1. Introduction

   National Highway Traffic Safety Administration suggests that the economical and societal harm from car accidents can cost up to $871 billion in a single year. Specifically, NYC has the most heavy traffic in the US. Therefore, this project aims to predict how severity of accidents in NYC can be reduced based on certain factors, which would help NYC citizens to better plan their travelling as well as NYC traffic administration.

2. Data

   The dataset used for this project is based on car accidents which have taken place within the New York City from 2012 to 2020. This data involves the severity of each car accidents along with the time and conditions under which each accident occurred. The data set used for this project can be found at https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95. The model aims to predict the severity of an accident, where, the variable of Severity Code was in the form of 1 (Property Damage Only) and 2 (Physical Injury) which were encoded to the form of 0 (Property Damage Only) and 1 (Physical Injury). Following that, 0 was assigned to the element of each variable which can be the least probable cause of severe accident whereas a high number represented adverse condition which can lead to a higher accident severity. Whereas, there were unique values for every variable which were either Other or Unknown, deleting those rows entirely would have led to a lot of loss of data which is not preferred.

   The Motor Vehicle Collisions crash table contains details on the crash event. Each row represents a crash event. The Motor Vehicle Collisions data tables contain information from all police reported motor vehicle collisions in NYC. The police report (MV104-AN) is required to be filled out for collisions where someone is injured or killed, or where there is at least $1000 worth of damage (https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/ny_overlay_mv-104an_rev05_2004.pdf). It should be noted that the data is preliminary and subject to change when the MV-104AN forms are amended based on revised crash details. For the most accurate, up to date statistics on traffic fatalities, please refer to the NYPD Motor Vehicle Collisions page (updated weekly) or Vision Zero View (updated monthly).

3. Methodology

   Considering that the data size and the some variable are categorical variables with the likes of data size, road condition and light condition being an above level 2 categorical variables whose values are limited and usually based on a particular finite group whose correlation might depict a different image then what it actually is. Generally, considering the effect of these variables in car accidents are important hence these variables were selected. A few pictorial depictions of the dataset were made in order to better understand the data.

4. Result

   The results of each of the three models had variations among them, one worked very well at predicting the positives accurately while the other predicted the negatives better. Logistic Regression

	Precision	Recall	f1-score
0	0.72	0.67	0.69
1	0.35	0.41	0.38
Accuracy	0.59		
Macro Avg	0.53	0.54	0.53
Weighted Avg	0.61	0.59	0.60
Log Loss	0.68		

5. Recommendation

After assessing the data and the output of the Machine Learning models, a few recommendations can be made for the stakeholders. The developmental body for NYC can assess how much of these accidents have occurred in a place where road or light conditions were not ideal for that specific area and could launch development projects for those areas where most severe accidents take place in order to minimize the effects of these two factors. Whereas, the car drivers could also use this data to assess when to take extra precautions on the road under the given circumstances of light condition, road condition and weather, in order to avoid a severe accident, if any.