**Coursera Capstone Project - Car Accident Severity**

**Xinkai Yip**

The objective of this project is to predict the severity of an accident based on the data that were collected from the Seattle City for collisions happened from 2004 to 2019.

1. **Introduction**
   1. **Problem background**

According to the World Health Organization there were approximately 1.35 million people killed on roadways around the world. Averagely there are almost 3,700 people killed globally in road traffic crashes. It was estimated to be the eighth leading cause of death globally for all age groups and the leading cause of death for children and young people 5-29 years of age. National Highway Traffic Safety Administration (NHTSA) data shows that in United States alone in year 2018, there were around 50,000 deaths from car accident and 4 million accidents that causes disabling and non-disabling injuries. The causes of accident ranges from weather condition, driver condition to road condition. By being able to predict the severity of accident based on several conditions, it could help insurance company to calculate the risk of the insurer based on the route they frequently used and the weather condition of their area. This machine learning project could also suggest navigation systems on which route is the safest to the user.

* 1. **Audience**

It was published in the Bulletin of the American Meteorological Society that there is a 34 percent increase in the risk of a fatal crash when precipitation is falling. Seattle is considered one of the states that has a lot of precipitation compared to other states in United States. This machine learning project is directed towards the people of Seattle as what conditions are prone to accidents and what could be avoided to reduce the risk of getting in an accident. It could also be used by the Department of Transportation to find ways on how to improve the traffics within Seattle and provide safeness to its people.

1. **Data Description**

The data set used met the following criteria:

1. The target or label columns should be accident " severity" in terms of human fatality, traffic delay, property damage, or any other type of accident bad impact.
2. The machine learning model should be able to predict accident "severity"
3. To build a good model, the dataset should be rich and contain many observations (rows) and various attributes (columns)
   1. **Data source**

The data was obtained from Seattle City Geo Data. The data records all types of collisions from year 2004 to Present. Collisions are displayed at the intersection or mid-block of a segment. The data contains speed, light condition, road condition etc. The purpose of the project is to use machine learning to predict the severity of an accident based on the datasets that was given. The dataset "Data-Collisions.csv" was provided by SDOT Traffic Management Division, Traffic Records Group.

These were some of the metadata’s that were provided by the SDOT GIS Analyst.

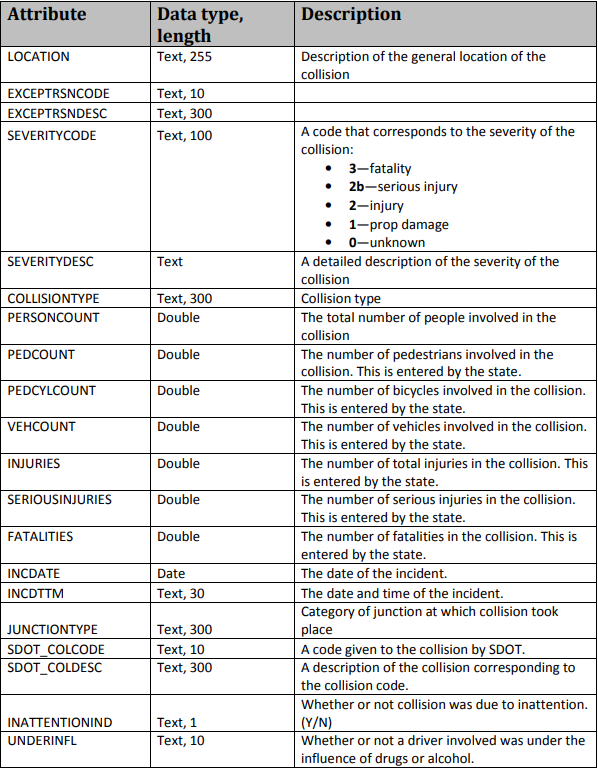


Figure : Metadata for Data-Collisions

1. **Methodology**
   1. **Data Preparation**

The dimension of the data was (194673, 38). Upon examining the features, it was found that there were a few features that does not contribute to the prediction of accident severity. For example, INCKEY, SEVERITYDESC, SDOT\_COLDESC which provides description to the respective incidents. These few features were drop instantly from data modeling consideration.

A screenshot of a computer

Description automatically generated

Figure : Description of the data

Further examination of the data shows that there were a few data sets that shows N, Y for No and Yes or unknown data. The speeding data consists of 185340 NaN and 9333 of Y data. There are 95% of unknown data in this feature. It was decided that the feature was dropped as it could not be assumed that the unknown data is N. There were a few other features that was dropped as well because of this consideration. They all yield more than 80% of unknown data which could cause bias in the prediction if it was assumed to be N. For features that has distinct N and Y data, it was transformed into 0 and 1.

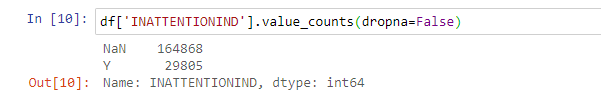
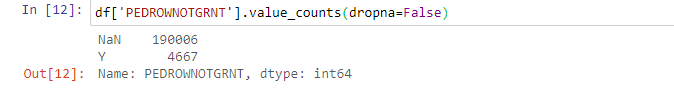
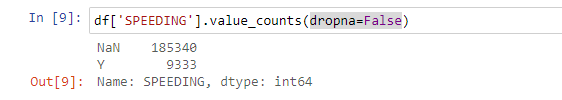


Figure : Data Analysis for Speeding, Pedrownotgrnt, and InattentionInd

The date data was examined, and it was found that it was not suitable to be considered in the data modeling. The date data was separated into date only and time only data. Date data was transformed into 0 – 6 where 0 is Monday and 6 is Sunday.

According to the bar graph and the data analysis, there is a 70% probability that the accident will have a severity code of 1 no matter the day and a 30% probability of severity code 2. This is an imbalance data which is why the date is dropped from consideration.

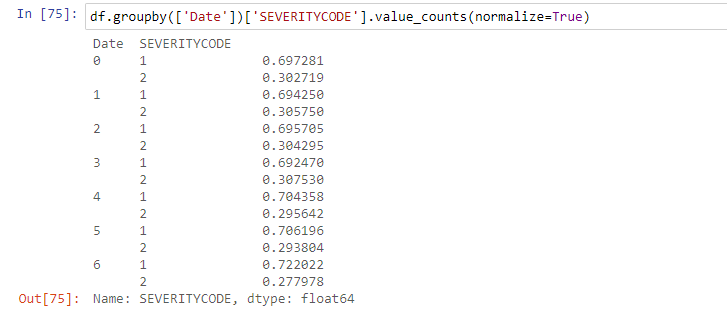
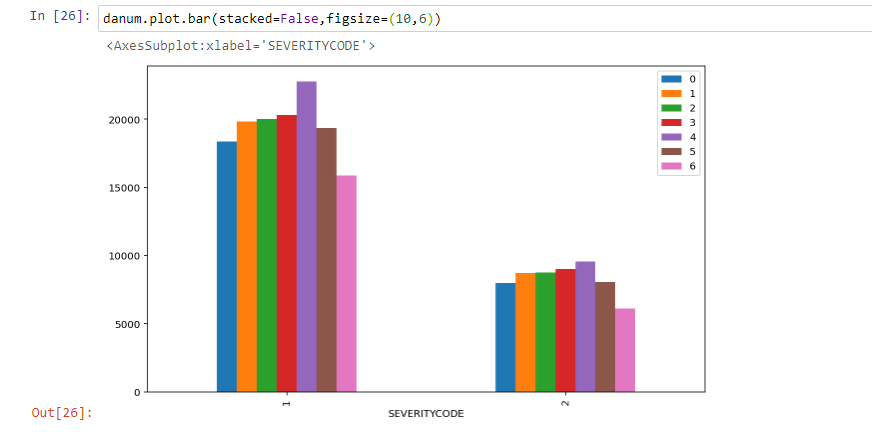


Figure : Date Data

Next is the Time data. Through the bar graph, the probability of accident from time 0 to time 23 with severity code 1 ranges from 65% to 77%. This data would be put into consideration for data modeling.

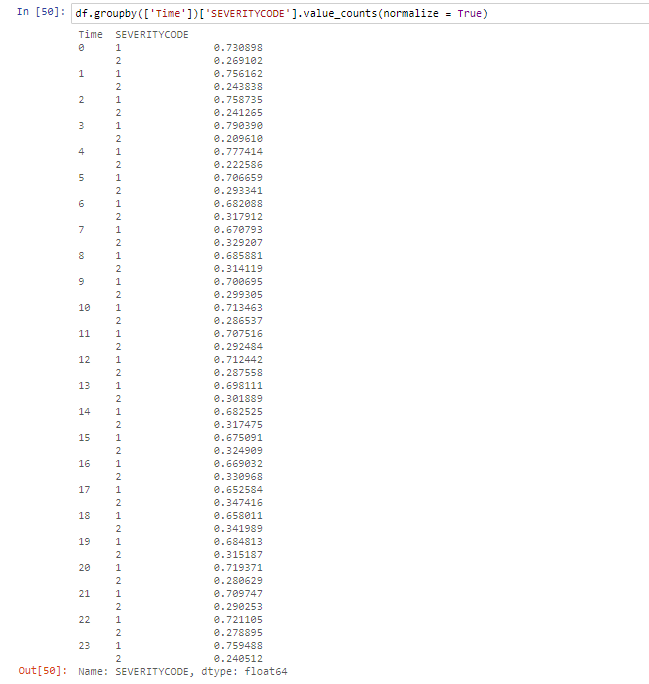
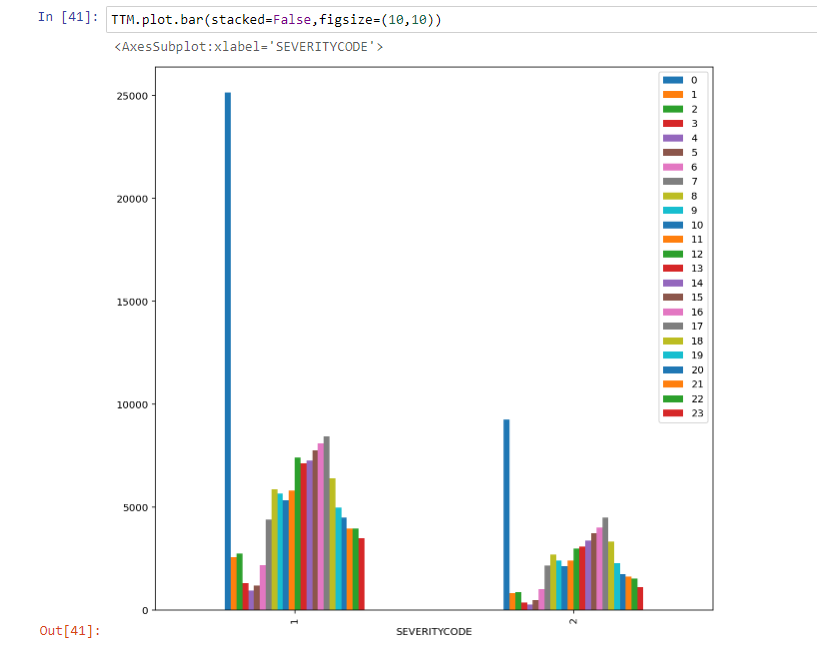


Figure : Time Data

The other datasets that were included in the data modeling were Junction Type, Collision Address Type, Collision Type, Light Condition, Road Condition, and Weather. All the data were grouped by the respective feature separating it by severity code and normalized data.

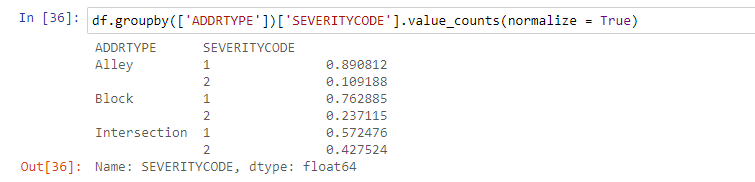
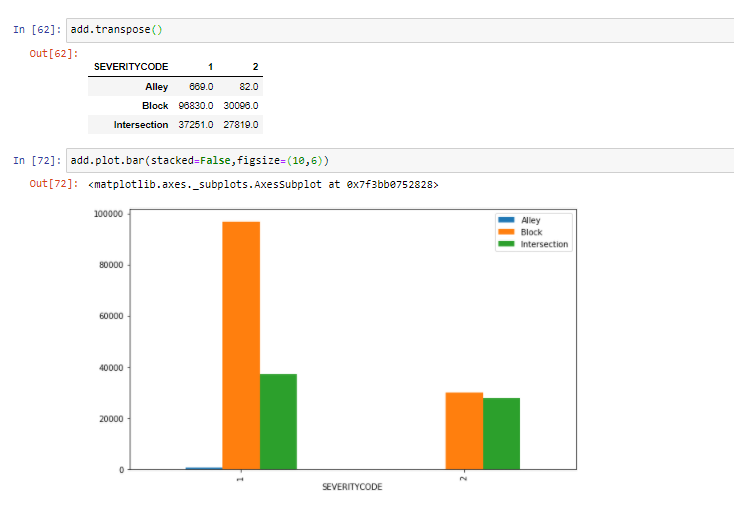


Figure : Collision Address Type

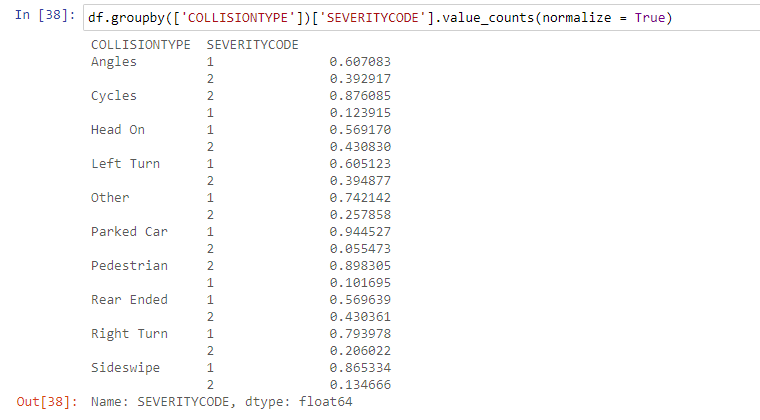
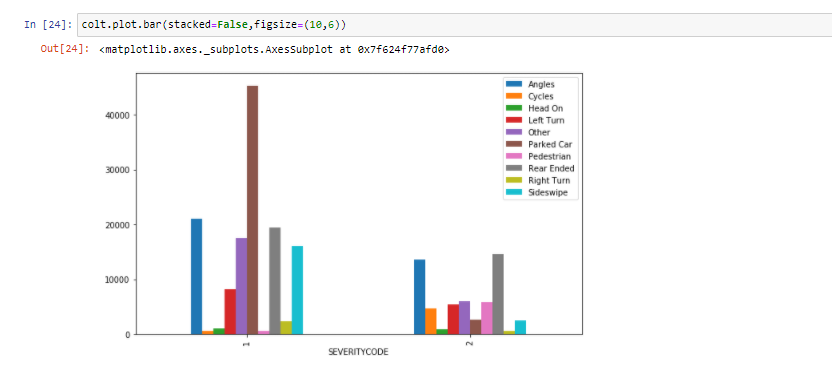


Figure : Collision Type

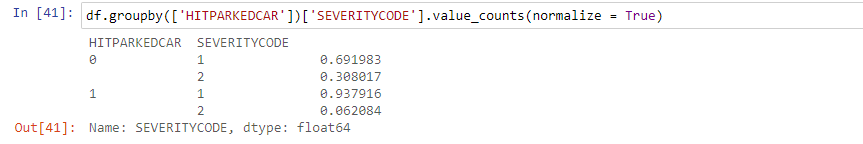
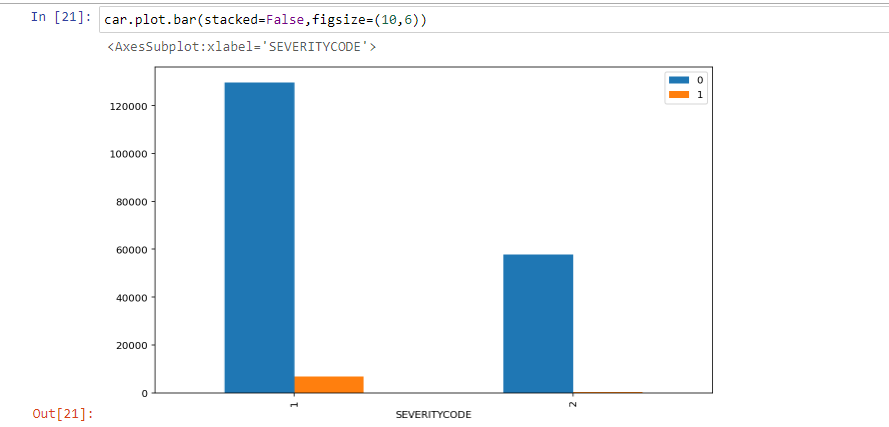


Figure : Hit Parked Car

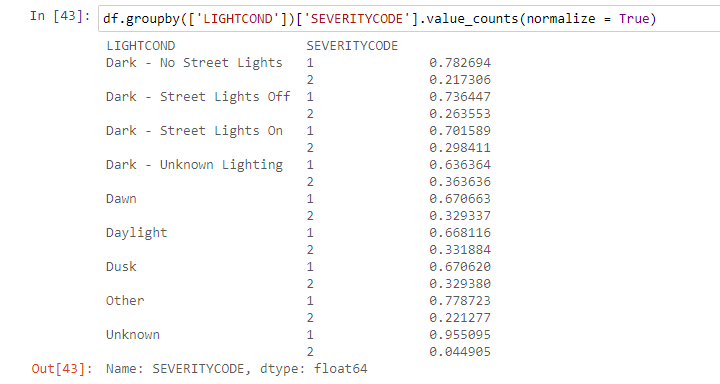
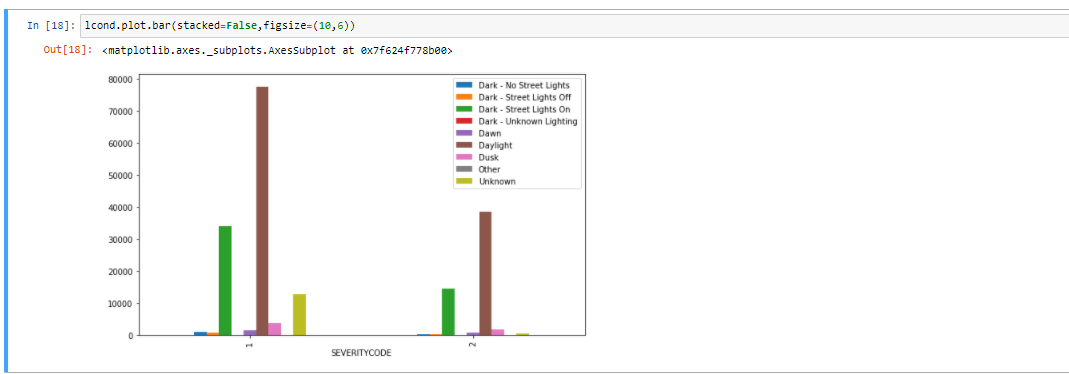


Figure : Light Condition

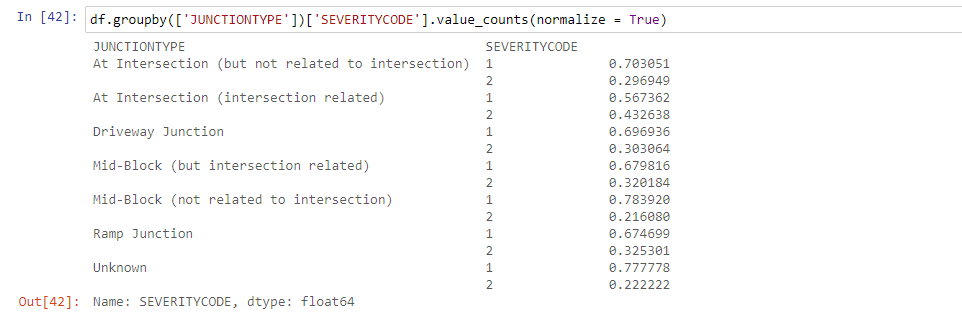
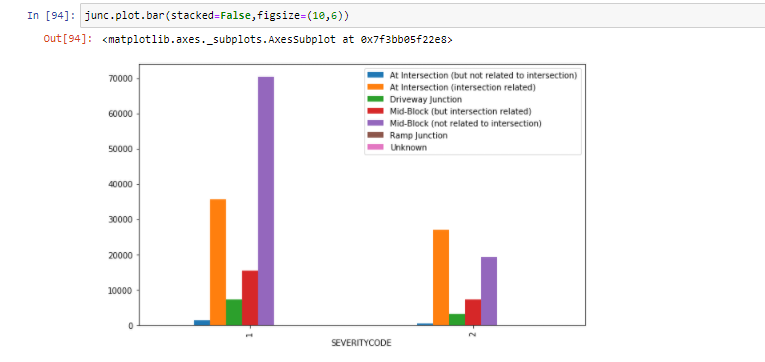


Figure : Junction Type

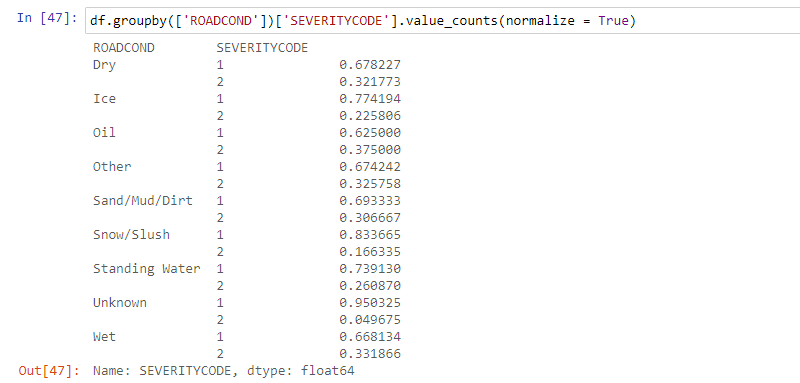
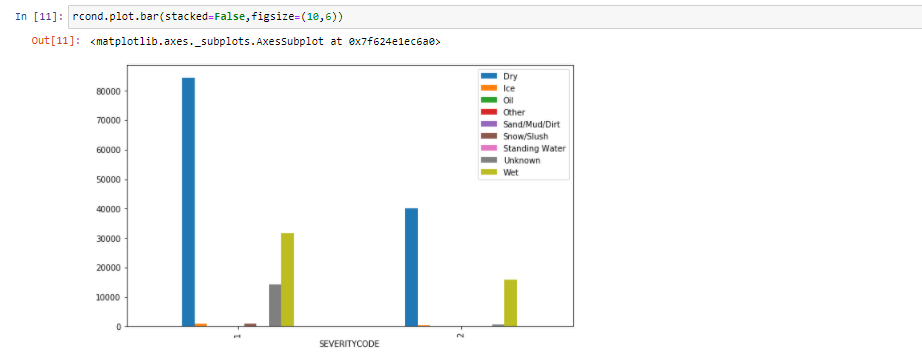


Figure : Road Condition

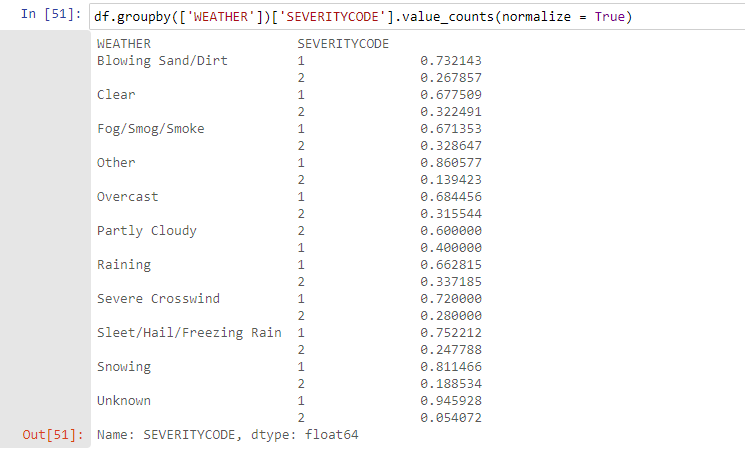


Figure : Weather

* 1. **Data Modeling**

Using one hot encoding method, the above data sets were split into dummies. The dataset was then split into training and testing data with size of 80/20. The data was split into train and test data as to not have overfitting in the results which leads to inaccurate prediction for general datasets.

The dimension of the data set was separated to Train set X: (155738,75), y: (155738,) and Test set X: (38935,75), y: (38935,). The data is fitted, and the model is trained.

There were a few classification methods that was considered for the modeling. KNN, Decision Tree and Random Forest as they are suitable for supervised learning. KNN was not used for the modeling as the kernel could not support the algorithm. To be able to get unbiased results, the data were normalized so that they are weighted appropriately. Decision Tree does not need so much effort for data preparation during pre-processing. The cons for Decision Tree indicate that it could overfit the data. That is why Random Forest was included in the classification method as well as Random Forest is similar to Decision Tree. It builds multiple decision trees and merges them together to get a more accurate and stable prediction. This would give a better performance compared to Decision Tree.

Since the Time data ranges from 65% to 77% for cases with severity code 1, the data modeling will be run twice. Once with Time data included and one without Time data included. The dimension of data without Time included would be Train set X: (155738,50), y: (155738,) and Test set X: (38935,50), y: (38935,) which has 25 less features.

In order to get the best depth for Decision Tree, F1 score was used to get the test accuracy.

It was calculated that the best depth for the Decision Tree was 3 or 4 since they had the same F1 score at 0.689341. To make sure the modeling was constant, a depth of 3 was chosen for the Decision Tree.

1. **Results**
   1. **Result Including Time**

By running the model and using Decision Tree and Random Forest, it was calculated that Random Forest has a training accuracy of 76.63% while it has a testing accuracy of 74.32%.

By having a Decision Tree of depth 3, the F1 score for the test set was calculated to be

84.93% compared to the F1 score for train set at 68.93%.

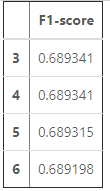


Figure : F1 Score including Time Data

* 1. **Result Excluding Time**

By running the model and using Decision Tree and Random Forest, it was calculated that Random Forest has a training accuracy of 75.28% while it has a testing accuracy of 75.03%. By having a Decision Tree of depth 3, the F1 score for the test set was calculated to be 84.93% compared to the F1 score for train set at 68.93% which is similar to when the modeled data including it.



Figure : F1 Score Excluding Time Data

Based on the four results yield from Decision Tree and Random Forest, Random Forest was chosen over Decision Tree as it returns a better result compared to Decision Tree when Time data was added and removed.

1. **Discussion**

Since the inclusion or exclusion of Time data has a difference of around 1%, it could be redundant as it does not show significant results. The remaining data could be recorded better such as Speeding or Inattention Individual as they could provide significant changes to the data modeled. Data such as Date were not included into data modeling as they were consistent throughout everyday and it will just add extra processing load onto the machine.

Further studies include reducing more data that does not contribute to the data modeling as it is not that important. Another important feature of Random Forest is that it could measure the relative importance of each feature. It computes the score for each feature and scale the result so that the sum of all importance were equal to 1.

A screenshot of a cell phone

Description automatically generated

Figure : Bar graph for feature importances

Based on the bar graph above, it could be concluded that some of the weather conditions were not important and could be dropped for further studies which could improve data modeling and prediction.

1. **Conclusion**

In this study, the relation between several data sets and the severity of accidents were found. The data used has a direct correlation on the severity of the accidents. This model could be very useful for the Seattle Department of Transportation as they could use it to determine red zones for accident and could deploy more staffs on maintaining the safeness of road users based on certain conditions.

The department responsible could also improve or maintain the infrastructure of the road such as the light conditions or the road conditions.

Road users could also determine the risk of driving when route planning as to avoid certain conditions that could result in an accident.