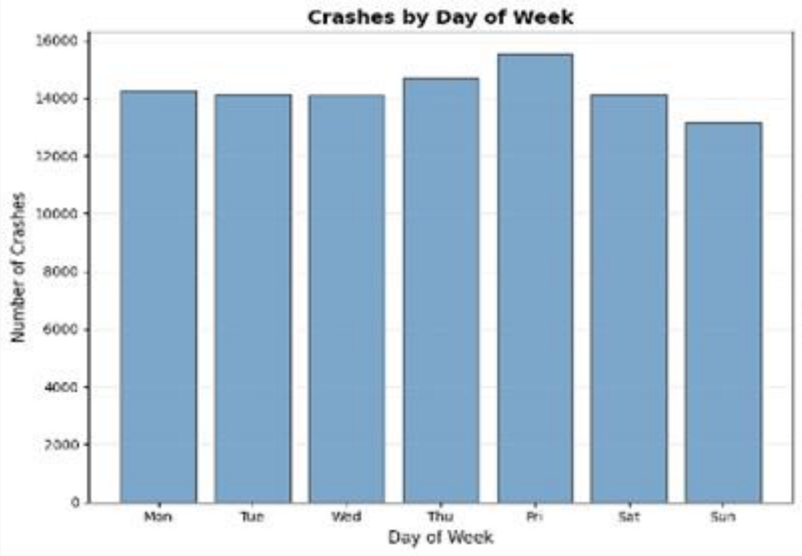
Temporal Patterns: Peak crash frequency is at <u>**5:00 PM**</u> (evening rush hour).

Severity by Hour: The highest severity rates occur during late night/early morning hours (<u>**9 PM, 11 PM, Midnight**</u>), suggesting reduced visibility and potentially higher speeds are linked to more severe outcomes

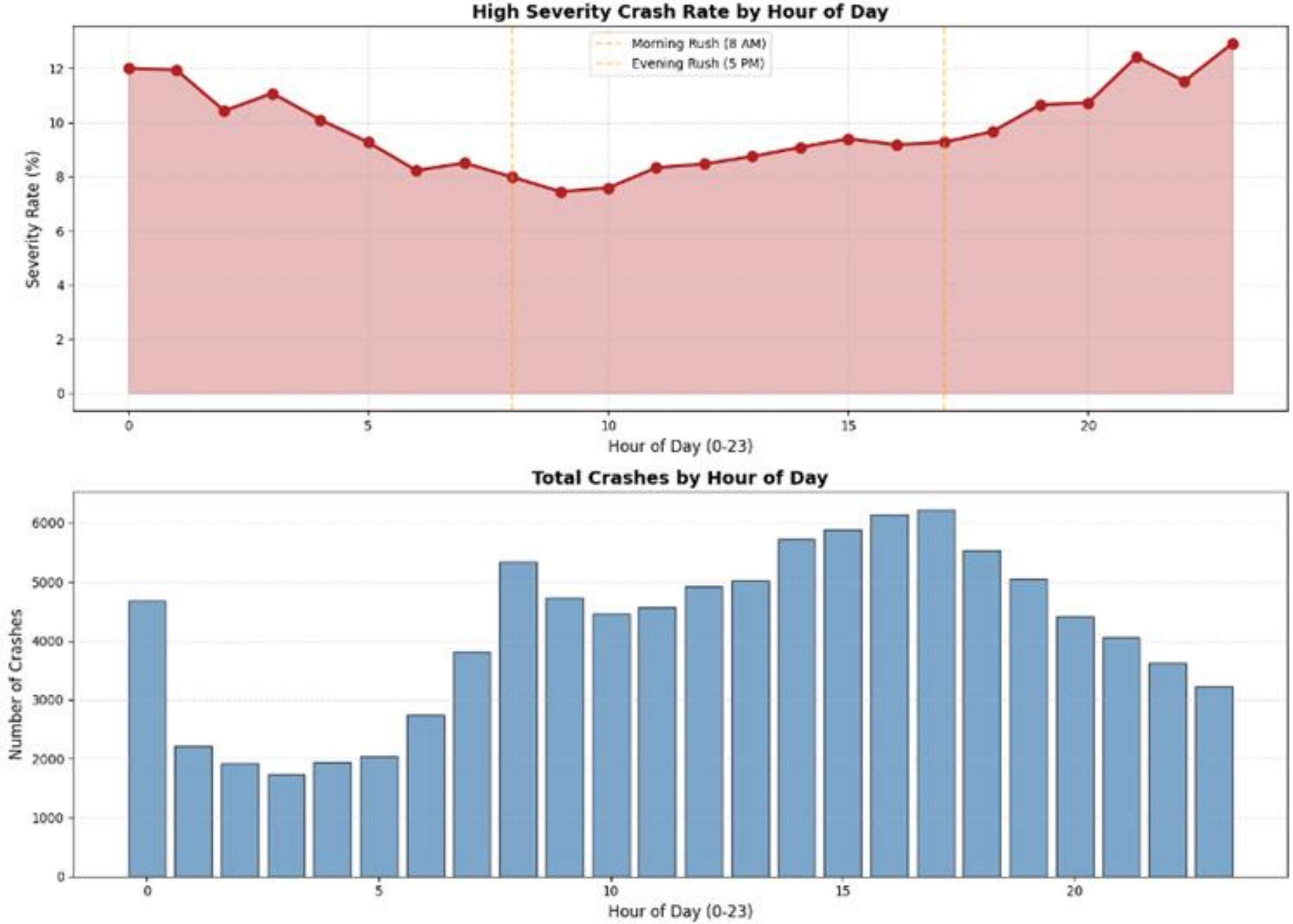
Contributing Factors:

- **Most Common:** Driver Inattention/Distraction
- **Highest Severity Rate:** Lost Consciousness, Illness, Unsafe Speed, and Unsafe Lane Changing

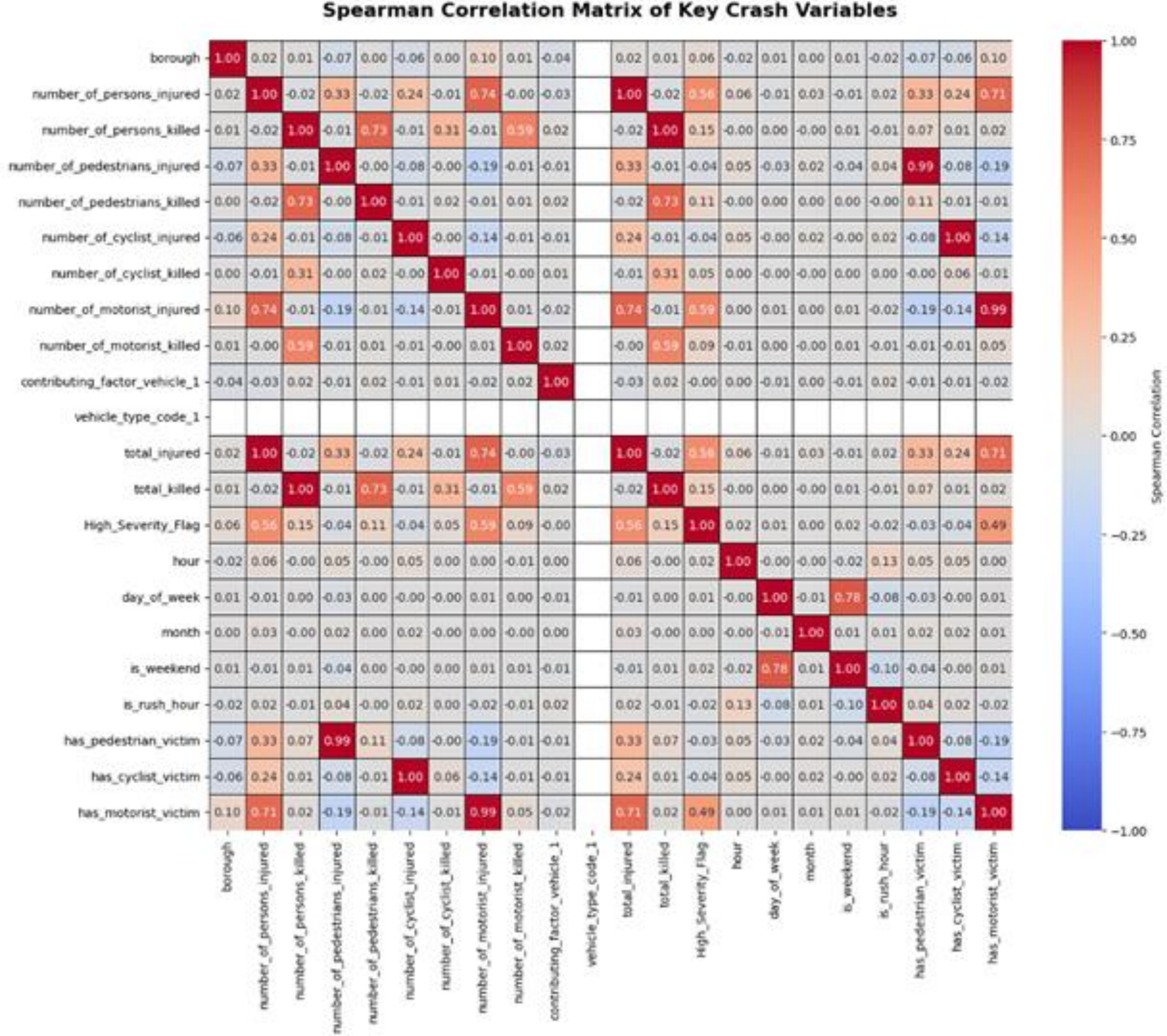

Total Crashes by Borough
BROOKLYN
QUEENS
UNKNOWN
MANHATTAN
BRONX
STATEN ISLAND
28,153
20,987
20,293
15,427
11,877
3,263
Borough
0
5000
10000
15000
20000
25000
Number of Crashes


Crashes by Day of Week
16000
14000
12000
10000
8000
6000
4000
2000
0
Number of Crashes
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Day of Week


Crashes by Month
14000
12000
10000
8000
6000
4000
2000
0
Number of Crashes
1
2
3
4
5
6
7
8
9
10
11
12
Month



High Severity Crash Rate by Hour of Day
Morning Rush (8 AM)
Evening Rush (5 PM)
Severity Rate (%)
Hour of Day (0-23)
Total Crashes by Hour of Day
Number of Crashes
Hour of Day (0-23)



Summary of the Correlation Matrix

```
Top Correlations with High_Severity_Flag:

Positive Correlations:
number_of_motorist_injured        0.593642
number_of_persons_injured         0.559409
total_injured                     0.559409
has_motorist_victim               0.490196
number_of_persons_killed          0.155193
total_killed                      0.155193
number_of_pedestrians_killed      0.112852
number_of_motorist_killed         0.092122
borough                           0.055329
number_of_cyclist_killed          0.049498
Name: High_Severity_Flag, dtype: float64

Negative Correlations:
is_weekend                         0.016115
day_of_week                        0.012210
month                              0.000646
contributing_factor_vehicle_1     -0.002118
is_rush_hour                      -0.015141
has_pedestrian_victim             -0.030961
number_of_pedestrians_injured     -0.039370
has_cyclist_victim                -0.042346
number_of_cyclist_injured         -0.044601
vehicle_type_code_1                     NaN
Name: High_Severity_Flag, dtype: float64
```

Machine Learning Methods

Binary Classification:

Target variable (High_Severity_Flag): Yes/No

Engineered 13 features:

- Time-based: Hour, Day of Week, Weekend Flag
- Victim Counts: Total Injured, Total Killed
- Categorical: Borough, Contributing Factor (One-Hot Encoded)

Models Tested

1. Logistic Regression (Baseline)
2. Random Forest Classifier
3. Gradient Boosting Classifier
4. Logistic Regression with SMOTE (for imbalance)

Machine Learning Methods

Evaluation Metrics

Primary Metric:

ROC AUC (Measures overall separability, robust to imbalance)

Secondary Metrics:

Precision, Recall, F1-Score (Crucial for class imbalance)

Model Selection Rationale

Safety Focus:

In public safety, Recall (correctly identifying all severe crashes) is paramount, as a False Negative (missing a severe crash) is costly

Results

Best Overall Model:

(ROC AUC) Gradient Boosting (ROC AUC: 0.9055)

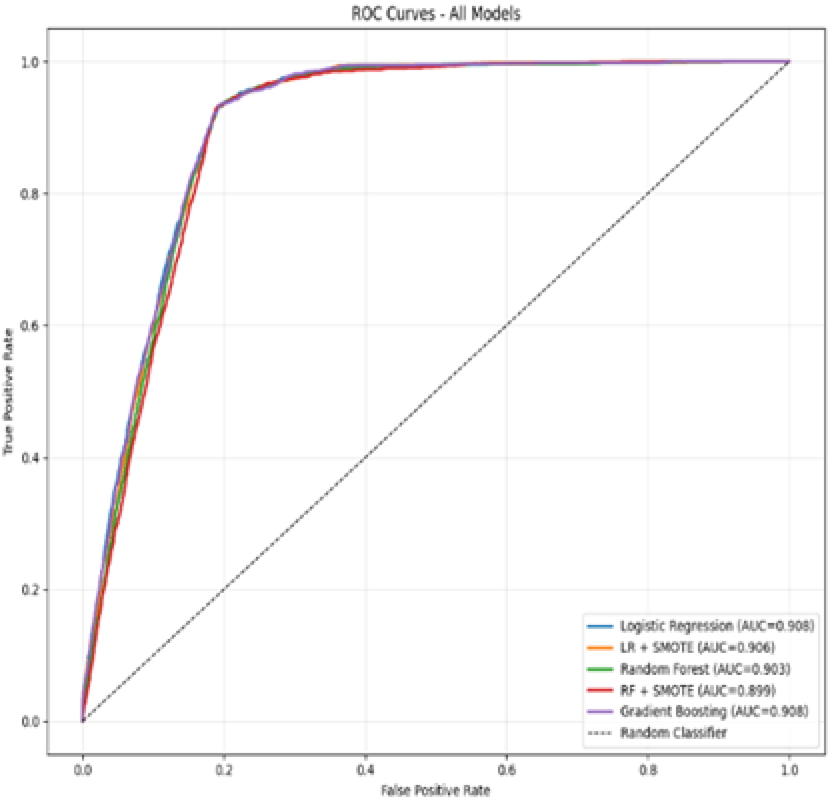
Model Chosen for Deployment:

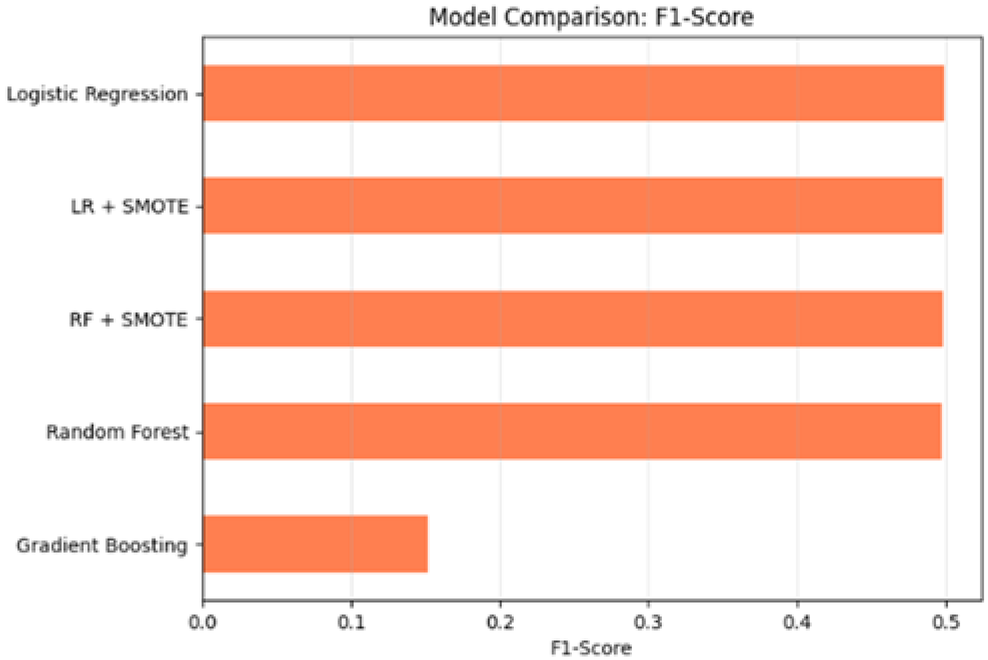
Logistic Regression (ROC AUC: 0.9040)

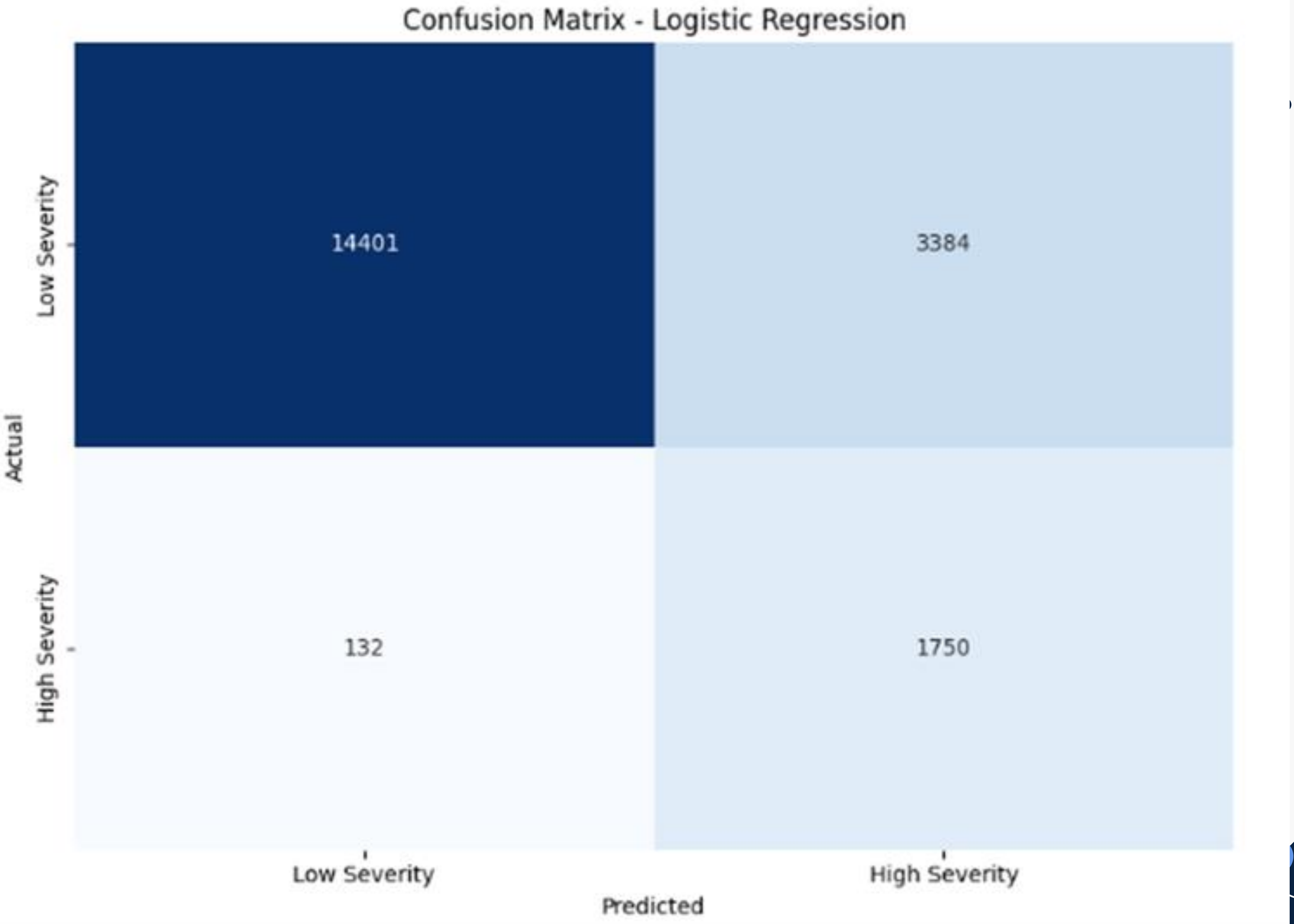
Final Model Performance:

(Logistic Regression)

- Accuracy: 90.5% (High due to imbalance)
- Precision: 0.47
- Recall: 0.53 (Best balance for our goal)
- F1-Score: 0.50


ROC Curves - All Models
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (AUC=0.908)
LR + SMOTE (AUC=0.906)
Random Forest (AUC=0.903)
RF + SMOTE (AUC=0.899)
Gradient Boosting (AUC=0.908)
Random Classifier


Model Comparison: F1-Score
Logistic Regression
LR + SMOTE
RF + SMOTE
Random Forest
Gradient Boosting
0.0
0.1
0.2
0.3
0.4
0.5
F1-Score


Confusion Matrix - Logistic Regression
Actual
Low Severity
High Severity
14401
3384
132
1750
Low Severity
High Severity
Predicted

Key Takeaway: **Features related to Total Injured, Total Killed, and the most severe Contributing Factors (e.g., Unsafe Speed) were the strongest predictors**

Reflections:

- Successfully collected and processed a large, real-world API dataset.

- Comprehensive EDA uncovered actionable patterns (time, geography, factors).

- Model selection was driven by a practical, safety-focused metric (Recall) rather than just the highest ROC AUC.

Future Work:

- **Feature Engineering:** Integrate external data like **weather conditions** and **road types** for stronger prediction.
- **Model Tuning:** Experiment with **class-weighted models** (e.g., XGBoost) or tuning the **prediction threshold** to further boost Recall.

Thank you!