CAR ACCIDENT SEVERITY MACHINE LEARNING MODELS

Sedat Kula

17 September 2020

PROBLEM STATEMENT

- According to the Global status report on road safety, conducted by World Health organization in 2013, based on the information gathered on road safety from 182 countries, accounting for almost 99% of the world's population, the total number of road traffic deaths are unacceptably high at 1.24 million per year.

- Every day, many accidents occur with various degree of severity based on different external and internal factors.

- It is important to define the factors leading the severity of accidents and to generate a model which predicts the severity type of accidents best

DATA

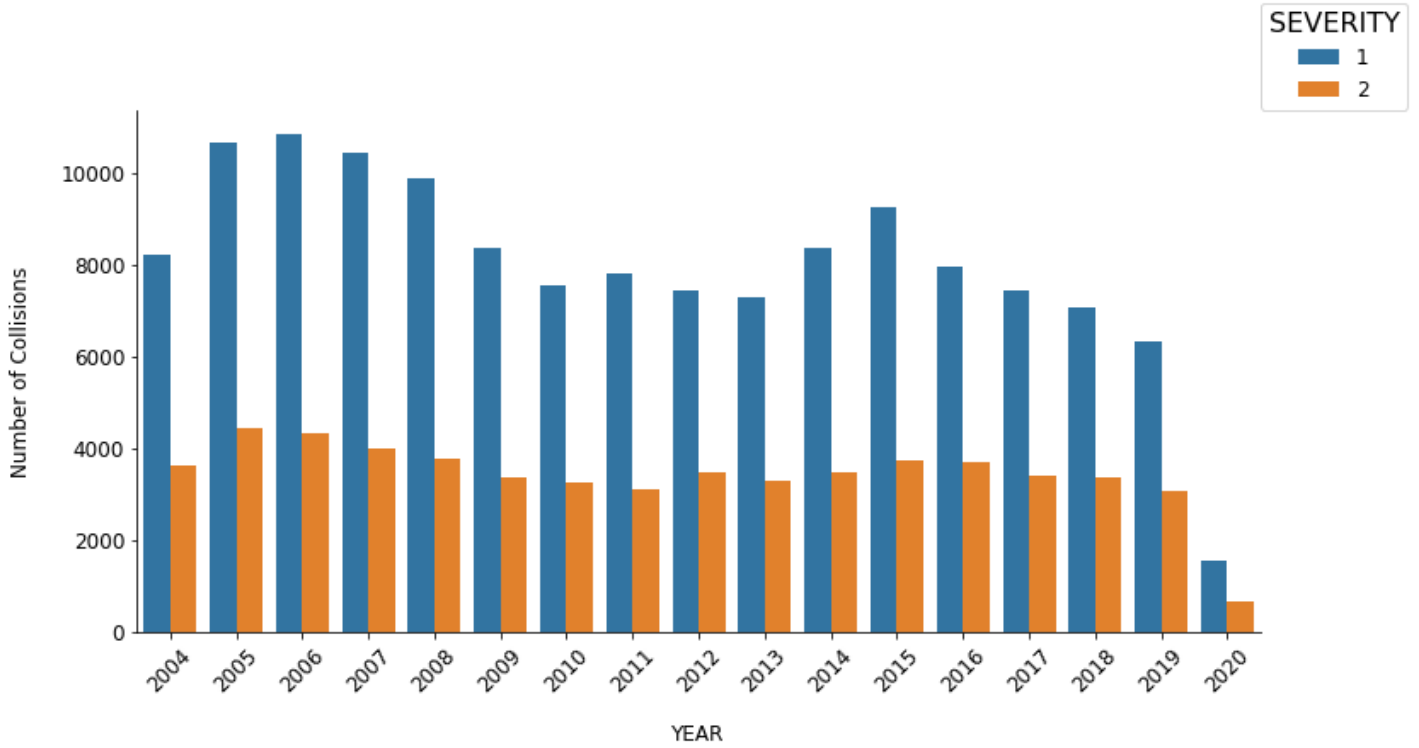
- The data for this project was provided by IBM's Applied Data Science Capstone Project.

- This data contains all types of collisions which occurred between the years of 2004 and 2020 in the city of Seattle from the Seattle Department of Transportation

- It has a total of 194673 records and 38 different attributes.

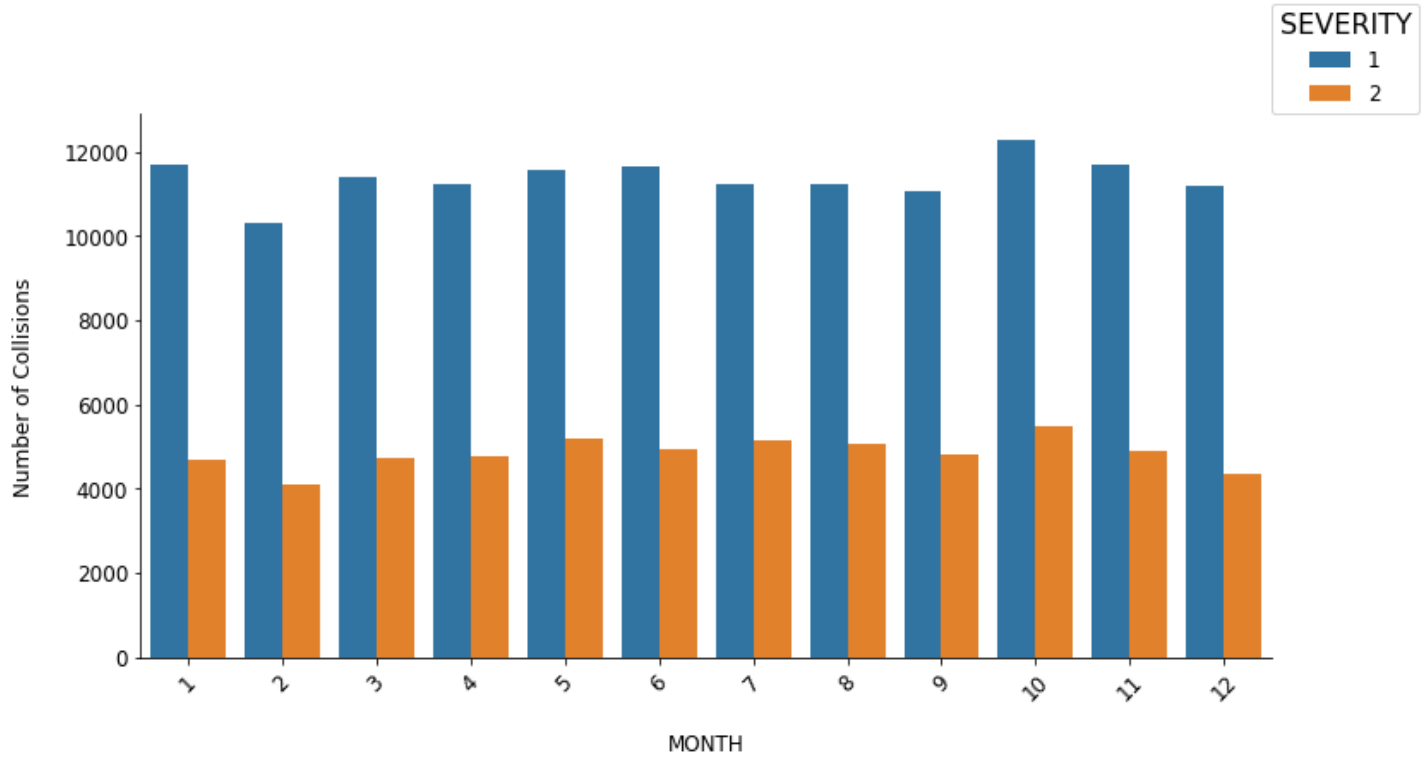
EXPLORATORY DATA ANALYSIS

Severity Levels by Year



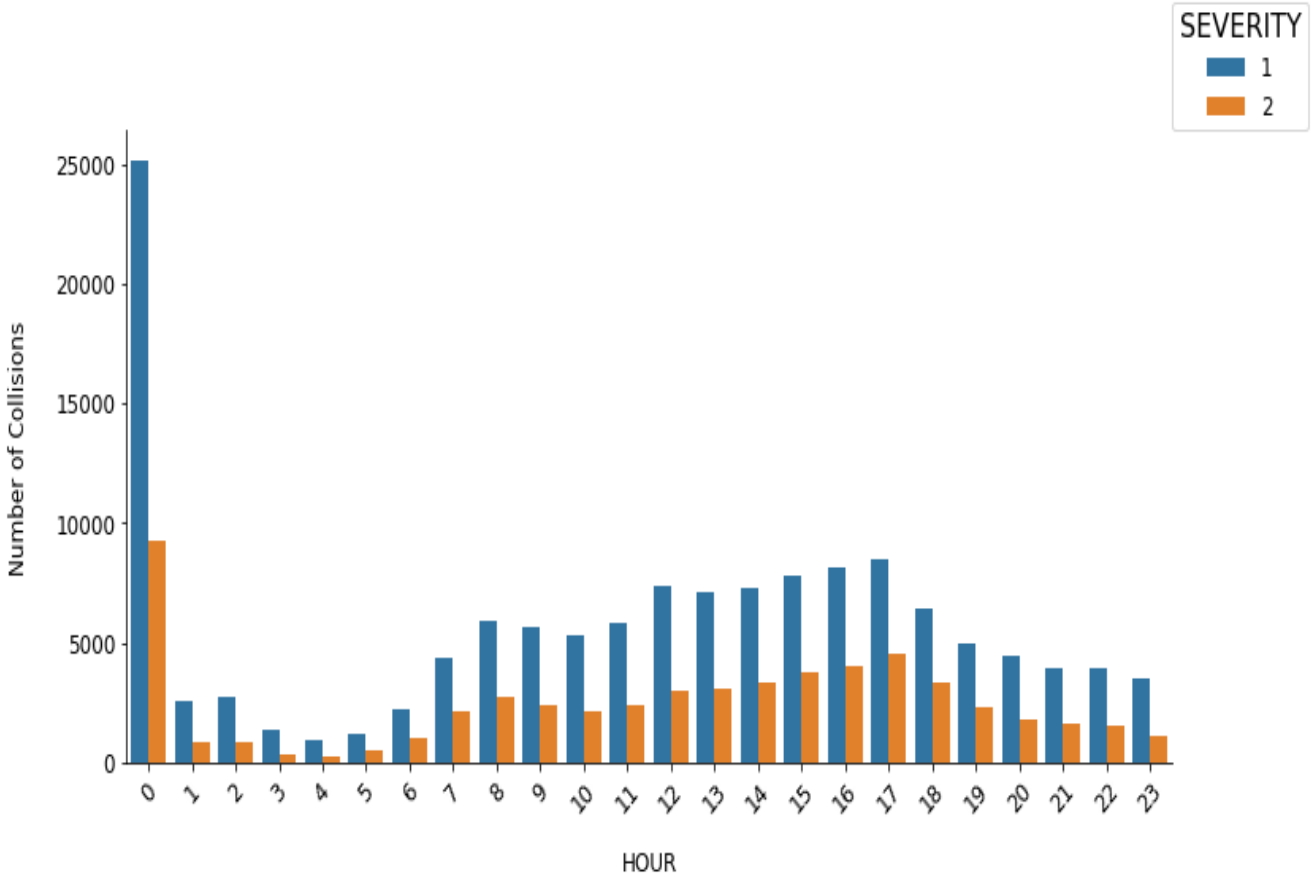
EXPLORATORY DATA ANALYSIS

Severity Levels by Month



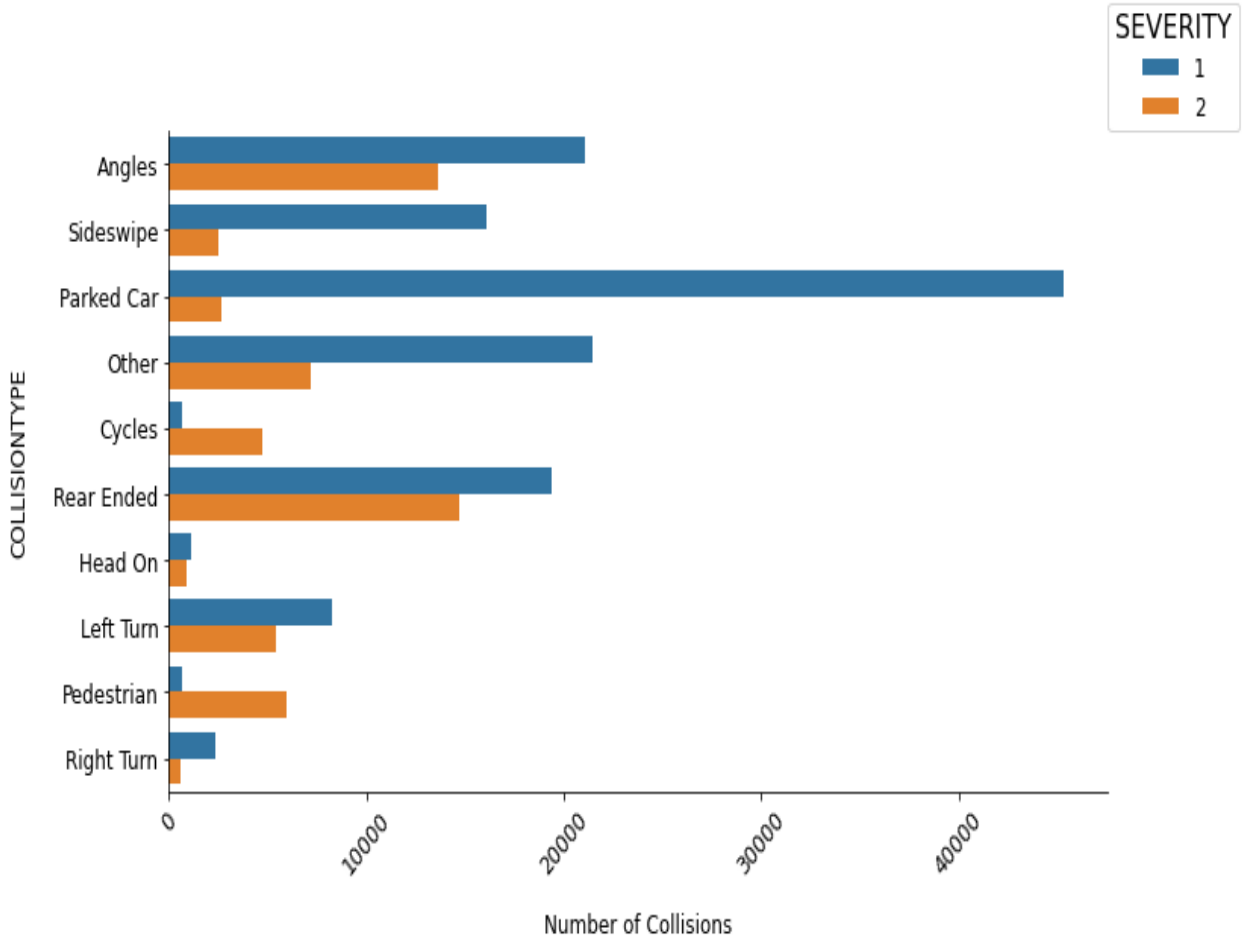
EXPLORATORY DATA ANALYSIS

Severity Levels by Hour



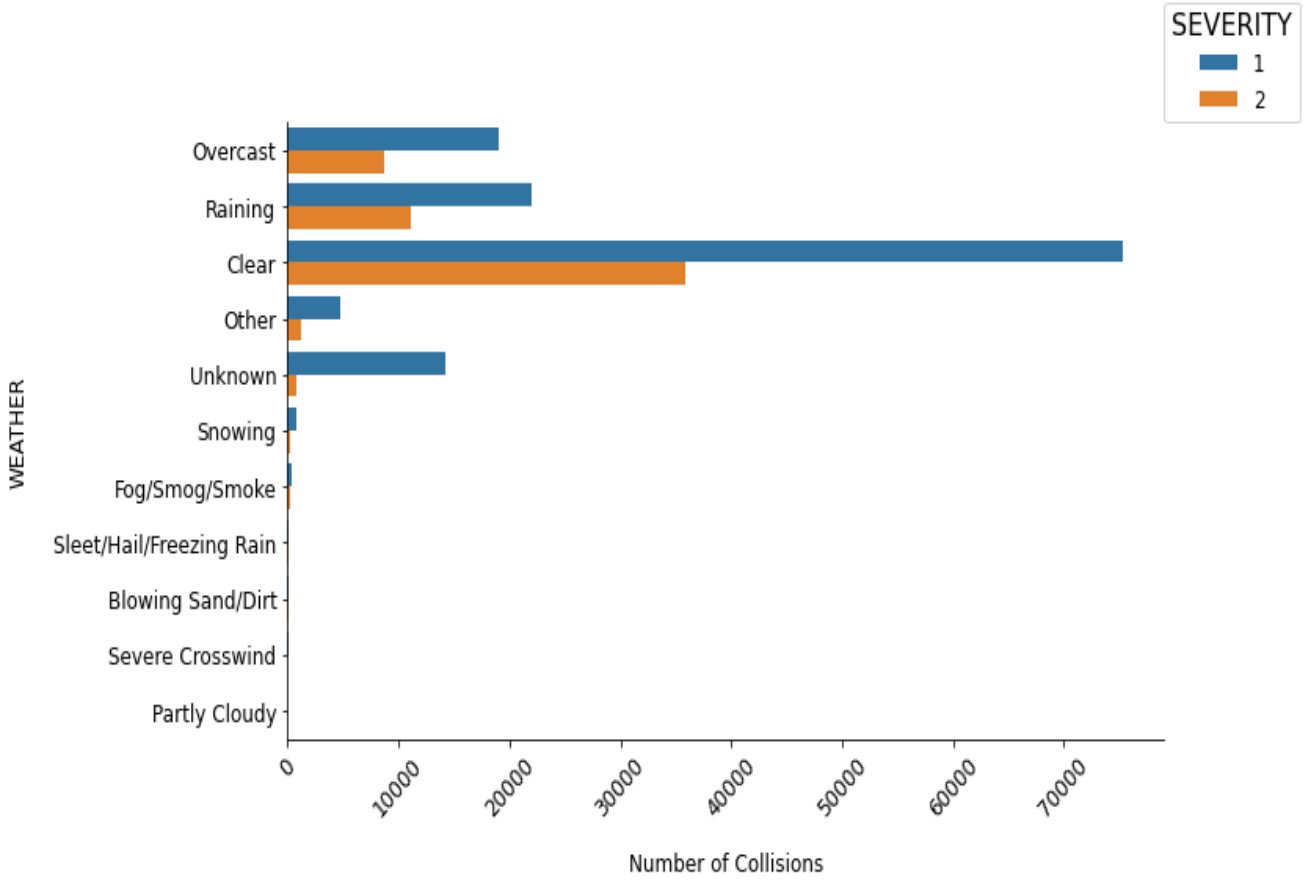
EXPLORATORY DATA ANALYSIS

Severity Levels by Collison Types



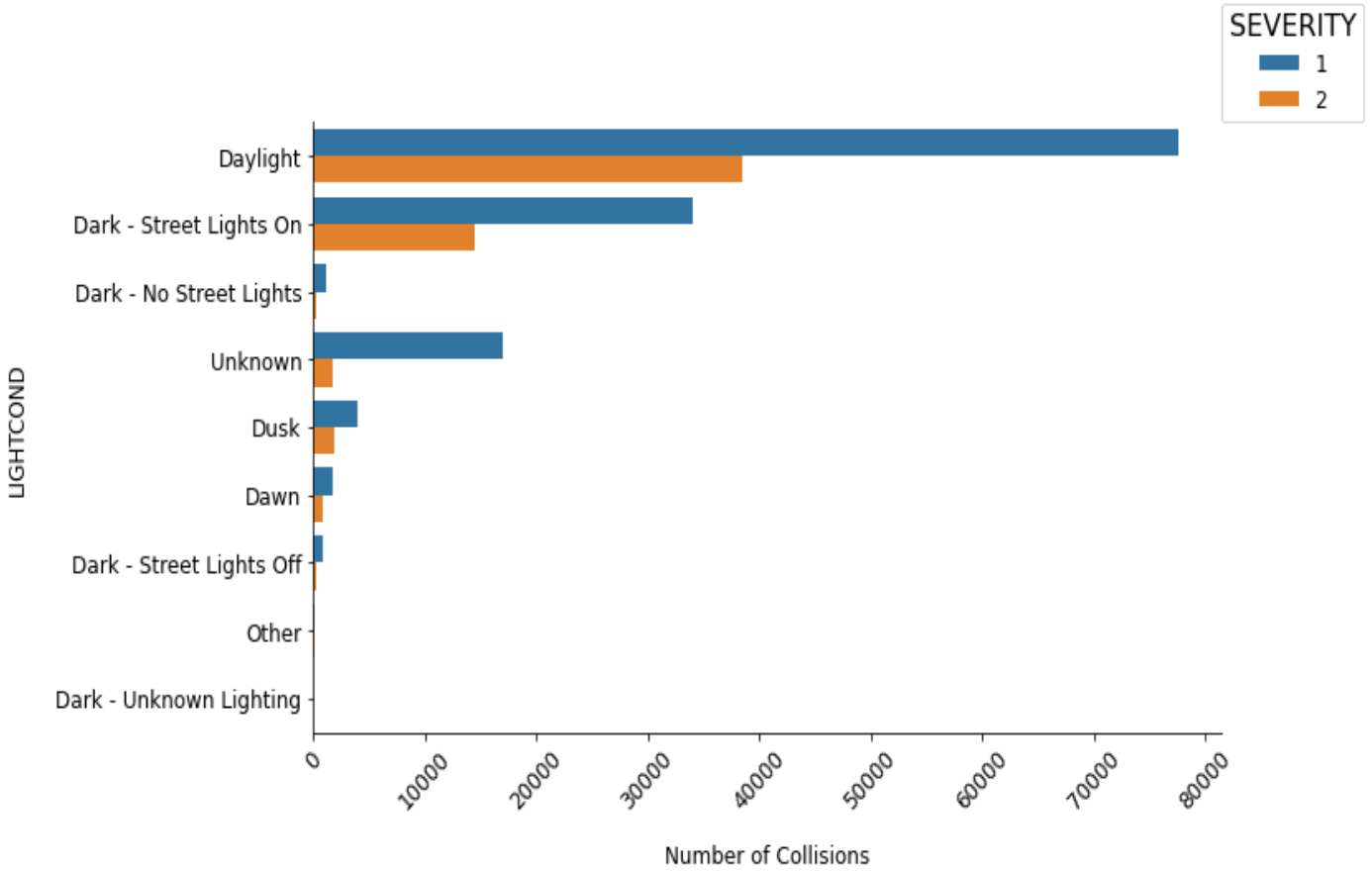
EXPLORATORY DATA ANALYSIS

Severity Levels by Weather Conditions



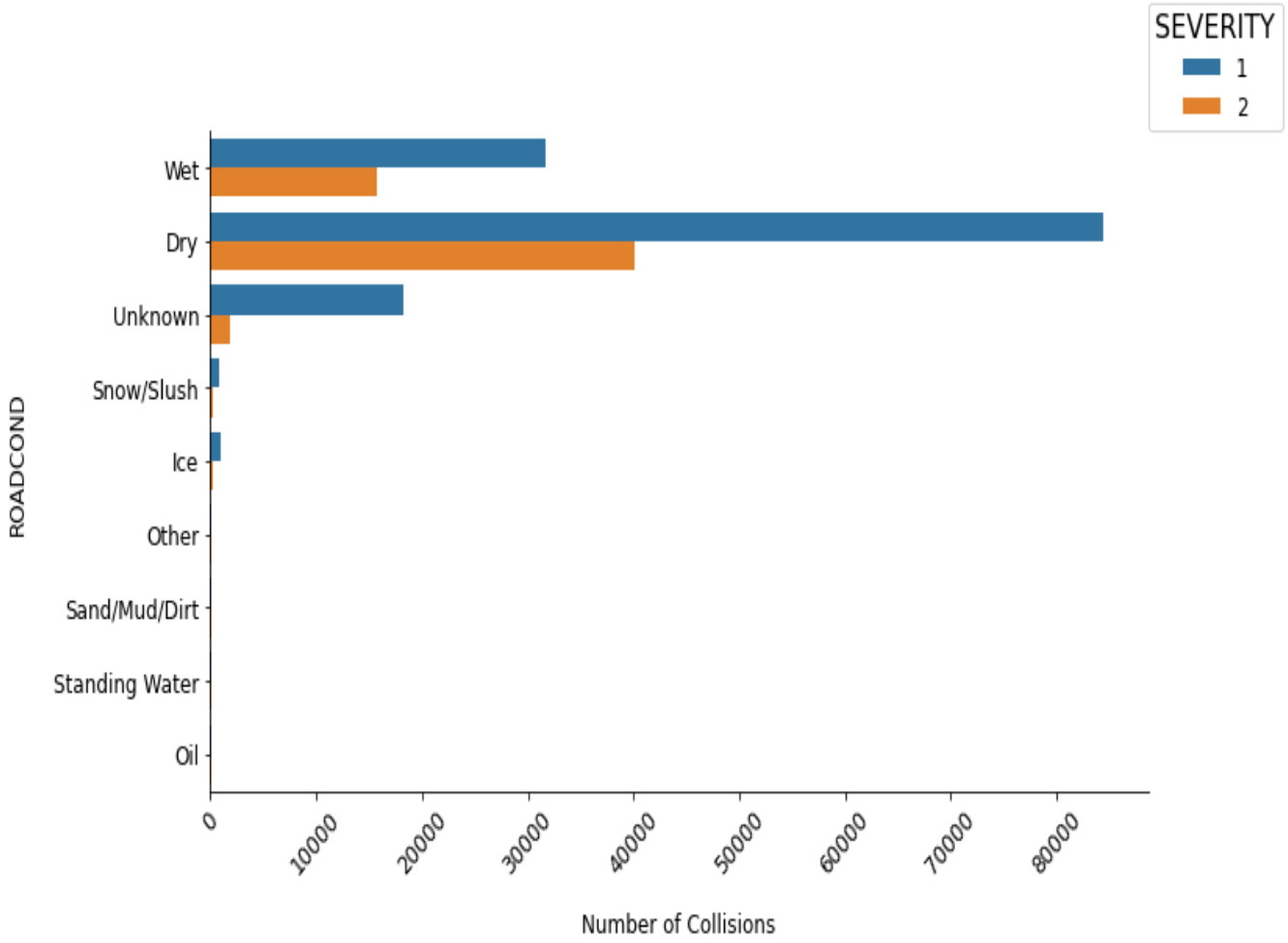
EXPLORATORY DATA ANALYSIS

Severity Levels by Light Conditions



EXPLORATORY DATA ANALYSIS

Severity Levels by Road Conditions



MODEL DEVELOPMENT AND EVALUATION

Five different classification models (supervised machine learning algorithms) were developed to classify the data.

1. K-Nearest Neighbour (KNN)
2. Decision Tree
3. Logistic Regression
4. Random Forests
5. Gradient Boosting (GBM)

	Model	F1 Score	Accuracy
0	Decision Tree	0.683414	0.748658
1	KNN	0.687520	0.729678
2	Logistic Regression	0.692575	0.747939
3	Gradient Boosting	0.688820	0.749120
4	Random Forests	0.683676	0.748761

DISCUSSION & CONCLUSION

- The selected features from the dataset have been used to classify the severity of the accidents.
- Five different machine learning algorithms were developed, namely, Decision Tree, KNN, Random Forests, Logistic Regression, and Gradient Boosting.
- Based on the accuracy scores, the Gradient Boosting algorithm offers the best performing model with an accuracy score of 0.7491.
- Exploratory data analysis indicates that the most vehicle accidents occur in a clear and dry road during the day. Therefore, it is important to emphasize that Seattle Transportation Agency should emphasis more on the drivers' training in terms of accidents.