CS777 Final Project – Ziwen Lyu

Prediction of Crash Severity in NYC by using Logistic Regression Model

# Introduction

Motor vehicle crashes are significant public safety issue, especially in population dense cities like New York City (NYC). Learning how various features associated with crash events can suggest authorities to take preventive measure to reduce the collision frequency and severity. This study aims to understand what features and how they influence crash severity in NYC, and build machine learning models to predict the severity based on raw data collected from the City of New York government. We will implement two machine learning models: Decision Tree and weighted Linear Regression, and evaluate their effectiveness through multiple metrics.

# Research Questions;

The primary goal of this research is to investigate how the time of day, months, locations, crash contributing factors, and vehicle types associates with collision severity level, and whether we can build a machine learning model based on basic crash information. If the model is able to predict severity with first-hand information, the authorities can prioritize and allocate emergency resources more effectively, as well as plan the management resourcefully.

1. Are certain boroughs in NYC and certain periods of a day more prone to have vehicle collisions in 2022?
2. How do certain crash factors like human behaviors (e.g., speeding, alcohol use) or environmental factors (e.g., road conditions) impact the severity in vehicle collisions?

# Methodology

Data Collection:

The data is from the City of New York government, which contains every crash event in New York City from 2012 to 2024 with a least $1000 worth of damage. This data contains 2,124,646 unique crash events (rows) and 29 columns. (*The data source:* [Motor Vehicle Collisions - Crashes - Catalog (data.gov)](https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes))

Since 2023 and 2024 missing numbers of crash events, this research will specifically choose year 2022 with 103,886 incidents, and use features Crash date, Crash time, Borough, number of persons injured, number of persons killed, contributing factor 1, vehicle type code 1.

|  |  |  |
| --- | --- | --- |
| Features | Description | value |
| Crash date | The date when crash happened | “1/25/2022” |
| Crash time | The time when crash happened | “15:34” |
| Borough | The location where crash happened | “BROOKLYN” |
| Number of persons injured | The number of people injured in crash | 5 |
| Number of persons killed | The number of people killed in crash | 1 |
| Contributing factor 1 | The reason why the crash happened | “Alcohol Involvement” |
| Vehicle type code 1 | The type of vehicle involved in crash | “Bus” |

For the dependent variable (“severity level”) is computed based on the number of persons killed/injured. If there’s any fatal or injuries equal or over to 5 persons, the crash will be categorized as “serious crash”( 0.06%), and if the injuries are limited from 1 to 4, the severity level will be “moderate crash”(36.6%), and if no one is injured, the severity level will be “no injury crash” (62.8%). This distribution suggests that our data is imbalanced, potentially leading to neglect minor classes. Since the number of serious crashes is minimal, we will combine the moderate crash and serious crash later and use weights in our machine learning prediction. For analysis purpose, we will keep these 3 categories.

Feature Engineering

Binning

Based on the first exploration of our data, there are 55 unique values in feature “Contributing factor 1” and 355 unique values in feature “Vehicle type code 1.” Redundant values will impact the performance of machine learning model, I binned those unique values into general categories as followed:

|  |  |
| --- | --- |
| Contributing factor 1 | General Category |
| Outside Car Distraction | Distraction |
| Driver Inattention/Distraction |
| Passenger Distraction |
| Following Too Closely | Driver's improper operation |
| Traffic Control Disregarded |
| Using On Board Navigation Device |
| Unsafe Speed |
| Passing or Lane Usage Improper |
| Aggressive Driving/Road Rage |
| Driver Inexperience |
| Texting |
| Reaction to Uninvolved Vehicle |
| Cell Phone (hand-Held) |
| Failure to Keep Right |
| Unsafe Lane Changing |
| Backing Unsafely |
| Eating or Drinking |
| Failure to Yield Right-of-Way |
| Turning Improperly |
| Passing Too Closely |
| Other Electronic Device |
| Cell Phone (hands-free) |
| Illness | Health Issue |
| Prescription Medication |
| Lost Consciousness |
| Fell Asleep |
| Fatigued/Drowsy |
| Physical Disability |
| Drugs (illegal) | Illegal behavior |
| Listening/Using Headphones |
| Alcohol Involvement |
| Shoulders Defective/Improper | Road issue |
| Lane Marking Improper/Inadequate |
| Other Lighting Defects |
| Pavement Defective |
| View Obstructed/Limited |
| Obstruction/Debris |
| Traffic Control Device Improper/Non-Working |
| Glare |
| Pavement Slippery |
| Other Vehicular | Third Party Issue |
| Pedestrian/Bicyclist/Other Pedestrian Error/Confusion |
| Animals Action |
| Driverless/Runaway Vehicle | Vehicle Defection |
| Accelerator Defective |
| Windshield Inadequate |
| Tinted Windows |
| Oversized Vehicle |
| Tow Hitch Defective |
| Tire Failure/Inadequate |
| Brakes Defective |
| Headlights Defective |
| Steering Failure |
| Vehicle Vandalism |
| Unspecified | Other |

The distribution of those crash contributing factor categories is: driver’s improper operation (40.4%), distraction (25.7%), third party issue (3.7%), illegal behavior (1.8%), road issue (2.4%), vehicle defection (1.5%), other (24.3%). We can see that driver’s improper operation (e.g. speeding, cellphone, following too closely, etc.) is the main cause of crash events in NYC.

As for the vehicle type, I binned all of 355 unique values into 10 categories: sedan(47.6%), station wagon(35.3%), bus(1.9%), truck(5.0%), bike(2.6%), motorcycle(0.9%), taxi(2.7%), van(0.6%), scooter(0.6%), other(4.4%).

Time

I added one more column “month” from feature “crash date”, to share information of how months change will impact the crash severity. I also turned the feature “crash time” into a continuous variable (e.g. 8:30 into 8.5, 23:15 into 23.25).

One-hot Encoding and Polynomial Expansion

All categorical features: month, borough, crash factors, and vehicle type will be one-hot encoding and assembled with other features into feature vectors for machine learning model. To capture and uncover the complex relationship among features, I set polynomial expansion with degree of 3.

Model Development

Two machine learning models were developed for predicting collision severity:

1. Weighted Logistic Regression:

This model is implemented to predict the likelihood of “Serious/Moderate injury crash” vs. “No injury crash,” with weights to deal with the imbalance classes. We set the regularization equals to 1, and maximum iteration to 150.

1. Decision Tree:

The decision tree model is adopted as reference. We set the hyperparameter max depth as 5, max bins as 32.

Model Evaluation

We will split the training and test dataset into 8:2. The performance of the model was evaluated by F1-score, which balances precision and recall and good at dealing with class imbalance. Additionally, a Chi-Square test was conducted to assess the significance of the features used in the models.

# Findings

*Association Analysis*

1. Borough vs. Severity Level (excluded “other” value in Borough)

As we can see, when severity level is moderate crash, the top three boroughs are Brooklyn (8510, 12.59%), Queens (6320, 9.35%), and Manhattan (4135, 6.12%). When severity level is serious crash, the top three boroughs are still Brooklyn (110, 0.15%), Queens (82, 0.12%), and Bronx (74, 0.11%). And these three locations remain their top 3 when measuring the crash counts for no injury crash (Brooklyn 14367, 21.26%, Queens 11721 17.34%, Bronx 7843 11.61%).

|  |  |  |  |
| --- | --- | --- | --- |
| BOROUGH | severity\_level | count | Percentage |
| BROOKLYN | no injury crash | 14367 | 21.26% |
| QUEENS | no injury crash | 11721 | 17.34% |
| BROOKLYN | moderate crash | 8510 | 12.59% |
| BRONX | no injury crash | 7843 | 11.61% |
| MANHATTAN | no injury crash | 7587 | 11.23% |
| QUEENS | moderate crash | 6320 | 9.35% |
| MANHATTAN | moderate crash | 4135 | 6.12% |
| BRONX | moderate crash | 4125 | 6.10% |
| STATEN ISLAND | no injury crash | 1774 | 2.63% |
| STATEN ISLAND | moderate crash | 878 | 1.30% |
| BROOKLYN | serious crash | 110 | 0.16% |
| QUEENS | serious crash | 82 | 0.12% |
| BRONX | serious crash | 74 | 0.11% |
| MANHATTAN | serious crash | 37 | 0.05% |
| STATEN ISLAND | serious crash | 15 | 0.02% |

1. Time of day vs. Severity level

From the following table we can know that, the serious crash happened most during the evening (204, 0.2%). This might be due to the lightening and sight limitation. Both moderate and no-injury crash prone to happen during the afternoon (13367, 13.07%, 21393, 20.91%).

|  |  |  |  |
| --- | --- | --- | --- |
| time\_of\_day | severity\_level | count | Percentage |
| afternoon | no injury crash | 21393 | 20.91% |
| morning | no injury crash | 17176 | 16.79% |
| evening | no injury crash | 15297 | 14.95% |
| afternoon | moderate crash | 13367 | 13.07% |
| evening | moderate crash | 10563 | 10.32% |
| midnight | no injury crash | 10369 | 10.13% |
| morning | moderate crash | 8583 | 8.39% |
| midnight | moderate crash | 4930 | 4.82% |
| evening | serious crash | 204 | 0.20% |
| afternoon | serious crash | 170 | 0.17% |
| midnight | serious crash | 138 | 0.13% |
| morning | serious crash | 119 | 0.12% |

1. Month vs. Severity level

Serious and moderate crashes concentrated in June, July, and August month, while most no-injury crashes are in May, June, and March. The temperature might be another factor contributing to the severity level.

|  |  |  |  |
| --- | --- | --- | --- |
| month | severity\_level | count |  |
| 5 | no injury crash | 5853 | 5.7% |
| 6 | no injury crash | 5705 | 5.6% |
| 3 | no injury crash | 5698 | 5.6% |
| 6 | moderate crash | 3593 | 3.5% |
| 8 | moderate crash | 3484 | 3.4% |
| 7 | moderate crash | 3449 | 3.4% |
| 7 | serious crash | 75 | 0.1% |
| 6 | serious crash | 63 | 0.1% |
| 8 | serious crash | 61 | 0.1% |

1. Contributing factors vs. Severity Level

Driver’s improper operation, other, and distraction are the top 3 in serious crash, moderate crash, and no-injury crash. This Dmight be due to the imbalance in the contributing factors. But we also can know that driver’s improper operation is the main cause of crashes in NYC.

1. Vehicle type vs. Severity level

Both Sedan and Station Wagon remain the top 2 in various severity level due to large number of these two types of vehicles. But it is surprisingly to see that truck has the third most serious crashes and no-injury crashes, and bike has the third most moderate crashes.

*Model Performance*

The F1-scores and confusion metrics from the models:

Decision Tree:

Accuracy 64%, Precision: 79%, Recall: 7.9%, F1: 14.3%

Confusion Matrix:

|  |  |  |
| --- | --- | --- |
| Actual/Predicted | Positive | Negative |
| Positive | 614 | 163 |
| Negative | 7198 | 12516 |

Weighted Logistic Regression:

Accuracy 56%, Precision: 44.8%, Recall: 65.5%, F1: 53.2%

Confusion Matrix:

|  |  |  |
| --- | --- | --- |
| Actual/Predicted | Positive | Negative |
| Positive | 5115 | 6311 |
| Negative | 2697 | 6368 |

By the comparison, we can know that decision tree is better in the accuracy but tends to predict negative classes instead of positive classes, so the F1 score is incredibly low at 14%. The weighted logistic regression fixed this issue by introducing the weight to classes, so the model’s F1 score of 53.2% is much better.

But overall, the model didn’t perform as expected before (75% of F1 score), I re-examine the relationships between features and target value by using the Chi-square test.

*Chi-Square test*

The Chi-Square test shows most features had p-values less than 0.001, suggesting a strong significance with crash severity. So, the features we selected for the model are relevant predictors.

|  |  |  |
| --- | --- | --- |
| pValues | degreesOfFreedom | statistics |
| 0 | 1 | 138.9307 |
| 0.310436283 | 1 | 1.028819 |
| 6.59E-09 | 1 | 33.6508 |
| 1.45E-08 | 1 | 32.124 |
| 3.50E-05 | 1 | 17.12675 |
| 0 | 1439 | 3778.64 |
| 0 | 23 | 650.9649 |
| 1.11E-04 | 1 | 14.9453 |
| 0.868358748 | 1 | 0.027471 |
| 1.02E-09 | 1 | 37.28449 |
| 0.028491389 | 1 | 4.798085 |
| 6.64E-05 | 1 | 15.91059 |
| 1.24E-09 | 1 | 36.91144 |
| 1.49E-06 | 1 | 23.1671 |
| 0.004779251 | 1 | 7.961132 |
| 0.188020025 | 1 | 1.733064 |
| 0.002436286 | 1 | 9.187837 |
| 0 | 1 | 192.5202 |
| 0 | 1 | 124.6473 |
| 0 | 1 | 94.16311 |
| 0 | 1 | 867.5756 |
| 0 | 1 | 101.4922 |
| 3.26E-12 | 1 | 48.52373 |
| 0.010511406 | 1 | 6.546104 |
| 3.56E-10 | 1 | 39.34144 |
| 0 | 1 | 118.7818 |
| 4.20E-11 | 1 | 43.5184 |
| 0 | 1 | 249.6201 |
| 5.01E-14 | 1 | 56.72716 |
| 0 | 1 | 148.7952 |
| 0 | 1 | 2924.608 |
| 4.12E-10 | 1 | 39.05681 |
| 0 | 1 | 447.1461 |
| 0.000446106 | 1 | 12.32845 |

# Discussions

By the comparison, we see the weights in logistic regression model greatly improve the performance, but there remains a considerable gap from ideal scenarios. In the later trials, by adjusting the hyperparameters in models, introducing polynomial expansion and regularization, and binning more granularly in the features, the F1-score didn’t improve a lot for both decision tree and weighted logistic regression model. My speculation is that the model may need additional features from other data source, and current data cannot fully describe the complexity of crash severity. Based on our previous association analysis, we may include weather, traffic, speed limit, emergency needed or not, etc. in the data for prediction.

Conclusion

This study is interested in association of crash severity and predicting crash severity in NYC using the first-hand crash event data. We see strong relationships in time of day, months, location, vehicle type, and crash factors with severity level. The weighted logistic regression model performs better than the decision tree as it handles class imbalance, but overall two models are still away from expected results due to limited data source. In the future, this study should focus on integrating more features like weather related to develop a more robust model.