Motor Accident Analysis in New York City with Predictive Modelling

- Kaushik Mellacheruvu
- Deepjyot Singh Kapoor
- Mehran Ali Banka

Background, Problem and Data

- The problem at hand involves the high rate of traffic accidents in New York City, jeopardizing public safety and incurring significant economic costs. We aim to address this issue through geospatial analysis, time series modeling and predictive modeling to help with resource planning, resource allocation and policy making
- Target variables: Monthly accident rate/Borough, Risk Severity, Geospatial Maps
- Dataset:- Spatial Dimensions: (2034305, 30), Temporal Dimensions: (2012,2023), Source: Dataset
- Preprocessing: Geopy API to impute missing values, Openmeteo API to integrate weather data
- Preprocessing: One hot encoding for feature processing, Scipy.spatial.distance for closest location

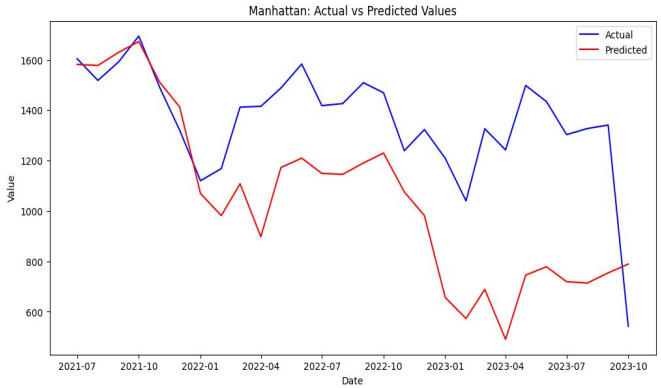
Feature	Type
Crash Date, Crash Time	Datetime
Zip Code, Number of persons injured, Number of persons killed, Number of pedestrians injured, Number of pedestrians killed, Number of cyclists killed, Number of cyclists injured, Number of motorists injured, Number of motorists killed, Collision Id	Numerical/Discrete
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

	BOROUGH WISE ACCIDENT COUNT PER MONTH (2012-2023)				
	Manhattan	Brooklyn	Queens	Bronx	Staten Island
Count	136	136	136	136	136
Mean	2909.9	4136.16	3786.94	2020.49	656.39
std	1169.9	1151.76	1259.87	622.79	258.04
min	511	1092	874	391	184
25%	1496.25	2970.75	2443	1592.25	423
50%	3423	4326.5	4065	1822	625.5
75%	3852.5	5037	4876.75	2619.5	911.25
max	4468	6147	5661	3231	1127

Model and Evaluation

Approach	Model/Tool	Reason	Evaluation	Reason
Geospatial Analysis	Folium	Interactive, Ease of Use	Visualization	N/A
Time Series	SARIMA, pyramid-arima, pcf/acf, Adfuller test	Seasonality, Stationary, Model Complexity, auto_arima	Visualization, RMSE	Identify seasonal pattern, comparison with mean of train data
Risk Severity Classifier	Random Forests	Non Linearity, Large Feature Space	Confusion Matrix	Multi Class Evaluation, model behaviour

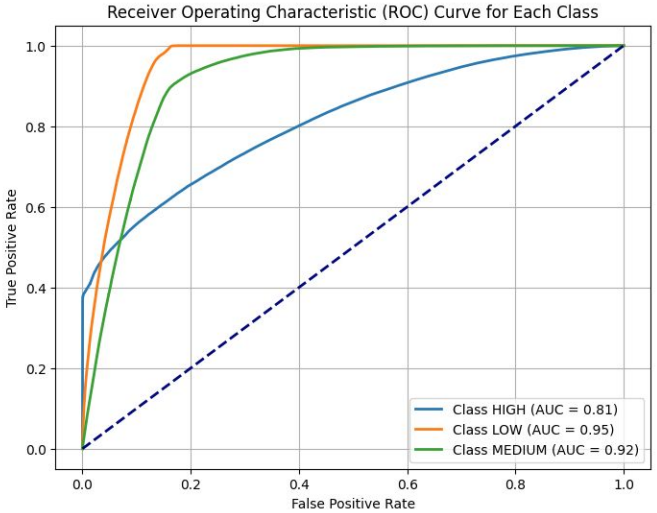
Time Series

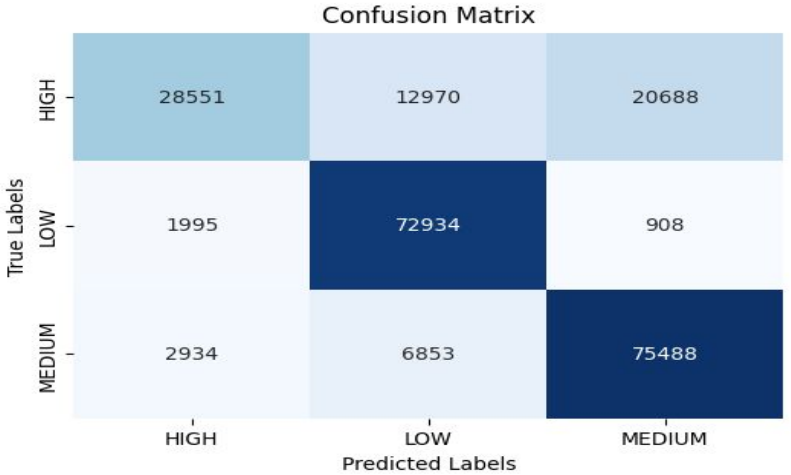


	Manhattan	Brooklyn	Queens	Bronx	Staten Island
RMSE(SARIMA)	410	798.84	205.56	472	69
RMSE(mean of training data)	1965.17	1887.72	892.25	2109.6623	336.56

Model and Evaluation

Risk Severity Classifier (Random Forests





Classification Report:

	precision	recall	f1-score	support
HIGH	0.85	0.46	0.60	62209
LOW	0.79	0.96	0.87	75837
MEDIUM	0.78	0.89	0.83	85275
accuracy			0.79	223321
macro avg	0.81	0.77	0.76	223321
weighted avg	0.80	0.79	0.78	223321

Assumptions, Limitations and Next Steps

Assumptions:

- Latitude and Longitude coordinates accuracy
- Data from Geopy and Openmeteo APIs for location-specific variables like weather conditions is justifiable due to their reliability
- Identifying accident hotspots based on geographic proximity is established. The dataset ensures ample data for pinpointing these hotspots.
- Usage of historical data for predicting future trends is standard practice, as is the reliance on 'NUMBER OF PERSONS INJURED/KILLED' to gauge accident severity
- Dropped columns such as ‘OFF STREET ADDRESS’, ‘CROSS STREET ADDRESS’ as we already have the address

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

Next Steps:

- Low level neighbourhood modelling
- Improve accuracy of models
- Include more features
- Improve imputations accuracy (Eg. Weather)
- Continuous modelling to adapt to changing patterns