

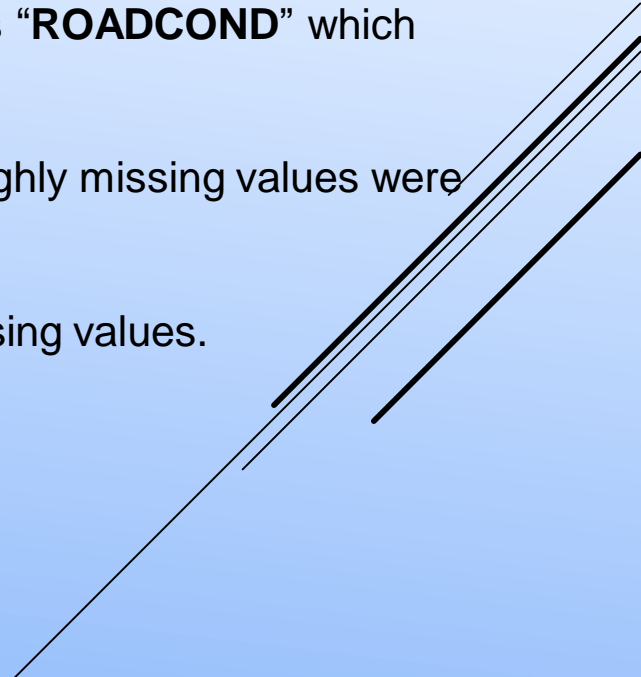
ACCIDENT SEVERITY ANALYSIS (IBM CAPSTONE)

By Iqbal Singh

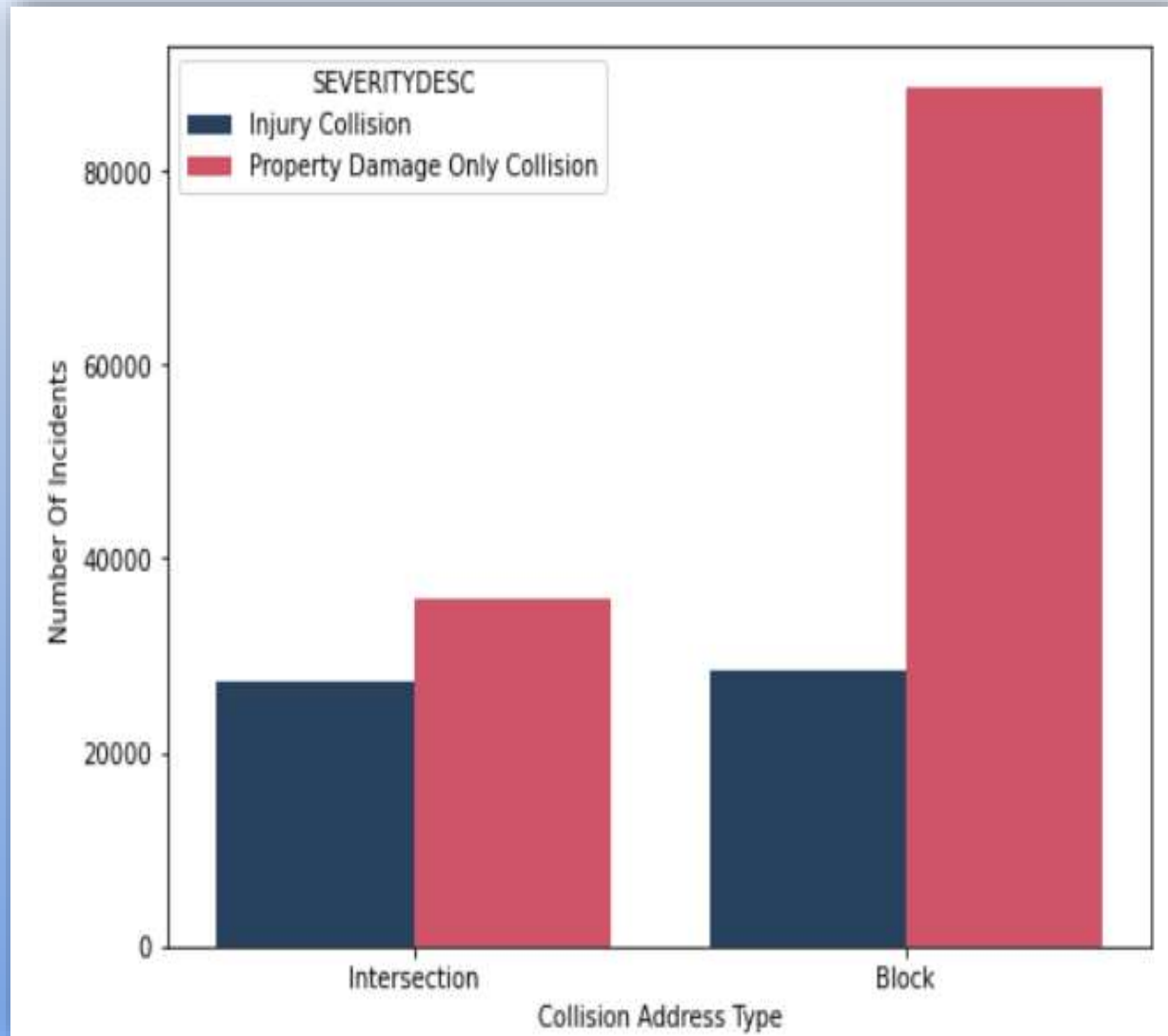
Predicting Accident Severity for Seattle City

- Motor vehicle accidents can happen quite suddenly, and all too often lead to fatal results. They can involve cars, trucks, motorcycles, and single or multiple vehicles. Some of these deaths occur at the scene of an accident, while others cause injuries that lead to fatalities after the victims receive medical care.
- In one report, the World Health Organization (WHO) articulates that the proportion of road accidents fatalities to total deaths in the world has grown by 2.2% from 1.255 million deaths in 2012. The accidents rate was increased by 0.3% from 2000 to 2012 despite the efforts served in terms of better road and laws enforcement. This is how WHO ranks road accidents as the 9th leading cause of deaths with 17.7 per 100,000 population in the world, which is sadly close to that of dangerous diseases like diabetes, diarrhea, HIV/AIDS, etc. In addition, approximately 30 to 50 million population are either injured or permanently disabled every year. Moreover, road accidents every year cause great financial havoc of \$518 billion and so costing countries 1% to 2% of their individual GDP alone.
- Now, wouldn't it be great if there is something in place that could warn you, given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be, so that you would drive more carefully or even change your travel if you are able to.

Data acquisition and cleaning

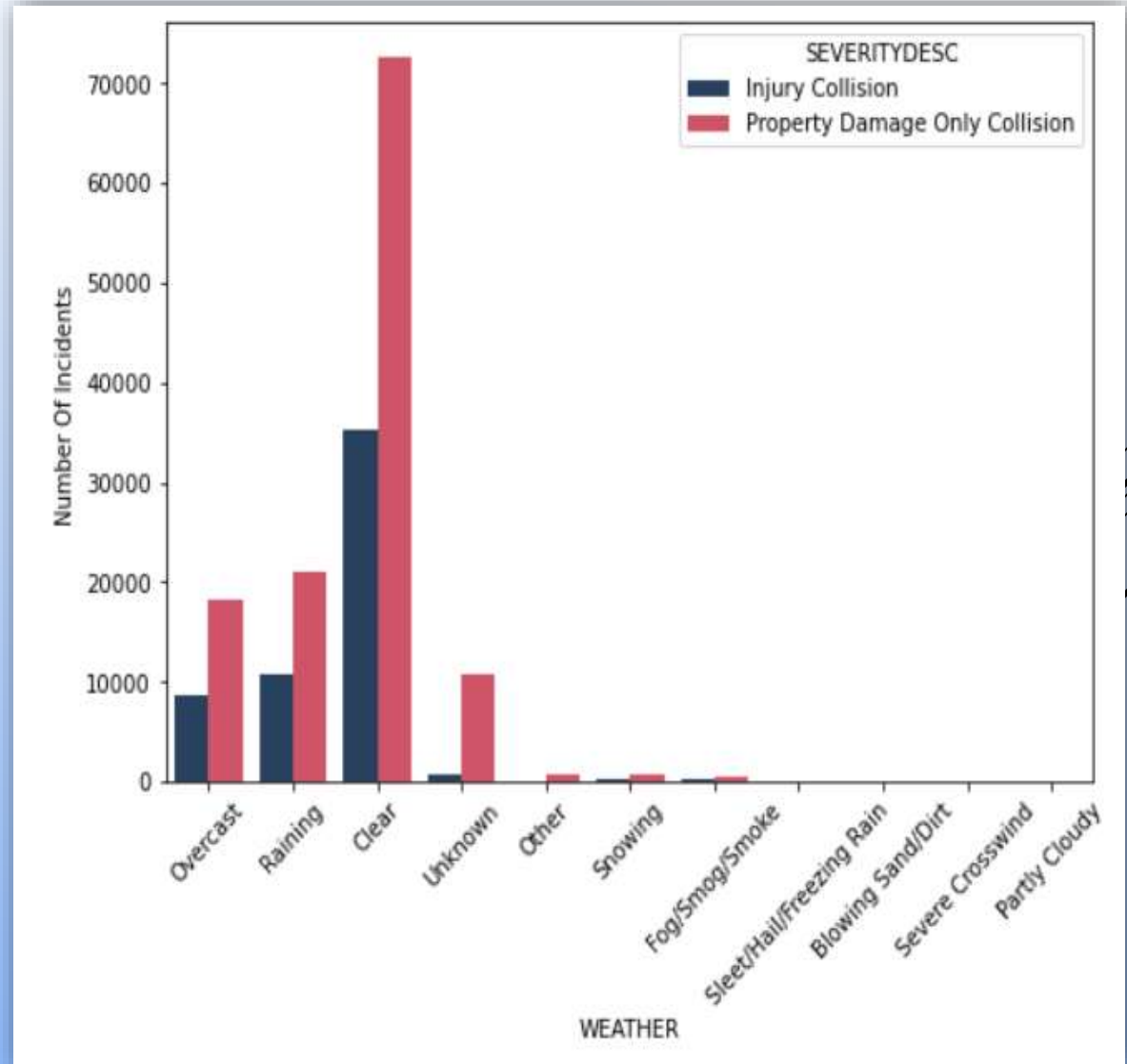
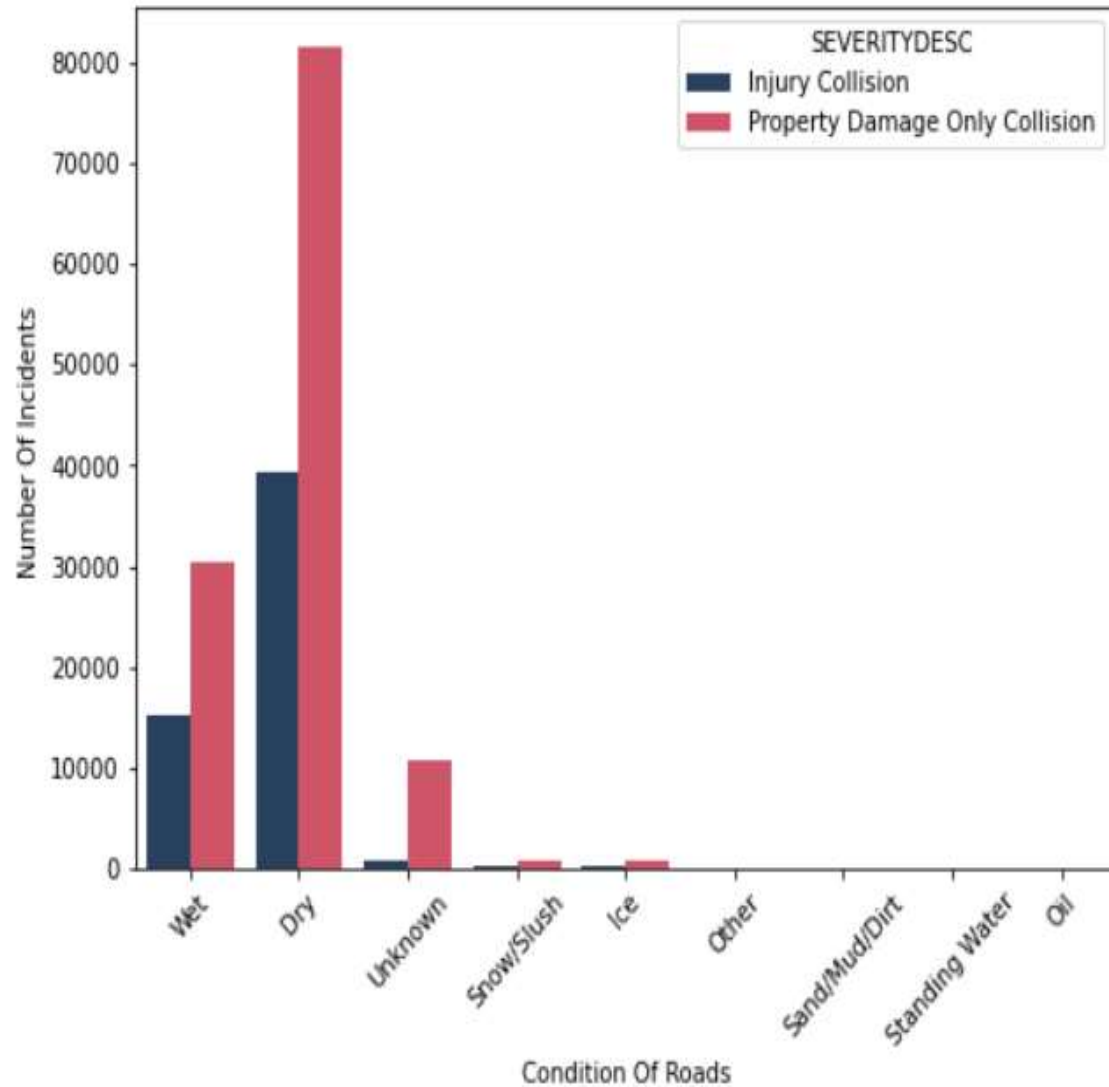
- For the purpose analysis, Data used in this capstone project was provided by IBM.
 - The collisions dataset includes statistics of accidents from 2004 to starting quarter of 2020. The dataset contains 194673 records in 38 columns. The first column in the dataset is “**SEVERITYCODE**” which is a labeled data and describes the severity of an accident, remaining columns have different attributes such as “**ROADCOND**” which describes about the condition of road at the time of accident etc.
 - The dataset has so many missing values in some of the attributes, all the features with highly missing values were dropped and then all the rows which can't be replaced were also dropped.
 - ‘**SPEEDING**’ was a crucial feature but needed to be dropped because it had 185340 missing values.
 - Some features were converted to correct data types.
 - After the preprocessing phase there were total 180067 rows and 22 columns.
- 
- A series of three parallel diagonal lines in the bottom right corner of the slide, extending from the middle of the right edge towards the bottom left.

Relationship between Collision address type and Severity

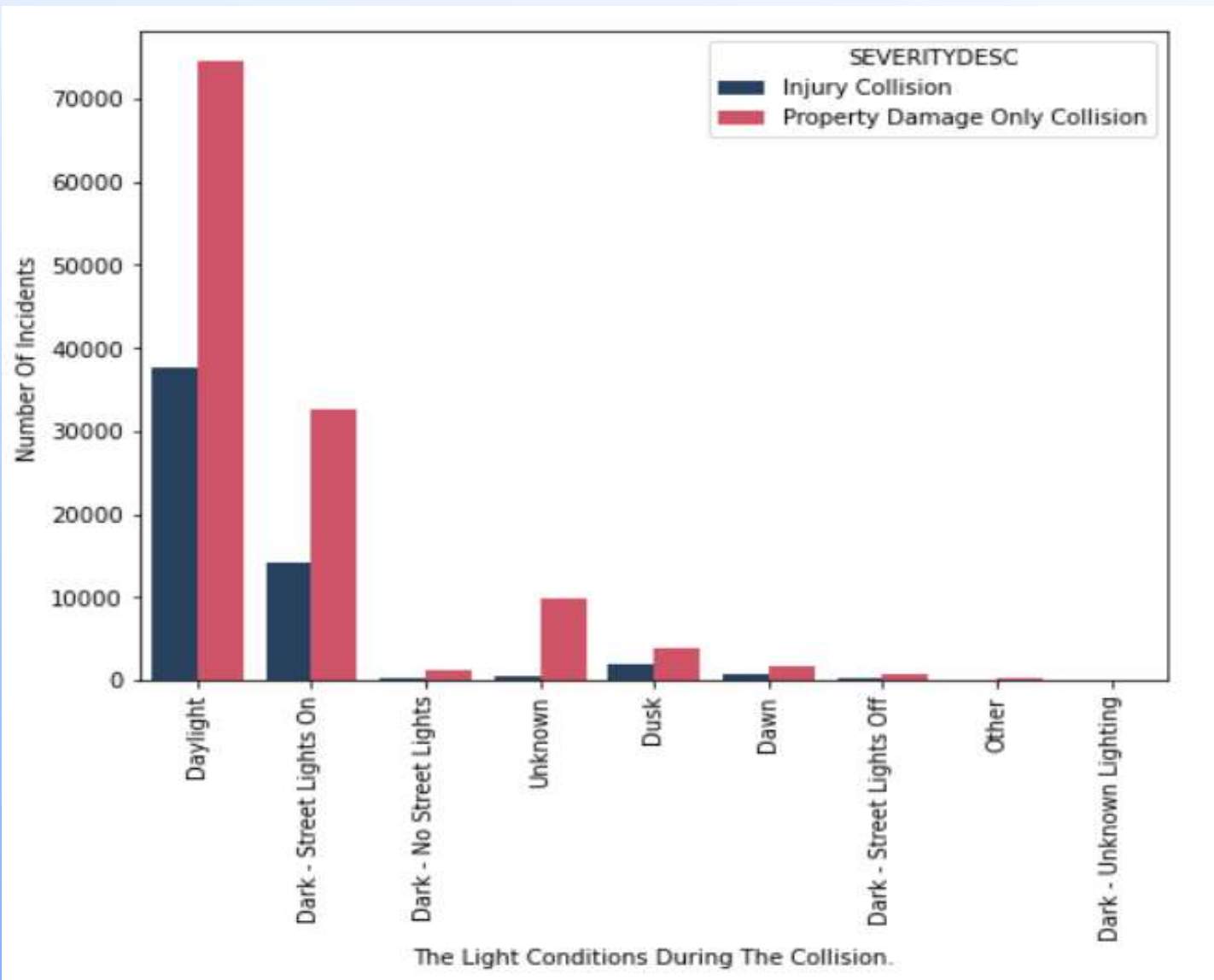


- From the figure we can say that Block leads to more accidents than intersections.
- Injury collision is almost the same in both the Collision address types i.e. around 25,000 cases.
- Property damage is more recorded in Block address with more than 80,000 cases recorded as compared to around 38,000 for intersection.

Affects of Weather and Condition of Roads on Severity

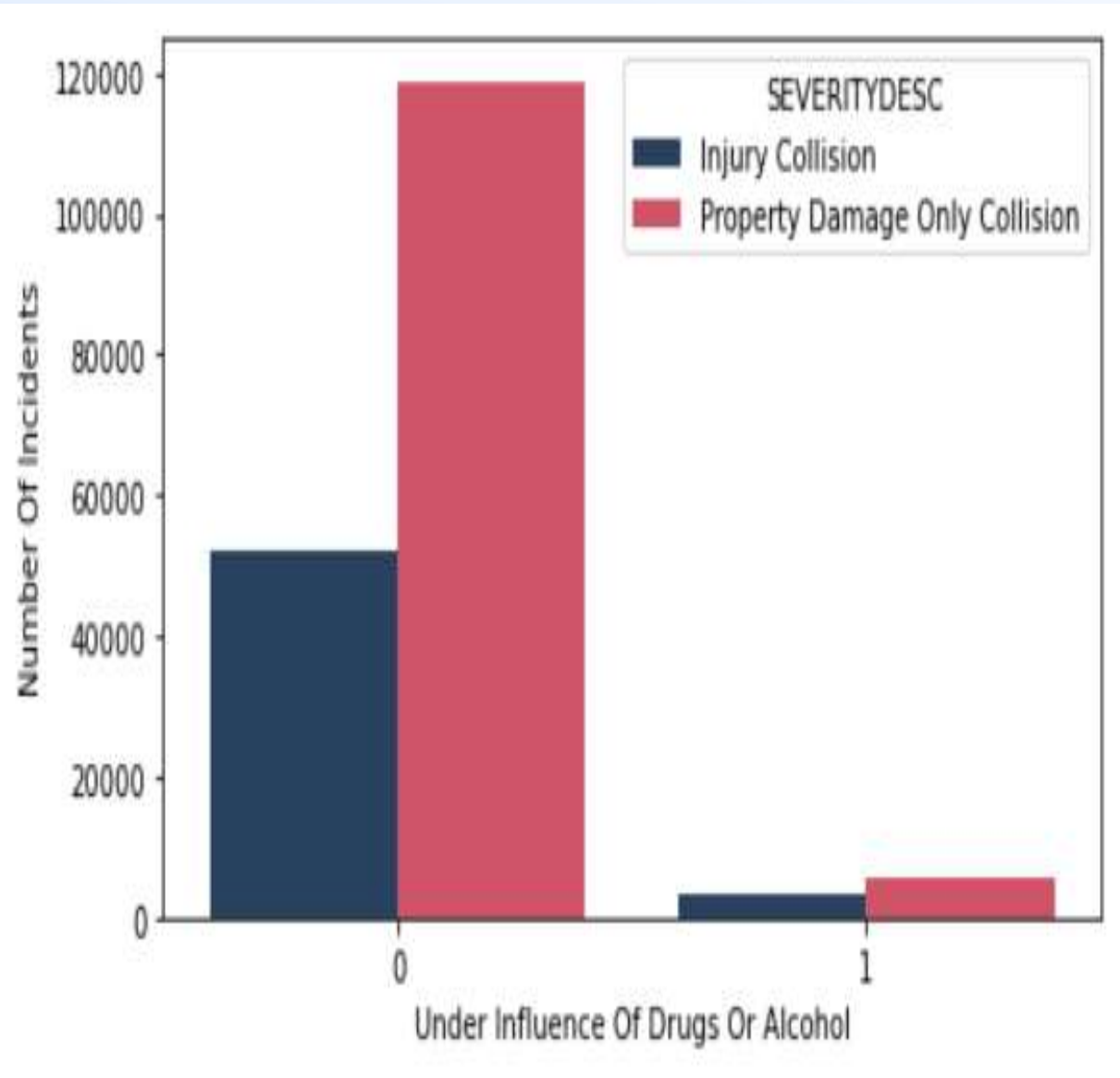


Relationship between Lighting Conditions and Severity



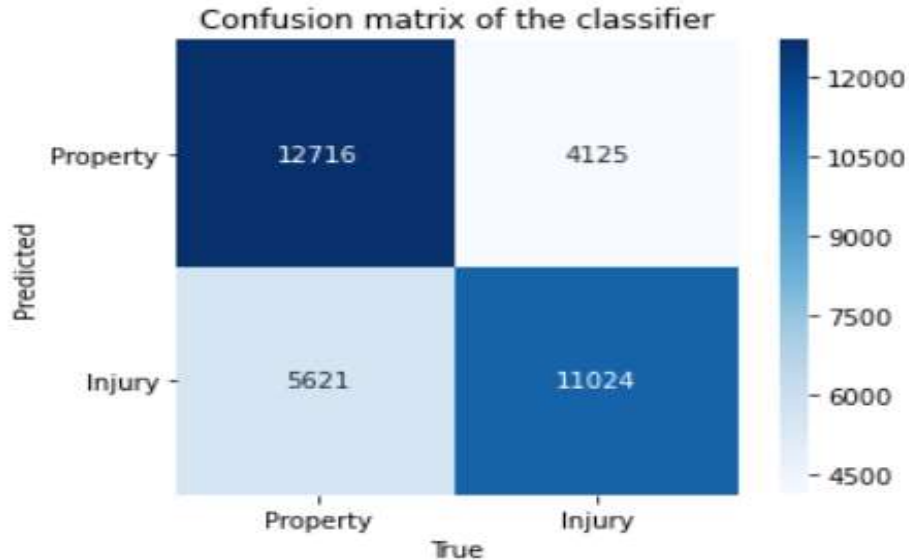
- From the figure we can observe that most accidents happen during Daylight or in Dark- street lights on, this indicates that bad light conditions do not lead to more accidents.

Relationship between Influence of Drugs and Severity



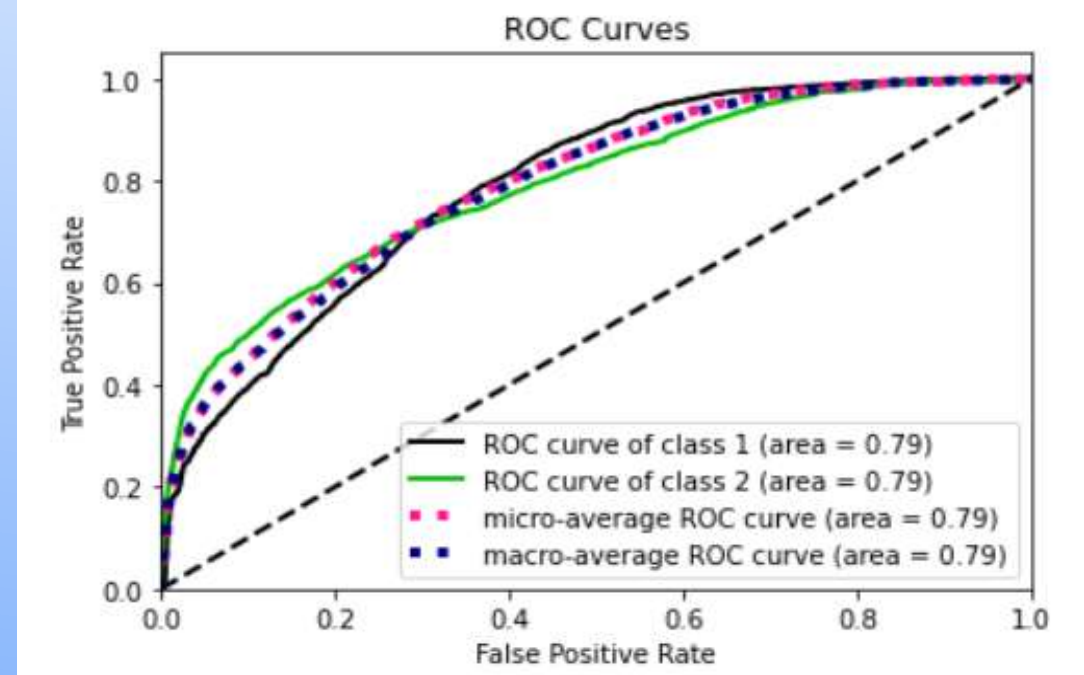
- From the figure we can observe that most accidents happen during Daylight or in Dark- street lights on, this indicates that bad light conditions do not lead to more accidents.

Performance of Logistic Regression

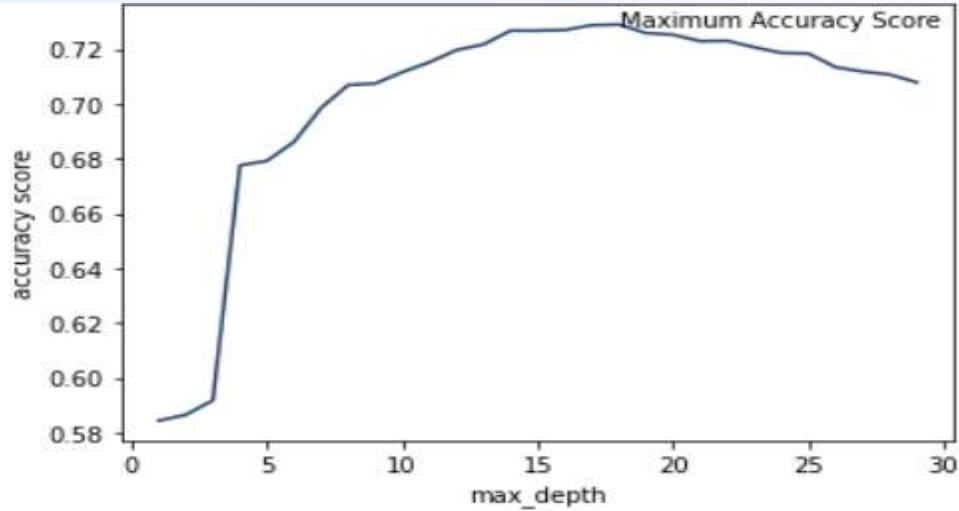


- Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).
- The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables.

- Logistic Regression's Accuracy:** 0.7089529952816103
- Logistic Regression's LogLoss:** 0.542224535278289
- Logistic Regression's F1-Score:** 0.7082943087673719
- Average_neg_mean_absolute_error:** -0.29749654430962985



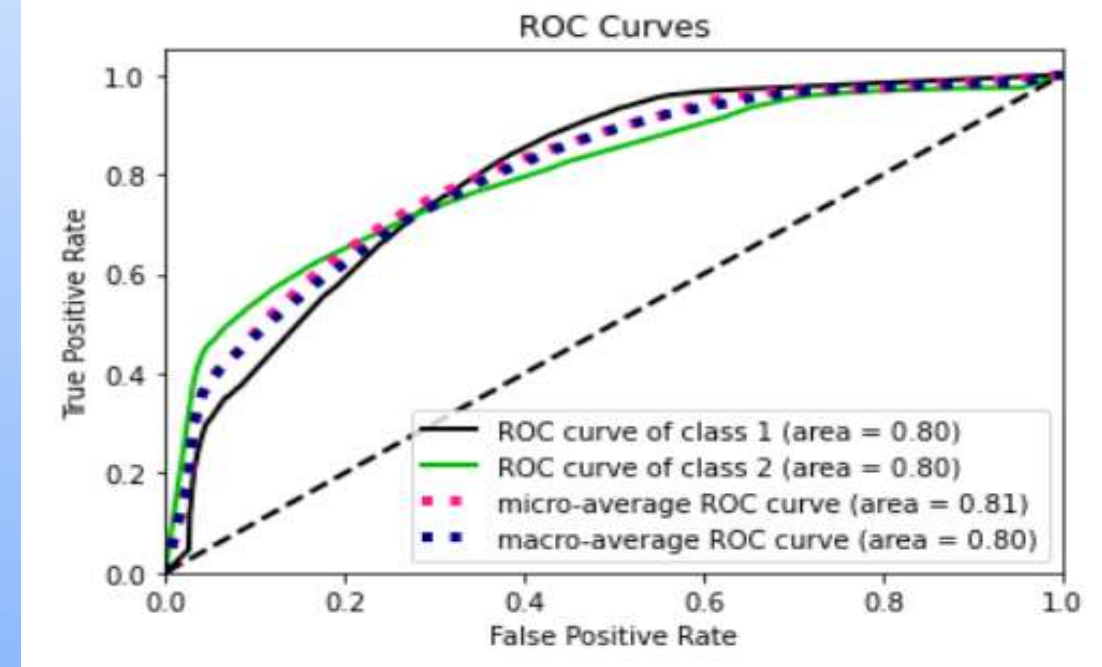
Performance of Logistic Regression



Max Depth of 18 shows better accuracy

- Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.
- I applied Decision tree algorithm to the training set and used criterion as “entropy” and max depth as 18. To find the max depth I applied for loop to get results of the depth that gets us the maximum accuracy.

- **DecisionTree's Accuracy:** 0.7281251866451651
- **DecisionTree's F1-Score:** 0.7249841884639812
- **Average_neg_mean_absolute_error:** -0.2799109199815697



Conclusion and future directions

- Build useful models to predict severity of an accident.
- Both models performed well with over 70% accuracy, but Decision Tree performed slightly better than Logistic Regression with 72.8% accuracy.
- Accuracy of the models can still be improved.
- Ideas include:
 - More Data can be recorded for Speeding as it is a very crucial Feature.

