Capstone Project - Car Accident Severity Report

Submitted By : Pavithra Kannan

30/9/2020

# 1. Introduction

## Background

In United states of America, in the year 2018, there were around 40,000 deaths, 4.5 million injuries and 10 million dollars in property damage associated with vehicular accidents. We are going to study the collision data provided by Seattle Police Department (SPD) recorded by Traffic Records.

## Problem

The purpose of this project is to develop an algorithm that will predict the severity of a car collision by factoring in various attributes. Some of the key attributes that are associated with the severity of accidents are weather, road conditions and Light Condition. This model will alert the drivers when any of these conditions are above danger limit.

# Data Acquisition and cleaning

## 2.1 Data Source

The data was sourced from SPD recorded by Traffic Records. The data consisted of around 37 attributes and 194,673 rows consisting of various factors recorded by SPD.

## 2.2 Data Cleaning

Upon studying the data, "SEVERITYCODE" is the is the target variable. It is the attribute that corresponds to the severity of the collision. The code values can be the following:

3 - Fatality

2b - Serious Injury

2 - Injury

1 - Prop damage

0 - Unknown

The data as it is was not fit for analysis. In order to use the data, all the irrelevant attributes were dropped. Then all the object data type was converted to numerical data types by using label encoding.

After understanding the data, attributes that contribute to the severity of the accidents according to me are

\* "WEATHER" - A description of the weather conditions during the time of the collision

\* "ROADCOND" - The condition of the road during the collision

\* "LIGHTCOND" - The light conditions during the collision

Once the contributing attributes were determined, I checked the values of different features which revealed that the target feature was in imbalance by almost three times. So I used statistical technique to balance it.

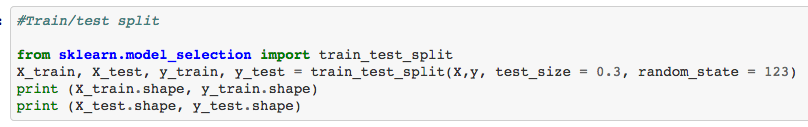
## Methodology

I have used Python as the language in jupyter notebook to do the analysis of the data. I first imported numpy, sklearn and pandas packages.

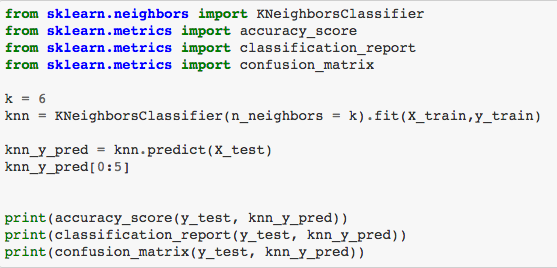
Three machine learning models are applied on this data:

1. K Nearest Neighbour (KNN)
2. Decision Tree
3. Linear Regression

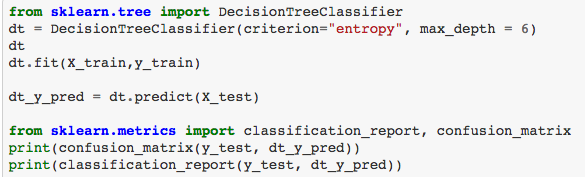
Then I split the dataset into training and predicting data set using the following code:



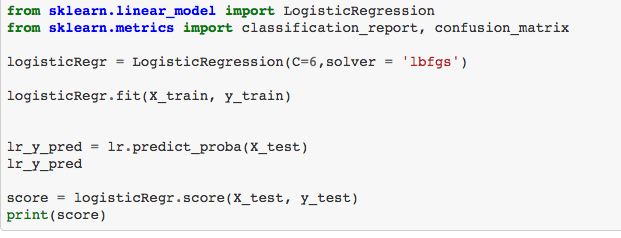
Then I applied the KNN method. KNN will help predict the severity code of an outcome by finding the most similar to data point within k distance.



Then I applied Decision Tree method. A decision tree model gives us layout of all possible outcomes so we can fully analyze the consequences of a decision. It context, the decision tree observes all possible outcomes of different weather conditions.



Then I applied Logistic Regression method. Because our dataset only provides us with two severity code outcomes, our model will only predict one of those two classes. This makes our data binary, which is perfect to use with logistic regression.



# 4. Result and Evaluation

The accuracy of our models is as below

|  |  |  |
| --- | --- | --- |
| **NO** | **Model Name** | **Accuracy** |
| 1 | KNN | 0.70 |
| 2 | Decision Tree | 0.70 |
| 3 | Linear Regression | 0.70 |

Based on the above table, all the three models give the same accuracy.