CS-GY 6053: FOUNDATION OF DATA SCIENCE

Project Report: Traffic Safety and Accident Analysis in New York City

Team: Kaushik Mellacheruvu, Deepjyot Singh Kapoor, Mehran Ali Banka

Problem: The problem at hand is the alarming rate of traffic accidents in New York City, posing a significant threat to public safety. These accidents result in injuries, fatalities, and substantial economic costs. The overarching problem is the lack of comprehensive insights and data-driven strategies to effectively mitigate and prevent accidents. This project aims to address this problem by analysing a dataset containing information about traffic accidents ißn the city.

Motivation: Identifying accident hotspots is crucial for prioritizing safety measures. Accurate time-series forecasting of accident trends can inform resource allocation and enhance overall safety. This project seeks to uncover patterns, contributing factors, and insights to inform policy decisions and enhance road safety.

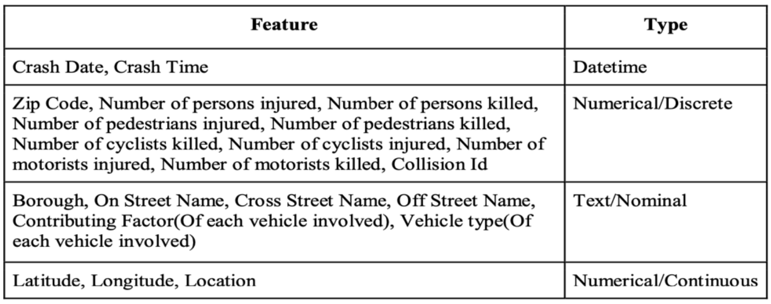
Background: We have researched previous works on how accident numbers have been analysed using Machine Learning. These primarily include:

- Machine Learning for Traffic Analysis
- Time Series Analysis of Accidents

Objective/Outcome: The objective of this comprehensive data science project is to enhance traffic safety and accident analysis in New York City by employing predictive modelling, geospatial analysis and time-series forecasting techniques.

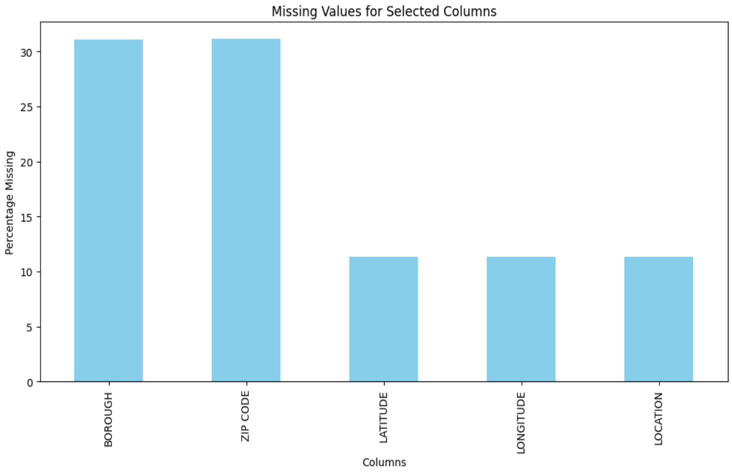
a. Prediction of seasonal variations in accident rates per borough using Time Series. This predictive model anticipates future accident rates by utilizing historical accident data.
b. Geospatial analysis that visually pinpoints areas with a high frequency of accidents, which is crucial for identifying geographical areas that demand focused safety interventions.
c. A classification model that can make predictions about the risk severity of a particular place with a given set of inputs, such as time, weather conditions, temperature etc.
d. A relevance graph of the most likely causes of accidents at a given place, given a set of inputs. For each neighbourhood, the authorities can plan measures to counter speeding appropriately.

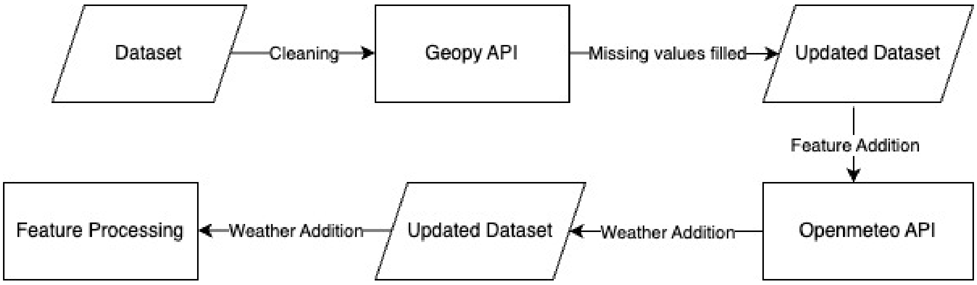
Dataset: The dataset we will use includes information on traffic accidents in New York City. The dataset is compiled and maintained by the New York City Police Department (NYPD).
Spatial Dimensions: (2034305, 30), Temporal Dimensions: (2012,2023), Source: Dataset



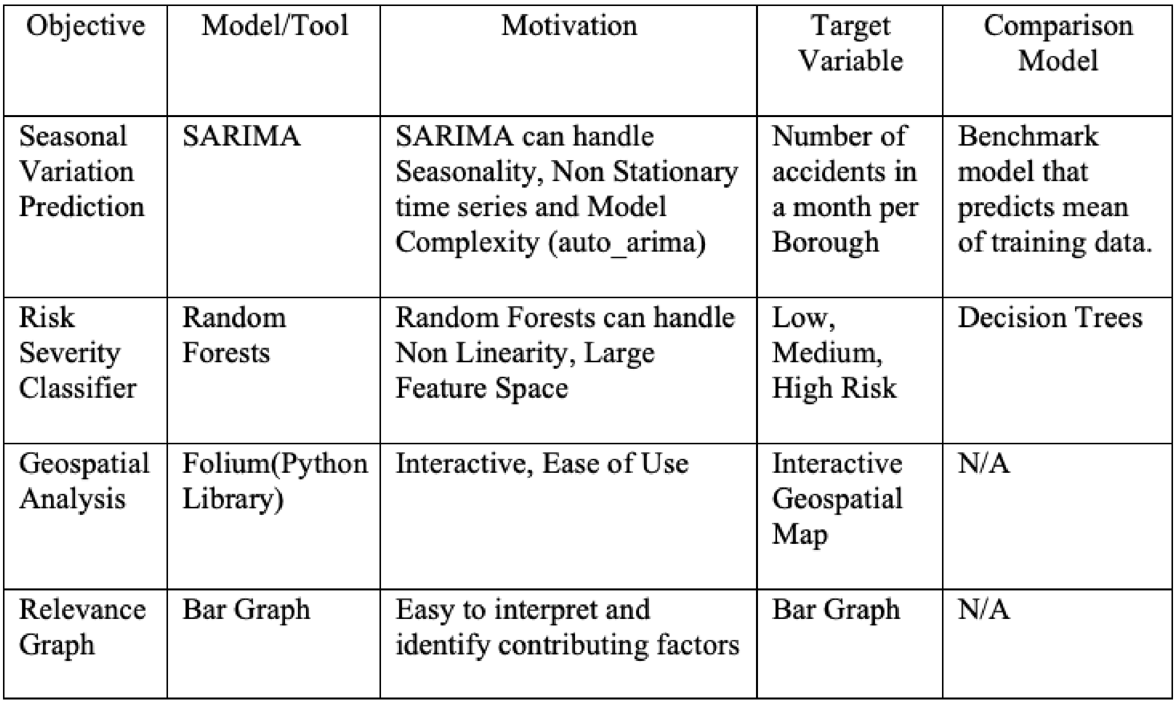
Feature	Type
Crash Date, Crash Time	Datetime
Zip Code, Number of persons injured, Number of persons killed, Number of pedestrians injured, Number of pedestrians killed, Number of cyclists killed, Number of cyclists injured, Number of motorists injured, Number of motorists killed, Collision Id	Numerical/Discrete
Borough, On Street Name, Cross Street Name, Off Street Name, Contributing Factor(Of each vehicle involved), Vehicle type(Of each vehicle involved)	Text/Nominal
Latitude, Longitude, Location	Numerical/Continuous

Data Pre-processing: Our approach involves enhancing the dataset by imputing missing values integrating new data, creating features and using proxy variables for modelling. We will be using Geopy API, Openmeteo API for imputation and one hot encoding for feature processing





Model/Tool Selection:



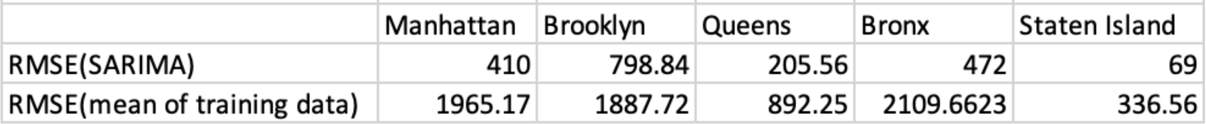
Objective	Model/Tool	Motivation	Target Variable	Comparison Model
Seasonal Variation Prediction	SARIMA	SARIMA can handle Seasonality, Non Stationary time series and Model Complexity (auto_arima)	Number of accidents in a month per Borough	Benchmark model that predicts mean of training data.
Risk Severity Classifier	Random Forests	Random Forests can handle Non Linearity, Large Feature Space	Low, Medium, High Risk	Decision Trees
Geospatial Analysis	Folium(Python Library)	Interactive, Ease of Use	Interactive Geospatial Map	N/A
Relevance Graph	Bar Graph	Easy to interpret and identify contributing factors	Bar Graph	N/A

Evaluation/Visualization Selection:

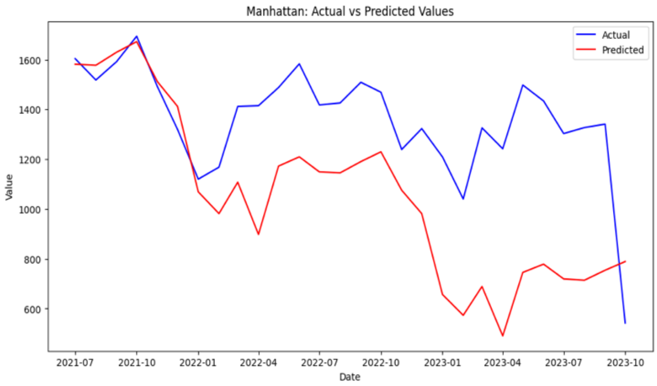
Objective	Evaluation	Motivation	Hyper-param Tuning	Visualization(Figures)
Seasonal Variation Prediction	RMSE	Provides a measure of the average magnitude of prediction errors & forecasting across the entire time series.	Seasonality order by ACF/PCF. ARIMA orders determined by auto_arima	Time Series Plot, RMSE comparison values with benchmark
Risk Severity Classifier	Confusion Matrix, ROC Curve	Detailed breakdown of model's performance, Easy to interpret multi-class evaluation	Grid Search to find the optimal trees and depth for best accuracy	ROC-AUC curve, Confusion Matrix
Geospatial Analysis	Graph	Interactive, Easy to Interpret	N/A	Geospatial graph
Relevance Graph	Graph	Interactive, Easy to Interpret	N/A	Bar graph

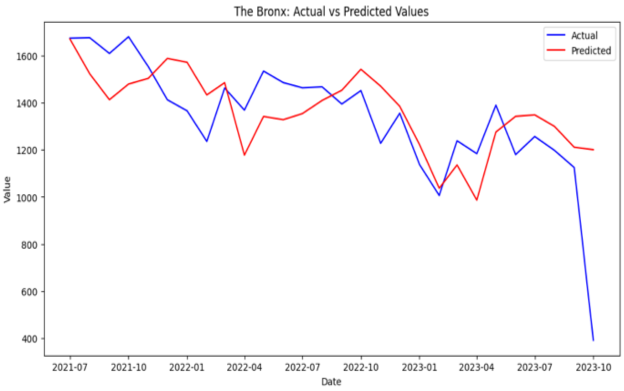
Results:

1. Seasonal Variation Prediction



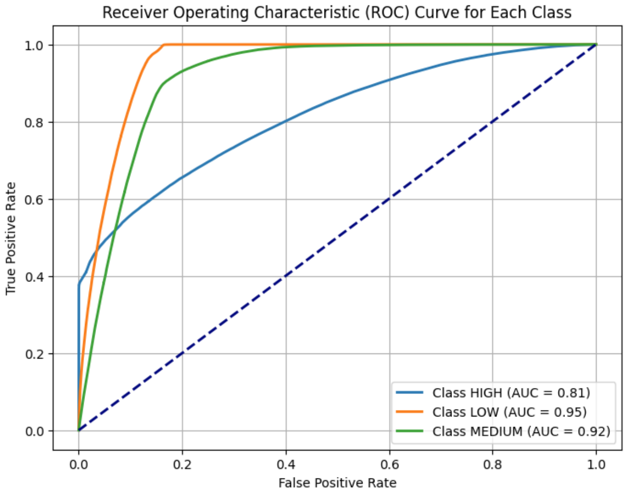
	Manhattan	Brooklyn	Queens	Bronx	Staten Island
RMSE(SARIMA)	410	798.84	205.56	472	69
RMSE(mean of training data)	1965.17	1887.72	892.25	2109.6623	336.56

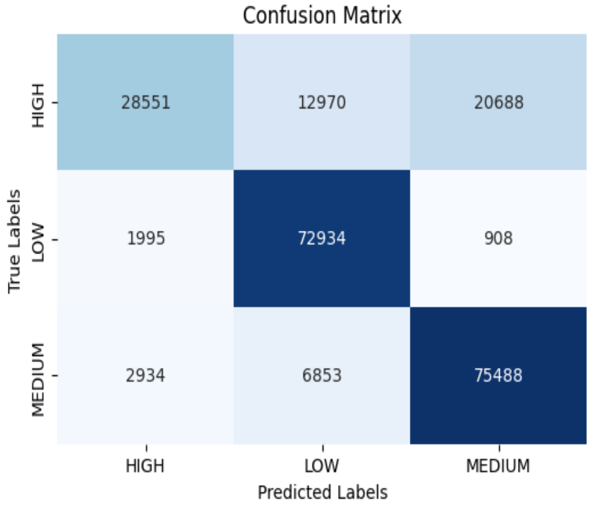




Analysis: We see that the time series models seem to closely forecast the patterns involved in the accidents. As future steps, we wish to improve accuracy by addressing outliers and anomalies in data as they negatively impact the model. We also wish to use exogenous variables in the model.

2. Risk Severity Classifier



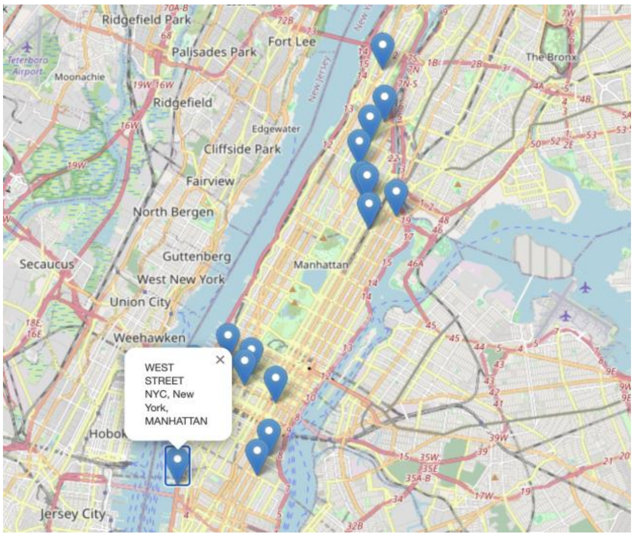

Confusion Matrix

True Labels \ Predicted Labels	HIGH	LOW	MEDIUM
HIGH	28551	12970	20688
LOW	1995	72934	908
MEDIUM	2934	6853	75488

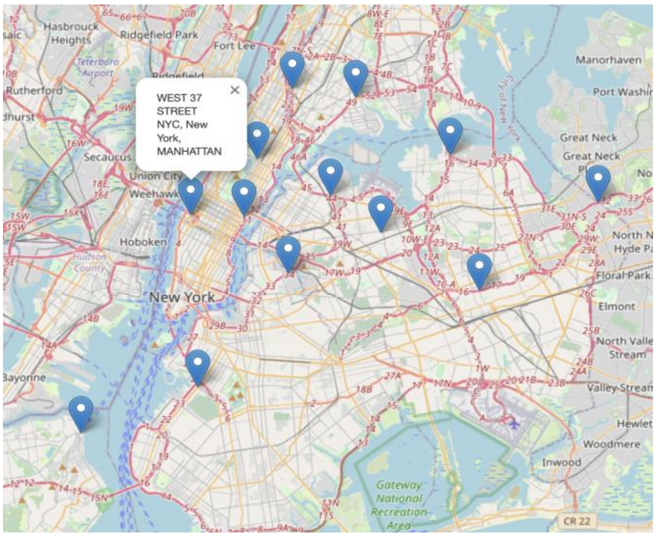
Analysis: The classifier does a fair job predicting the severity with an F1 score of 80%. The recall for high severity class was very low due to the highly imbalanced dataset. We wish to address this in the future by Data augmentation, Class weightage and Anomaly detection.

3. Geospatial Analysis

High Injury Severity in Manhattan

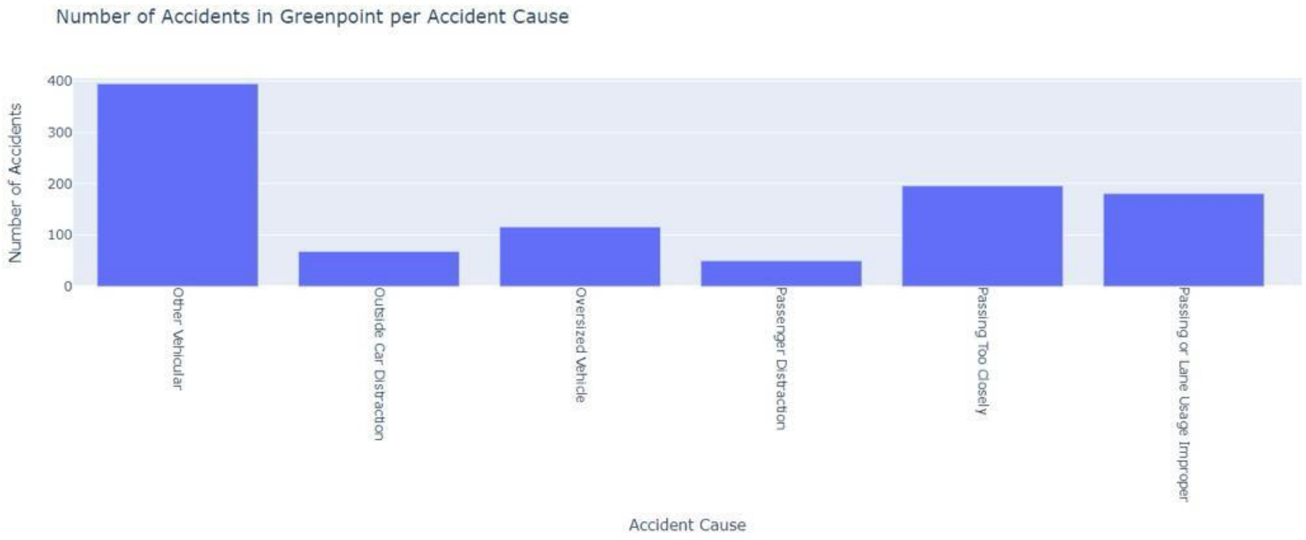


Road infrastructure issues in NYC



Note: These graphs can be interacted with in the python notebook

4. Relevance Graph



Note: Graphs can be generated for different neighbourhoods using the python notebook.

Assumptions:

- Latitude and Longitude coordinates accurately indicate accident locations
- Data from Geopy and OpenWeatherMap APIs for location-specific variables like weather conditions is justifiable due to their reliability
- Identifying accident hotspots based on geographic proximity is established, and a comprehensive dataset ensures ample data for pinpointing these hotspots.
- The available features are relevant to understanding accident causes as they are widely accepted factors influencing accidents
- Usage of historical data for predicting future trends is standard practice, as is the reliance on 'NUMBER OF PERSONS INJURED' to gauge accident severity.

Limitations:

- Under-reporting bias of accidents
- Weather data is w.r.t nearest crash hour time
- Contributing factors has missing values
- Variables influencing accidents may not all be there
- Changing Urban Landscapes affect traffic patterns and accidents

Team Evaluation:

1. Kaushik Mellacheruvu – 4
2. Deepjyot Singh Kapoor – 4
3. Mehran Ali Banka - 4