Coursera Capstone Project

Seattle Car Collision Analysis

# Business Introduction

As the world is growing tremendously, demands of the people also increased exponentially. Technology is one of the most important driving factors behind the growth. For ease of transportation, hybrid /electric cars were introduced to the market which attracts the majority of the population. Therefore, more vehicles on the road led to the drastic leap in the collision rates as well.

The Seattle government is going to prevent avoidable car accidents by employing methods that alert drivers through alert display system and police to be more careful in critical situations. In most cases, not paying enough attention during driving, drugs and alcohol or driving at very high speed for fun/competition are the main causes of occurring accidents that can be prevented by deploying harsher regulations. Besides the causes, weather, visibility, or road conditions are the major uncontrollable factors that can be prevented by revealing hidden patterns in the data and announcing warning to the local government, police and drivers on the targeted roads.

The target audience of the project is local Seattle government, police, rescue groups and insurance organisation as well. This model results will provide them to make insightful decisions for reducing the number of accidents for the city

.

# Understanding the Data

The data was collected by the Seattle Police Department and Accident Traffic Records Department from 2004 to present. The data consists of 37 independent variables and 194,673 rows. The dependent variable, “SEVERITYCODE”, contains numbers that correspond to different levels of severity caused by an accident from 1 to 2. Severity codes are as follows:

• 1: Very Low Probability — Chance or Property Damage

• 2: Low Probability — Chance of Injury

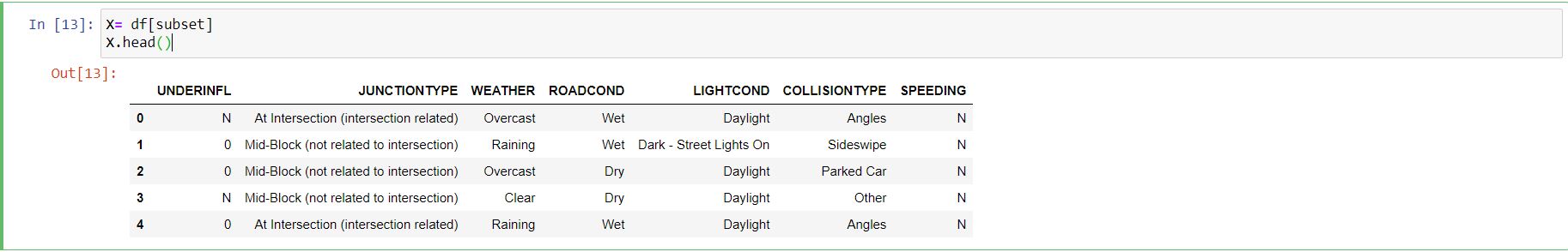
Furthermore, because of the existence of null values in some records, the data needs to be pre-processed before any further processing. Among all the features, I believe the following features have the most influence in the accuracy of the predictions:

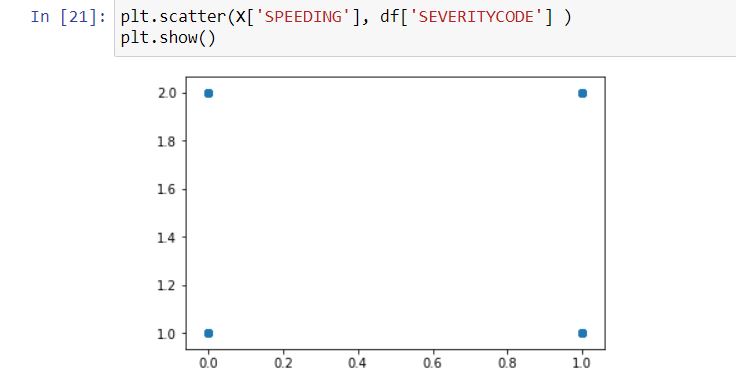
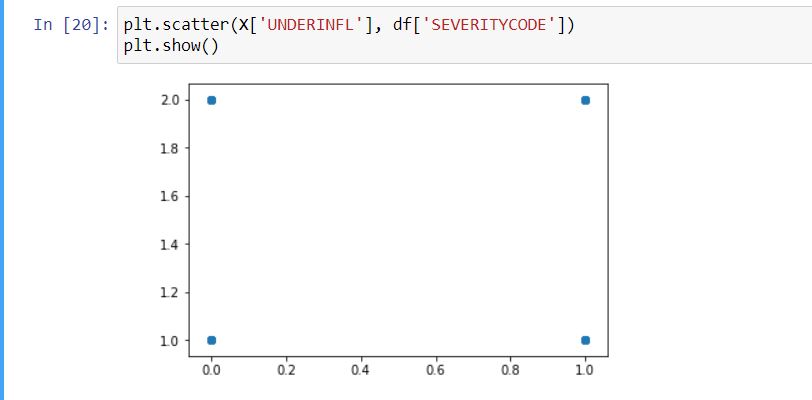
* “'UNDERINFL',
* 'JUNCTIONTYPE',
* ‘WEATHER',
* 'ROADCOND',
* 'LIGHTCOND',
* 'COLLISIONTYPE',
* 'SPEEDING'

But it need not be the correct assumptions. Well let's see what we have here

The target variable is “SEVERITYCODE”.

The features I have selected are shown below:-





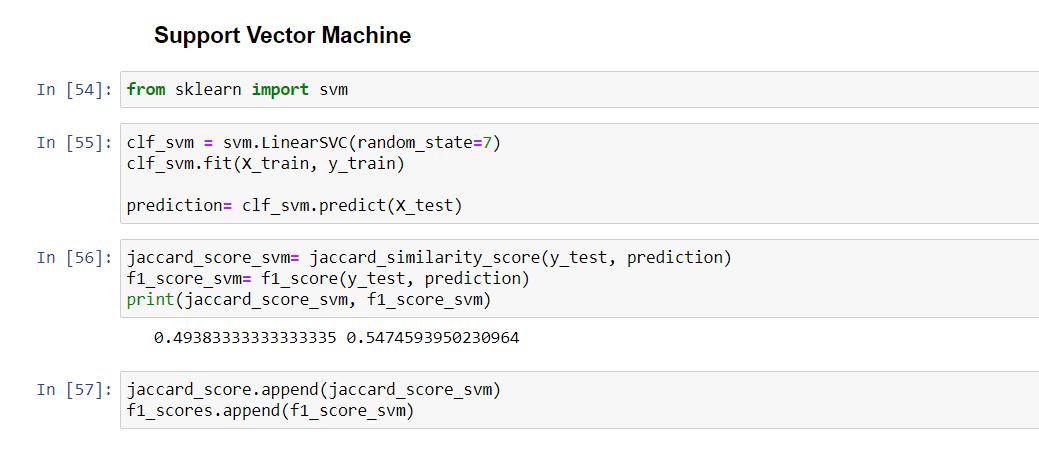
# Methodology

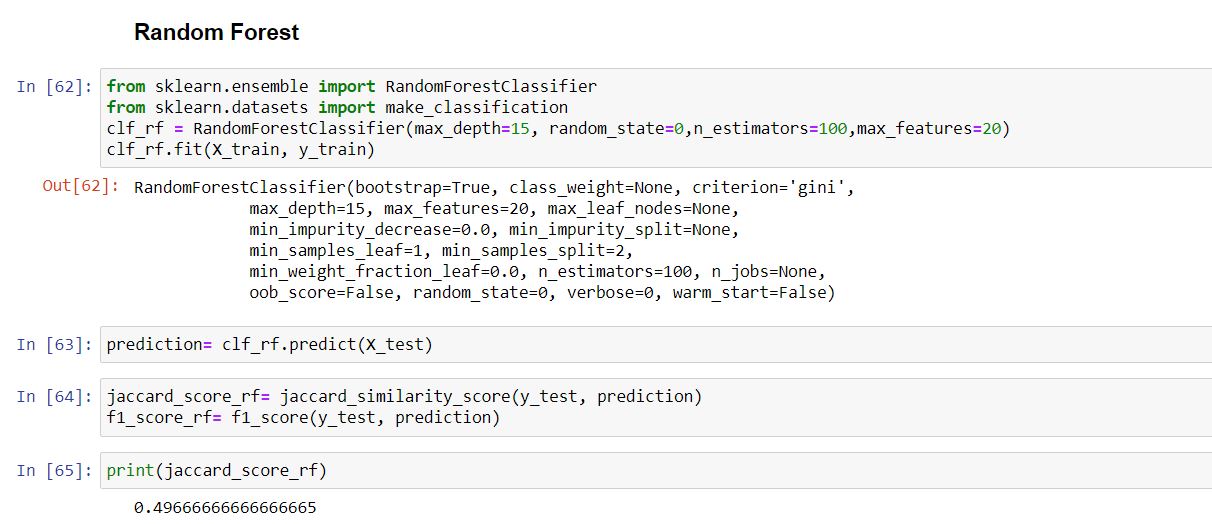
After one hot encoding the categorical data and balancing the whole data sets, the data sets where into the following supervised classification machine learning algorithm and their respective **‘F1-Score’** and ‘**Jaccard-Similarity Score**’ were calculated and the one with best scores was selected.

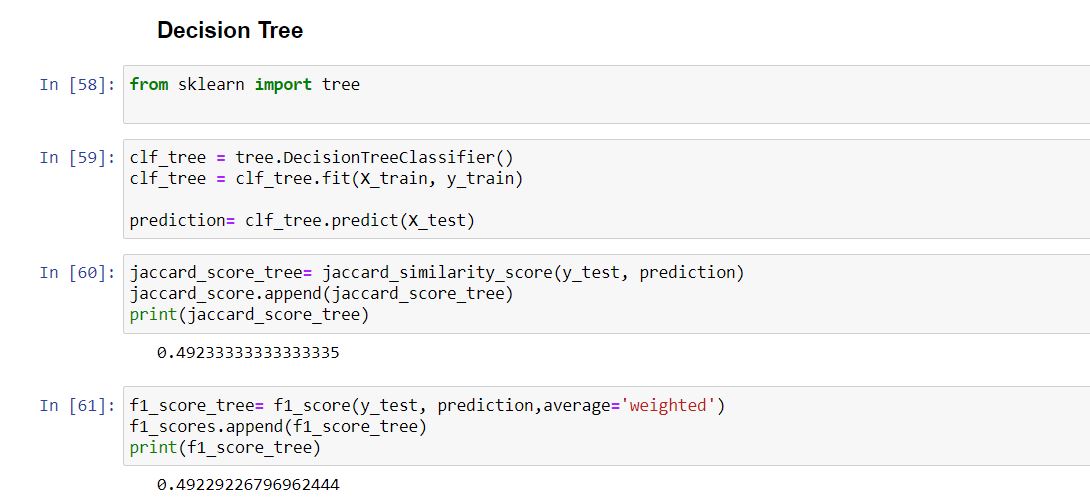
The supervised classification machine learning algorithm used are:

1. **K-Means Algorithm**
2. **Support Vector Machine**
3. **Decision Tree**
4. **Random Forests**



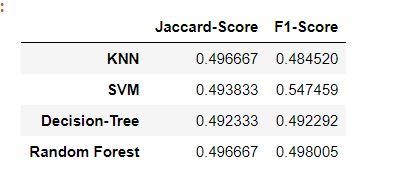






# Results

After successfully training and modelling the different supervised classifier, I have obtained their respective Jaccard Similarity scores and F1-Scores. It is shown below



From the table I have inferred that Random Forest will be a good classifier among the other classifiers with an accuracy score of ~50%.

# Conclusion

Model designed to automate the accident predictions based on the locality, weather, road conditions and other attributes with 50 % accurate making it easy for the Traffic Departments & Residents. The model can be made more effective by adding more datasets and bringing in more flexibility in the attributes of the given datasets.