**Capstone project – Predict car accident severity**

**Introduction:**

Capstone Final project based on Seattle Accidents.

Objective is to Predict “Severity Code” based on human fatality, traffic delays, weather conditions, road, and visibility conditions. The data is available on IBM Capstone project in Coursera.

Depending on weather conditions the model will predict and alert the drivers and caution them reducing accidents and injuries.

**Data:**

**Data contains 194,673 rows and 38 columns.**

Target variable is “Severity Code”. It is the dependent variable and contains numbers from 1 to 2 correspond to different levels of severity caused by accidents.

Data has null values, missing values, deleting unrelated columns and cleansed during data pre-processing.

Next step is EDA to visualize data. Process followed was selecting attributes that have higher impact in accident occurrence. Attributes used to weigh the severity of an accident are 'WEATHER', 'ROADCOND' and 'LIGHTCOND'.

**Data Pre-processing:**

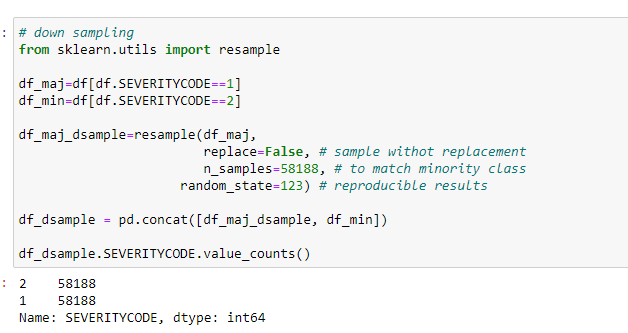
To prepare for data analysis data is cleansed by removing unrelated columns and missing values and null values and as most of attributes are object data types they were converted into numeric data types.

For Analyzing data, following features were selected

WEATHER, ROADCOND, LIGHTCOND, SEVERITY CODE.

After analysis the Target variable SEVERITY CODE is not perfectly balanced. Severity code 1 is three times volume of severity code 2.

The issue was resolved by sampling majority class.



Feature selection:

Selected attributes are categorical data types. As such they must be changed to numerical data types by replacing categorical variables with mode.

**METHODOLOGY:**

Jupyter notebook was used to do the data analysis. GitHub is used as Repository.

Python libraries NumPy, Pandas, Seaborn, Matplotlib were used. Data is imported from the sheet pd.read\_csv. We check attributes names and data types.

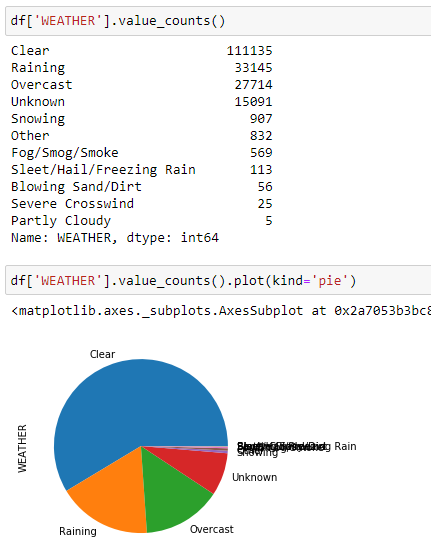
I have checked shape of data (194,673 ,38) total rows and columns. I have selected important Attributes to predict accident severity in Seattle.

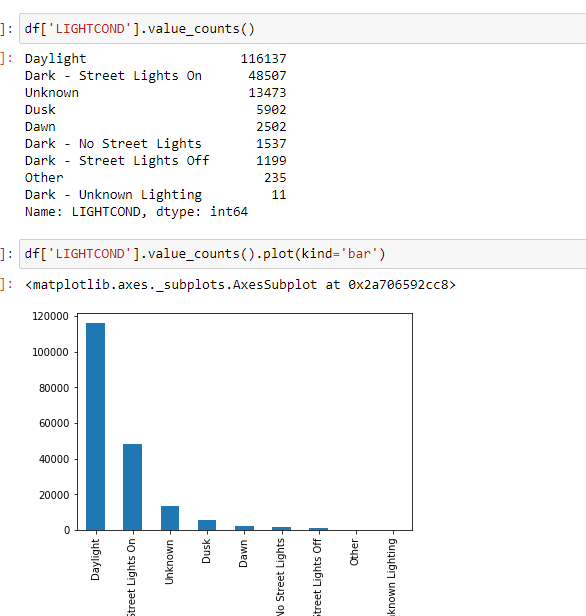
The following features have best accuracy Prediction.

* WEATHER
* ROADCOND
* LIGHTCOND

As mentioned, SEVERITY CODE is Target variable.

Data is visualized using Bar graphs, pie charts, Line charts. Value counts were done on ROADCOND, WEATHER to check road conditions during bad weather. Value counts were done on LIGHTCOND to check for accidents different light conditions.





**One hot coding:**

One hot coding converts selected numerical datatypes into dummies.

Machine learning model: Data is now ready for import ML models. Below are the ML models used

**KNN (K Nearest Neighbor)**

will help us predict the severity code of an outcome by finding the most like data point within k distance.



**Decision Tree**

A decision tree model gives us a layout of all possible for analyzing the consequences of a decision. Its context, the decision tree observes all possible outcomes of different weather conditions.



**Logistic Regression**

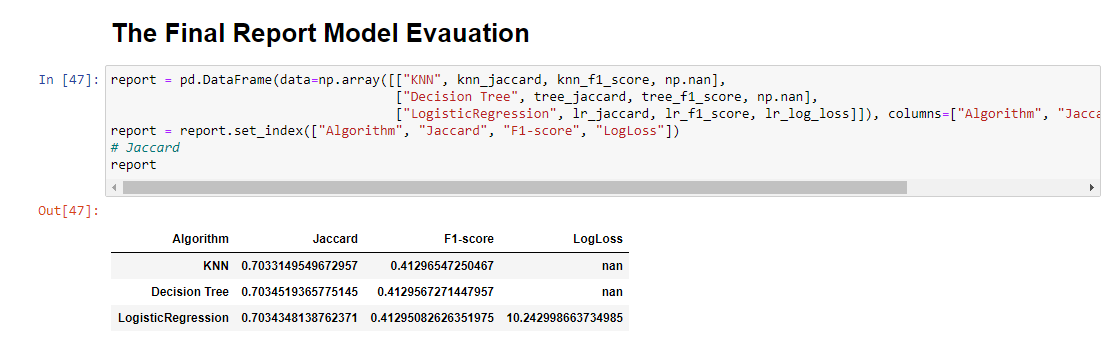
As the dataset only provides us with two severity code outcomes, our model will only predict one of those two classes. This makes our data binar which is perfect to use with logistic regression.



**Results & Evaluation:**

The results of the model evaluations are summarized in the following table:





**Discussions:**

In the Seattle Accident data set all categorical datatypes must be converted into numerical and again into dummy variables or label encoding to create into new datatypes as int.

In the target variable which is imbalanced mentioned before, Severity Code 1 is three times volume than Severity code 2

The solution to this was down sampling the majority class with sklearn's resample tool. We down sampled to match the minority class exactly with 58188 values.

After cleaning data, null values were converted to NAN and categorical values were converted into numeric data types. ML models KNN, Decision -Tree, Logistic-Regression were then applied for analysis

**Based on the model Decision-Tree is best model to predict Car Accident.**

**Conclusion:**

Based on data provided by Coursera for Capstone project from weather conditions certain classes of data were selected for analysis. It can concluded that weather conditions have an impact on whether travel could result in property damage (class 1) or injury (class 2).

