Capstone Project

Title - Car accident severity

Publication Date - 2020

Author – Yugal kumar

**Introduction**

Car accidents can occur all the time, however there are some conditions were the probabilities of have an accident arise due multiple variables.

This report has as purpose develop a model for Seattle government to predict the probabilities of have a car accident and severity, based on different conditions as weather or road conditions.

The information was provided by Seattle Police Department from 2004 to 2020.

This analysis has multiple applications like an app that will prompt the drivers to be more careful depending on the weather and road conditions on any given day or a way for the police to enforce more safety protocols. In this instance, I am using the data from the City of Seattle’s police department showing all the collisions from 2004 till present.

# ****Business Problem****

Identify the conditions that can cause future car accidents in order to alarm the people with anticipation to be aware and drive more carefully.

The avoid of accidents will result in multiples benefits:

* Save lives as main benefit
* Reduce costs in damage infrastructure
* Reduce cost from police and paramedics to attend each accident

# ****Methodology****

I used Jupiter Notebooks to conduct that analysis and imported all the necessary Python libraries like Pandas, Numpy, Matplotlib, and Seaborn. The data was mostly categorical so I stuck to graphical representation to see correlation between various variables.

I started by importing the csv file and to prepare the data, I dropped the columns we do not need from the dataset, i.e., columns that do not have values or where the values are unknown. Even though this is an important factor, I dropped Speeding entirely because it is missing over 180,000 values and this can hamper the results.

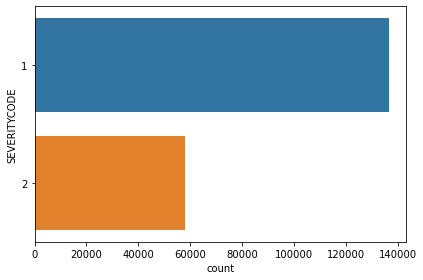
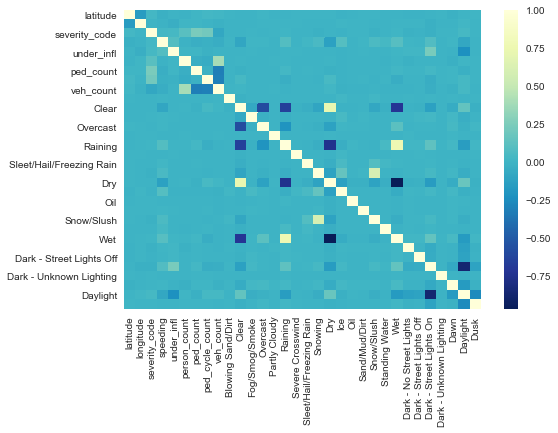
### **Data Cleaning & Pre-Processing**

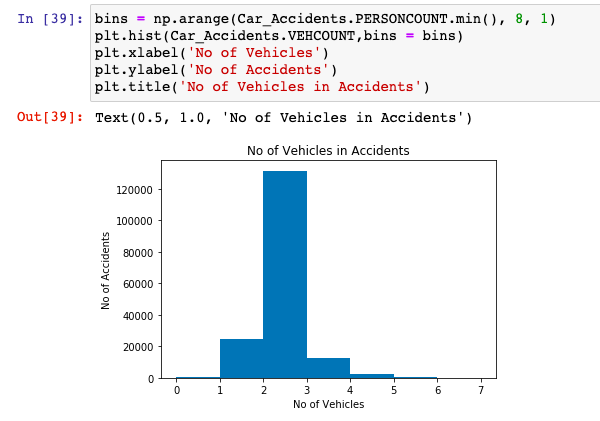
#### This dataset is obtained from an external source and usually any dataset obtained will not be in a clean format ,it will always contain some missing values and some irrelevant data which are just trash to the dataset. This dataset too consist of a lot of missing values and useless datas .so before further processing the data it is important to clean the dataset

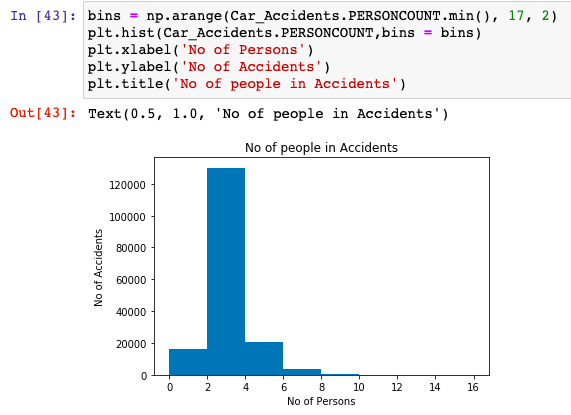
#### STATUS, INTKEY,OBJECTID, INCKEY, COLDETKEY, REPORTNO, EXCEPTRSNCODE, EXCEPTRSNDESC, INCDATE, INATTENTIONIND, PEDROWNOTGRNT, SDOTCOLNUM, SEGLANEKEY, CROSSWALKKEY has a lot of missing values and they are not useful and irrelevant for this dataset. so they are too dropped

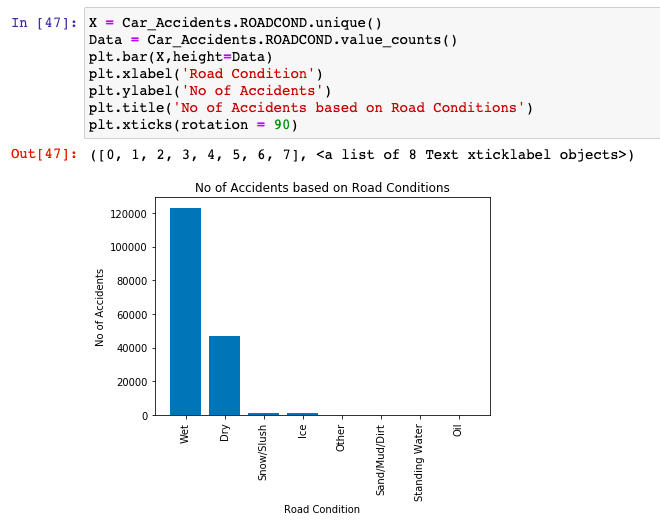
**EDA – Exploration Data Analysis**

##### Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. Finding the correlation among the features of the dataset helps understand the data better. For example, in the heatmap shown below, it can be observed that some features have a strong positive / negative correlation while most of them have weak / no correlation.









### **Modelling and Evaluation**

#### The process of modeling means training a machine learning algorithm to predict the labels from the features, tuning it for the business need, and validating it on holdout data.The output from modeling is a trained model that can be used for inference, making predictions on new data points.

#### Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. To avoid overfitting, both methods use a test set (not seen by the model) to evaluate model performance

#### Decision Tree makes decision with tree-like model. It splits the sample into two or more homogenous sets (leaves) based on the most significant differentiators in the input variables. To choose a differentiator (predictor), the algorithm considers all features and does a binary split on them (for categorical data, split by category; for continuous, pick a cut-off threshold). It will then choose the one with the least cost (i.e. highest accuracy), and repeats recursively, until it successfully splits the data in all leaves (or reaches the maximum depth).

#### Information gain for a decision tree classifier can be calculated either using the Gini Index measure or the Entropy measure, whichever gives a greater gain. A hyper parameter Decision Tree Classifier was used to decide which tree to use, DTC using entropy had greater information gain; hence it was used for this classification problem.

# ****Discussion****

At the start of our analysis, I was trying to figure out the severity and frequency of road accidents based on weather conditions, road conditions, and other factors. Even though our data was a good size, there were a number of missing elements and we needed to clean the data in order to get a good result. We had to drop ‘SPEED’ because there were too many missing elements but I think that is an important factor that should be considered. From the analysis, it is clear that most accidents involve solo drivers, on wet roads, bad weather, at intersections, and are minor in nature. This could be helpful to the police department in understanding where to install more stop signs, or maybe adding cameras to intersections to compel people to slow down. We also live in a technologically friendly world so maybe we can develop some inbuilt technology in our cars that warn us when the road and weather conditions are bad, or the car is approaching a stop sign.

# ****Conclusion****

Although this analysis has given us some good insight, there needs to be a closer inspection of certain other variables. It seems like a lot of these accidents are minor and avoidable. Having said that there is still a considerable amount of loss of property and these findings can be helpful to the Seattle PD in enforcing some new measures to prevent future accidents.