Predicting Car Accident Severity

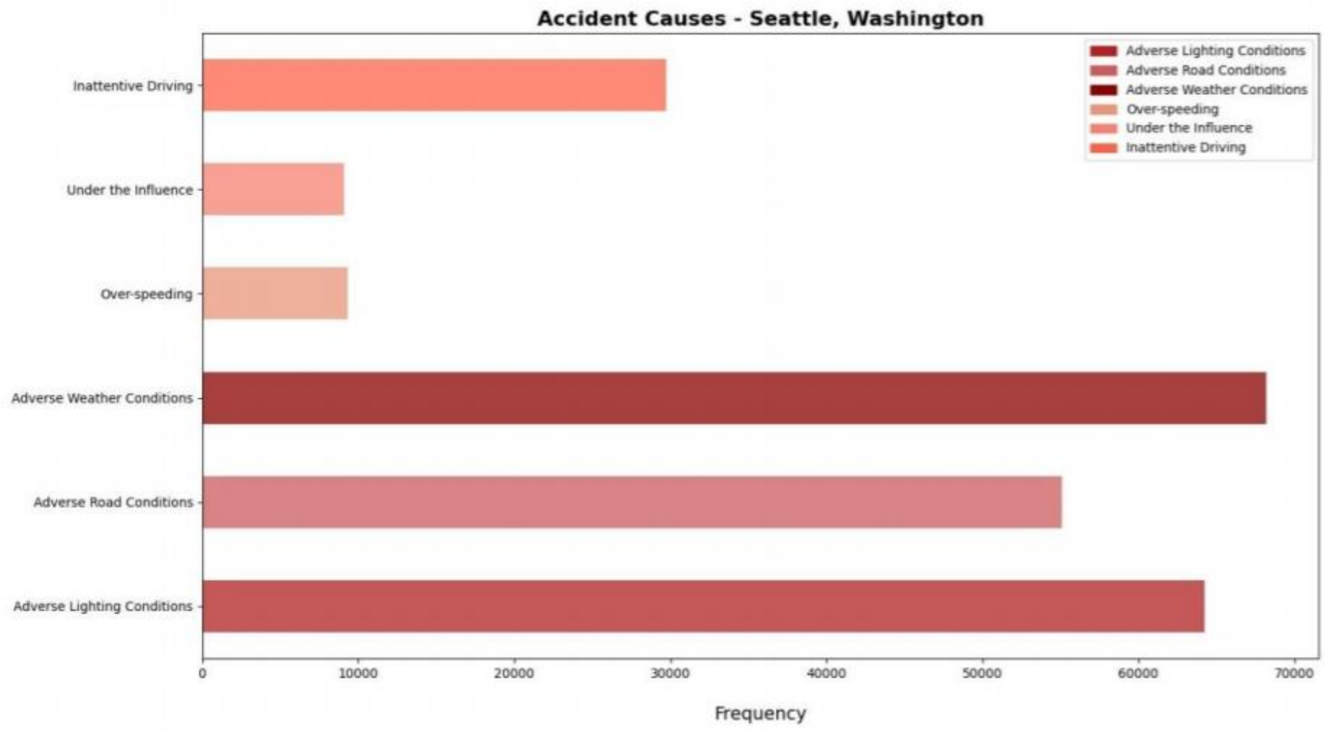
Introduction

- The world as a whole suffers due to car accidents, including the USA. National Highway Traffic Safety Administration of the USA suggests that the economical and societal harm from car accidents can cost up to $871 billion in a single year.
- According to 2017 WSDOT data, a car accident occurs every 4 minutes and a person dies due to a car crash every 20 hours in the state of Washington while Fatal crashes went from 508 in 2016 to 525 in 2017, resulting in the death of 555 people.
- The project aims to predict how severity of accidents can be reduced based on a few factors.
- The prediction in severity of accidents can be beneficial to the Public Development Authority of Seattle which works towards improving those road factors and the car drivers themselves who may take precaution to reduce the severity of accidents.

Data acquisition and cleaning

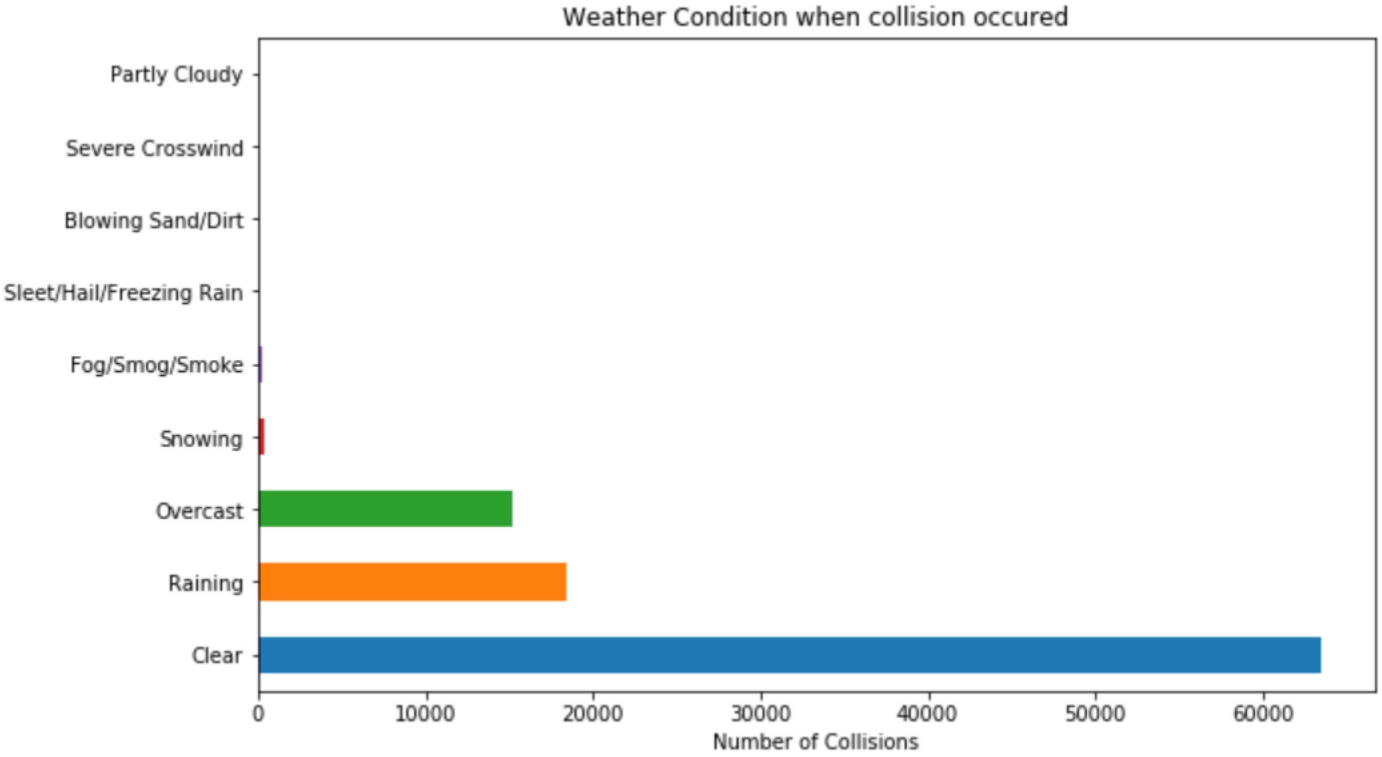
- Dataset has been collected and shared by the Seattle Police Department (Traffic Records).
- Dataset contains records of 200,000 accidents in the state of Seattle, from 2004 to the date it is issued.
- In total, raw dataset contains 194,673 rows and 37 features.
- Duplicate, highly similar or highly correlated features were dropped.
- Cleaned data contains 15 features.

Frequency of accidents which took place under adverse conditions



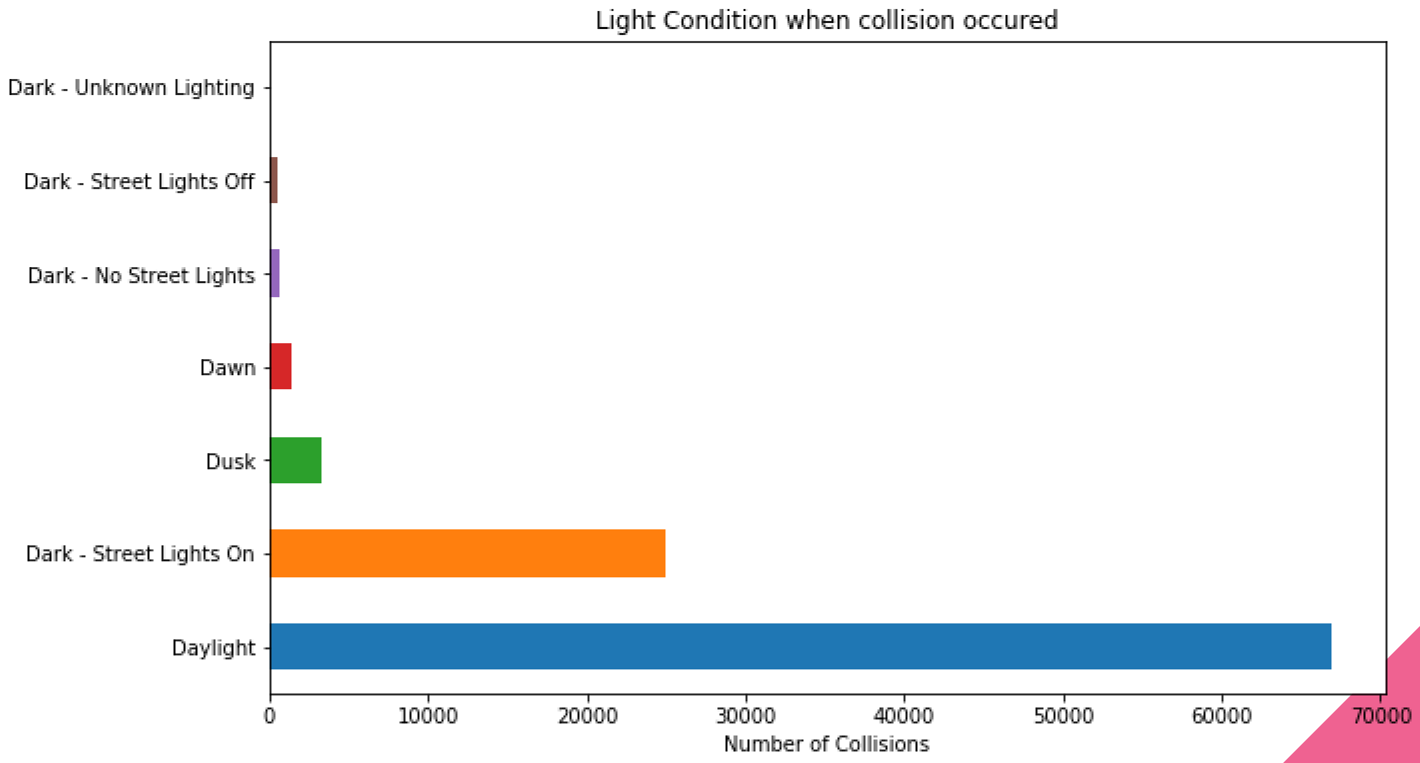
The factor which had the greatest number of accidents under adverse conditions was adverse weather conditions

Weather Condition when accident occurred



Almost 70% of the accidents have occurred when the weather is clear.

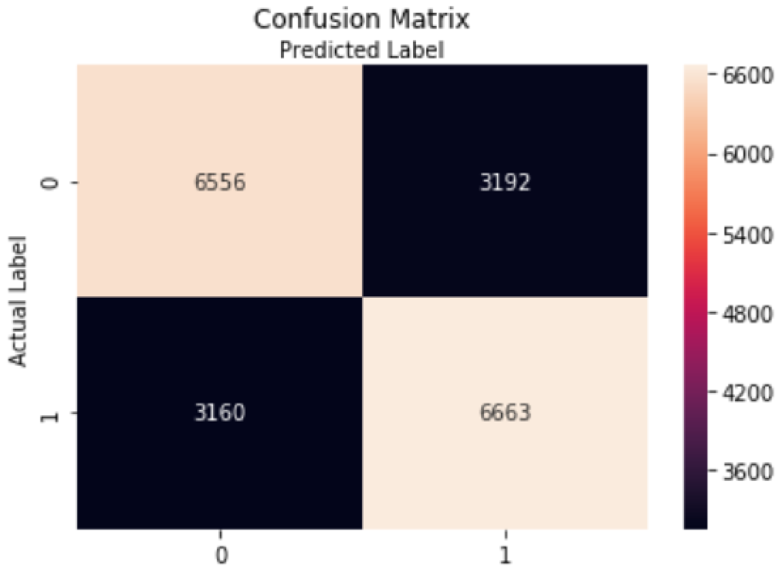
Light condition when accident occurred



Nearly 70% of the accidents have occurred during daylight

Classification Models Performance

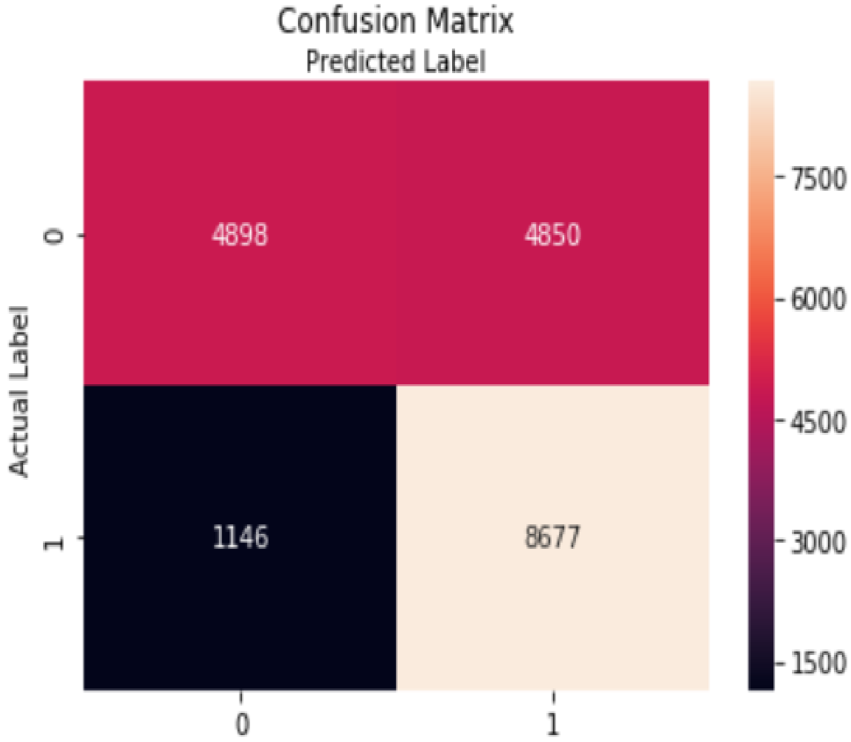
1. K Nearest Neighbors (KNN)



The best accuracy was obtained for 7 nearest neighbors with Euclidean being the distance metric .

The accuracy that was obtained using KNN classifier is 67%.

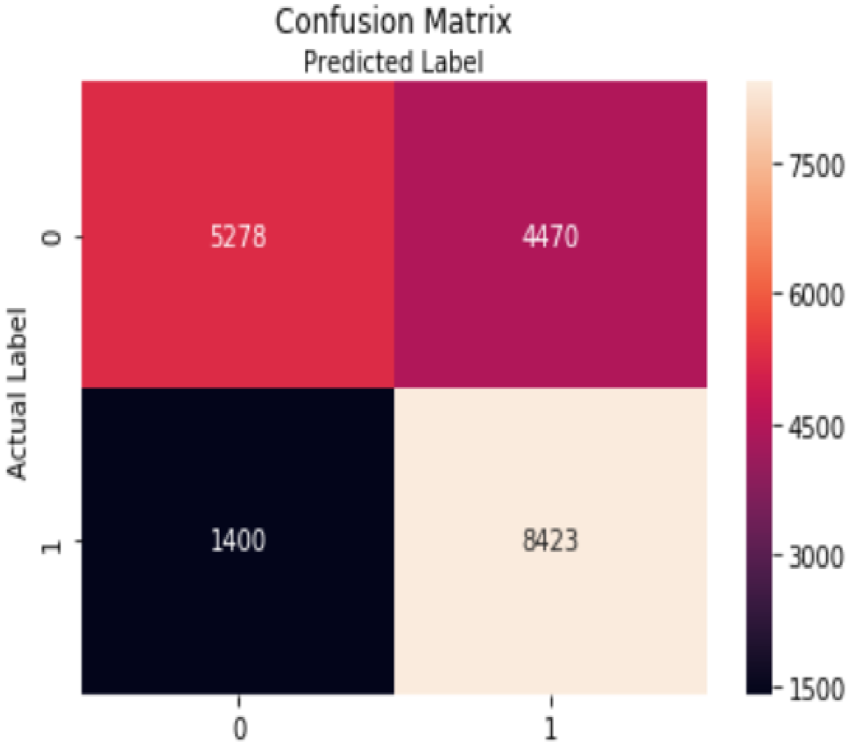
2. Decision Tree Classifier



Decision tree classifier using entropy had greater information gain; hence it was used for this classification problem.

The accuracy that was obtained using decision tree classifier is 69%.

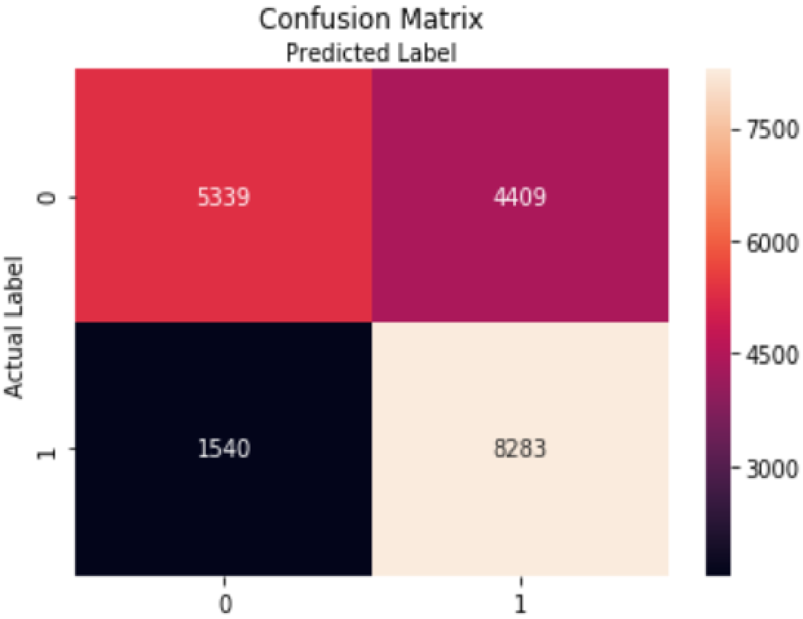
3. Support Vector Machine Classifier



Hyper parameter SVC was used to choose between Linear SVC and a Kernel SVC and the latter arrived on top with a greater accuracy . It used the ‘radial basis function’ kernel for performing the classification.

The accuracy obtained using SVM classifier is 70%.

4. Logistic Regression Classifier



	0	1
0	5339	4409
1	1540	8283

The accuracy that was obtained using logistic regression classifier is 69.5%.

Conclusion and future directions

- The accuracy of the classifiers is not great, highest being 70%. This usually means that the model is under fitted i.e. it needs to be trained on more data.
- Accuracy of the models has room for improvement.
- A better effort has to be made to collect data to reduce the number of missing values
- As mentioned above, the amount of data available to train the mentioned models is not sufficient and it does not seem to have enough data of all varieties. Hence, integrating Cross Validation methods with hyper parameter model would help in training and possibly increase the accuracy of every classification model.