**Introduction:**

The problem statement is to identify the probability and severity of an accident on a particular day and in a particular weather. Various factors like weather, time, day, road conditions etc will be analysed to predict what type of accidents can occur during that time. Moreover, severity of an accident will also be predicted.

**Target Audience:** This problem will be of interest to residents of Seattle and people who are travelling to/from Seattle. Solution to this problem would help these people to travel safely. It would help people in understanding the risks of travelling on a particular day depending upon the weather conditions. For ex: If a person who wants to travel to Seattle on a rainy day at night, he/she can check chances of an accident and its severity. Consequently, making the decision of travelling or not.

**Data Gathering**:

“Collisions—All Years” dataset was used for this problem. Metadata of this dataset can be found on this link : <https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf>

Dataset can be downloaded from this link : <https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv>

This data set has around 200k observations and 37 attributes. SEVERITYCODE of the accident is used as a label and this will be used in the end to check the severity of the accident that can occur given other conditions.

**Understating data:** Some attributes in the dataset like LOCATION, EXCEPTRSNCODE, EXCEPTRSNDESC, SEVERITYCODE, PEDROWNOTGRNT or SPEEDING cannot be used directly. Some of these columns have missing values that need to addressed properly and some columns like REPORTNO and LOCATION cannot be used entirely.

Moreover, few columns like LIGHTCOND, ROADCOND, JUNCTIONTYPE, ADDRTYPE, WEATHER will have to be hot encoded for proper data modelling.

Columns like OBJECTID, INCKEY, COLDETKEY, REPORTNO cannot be used in model creation as these columns do not have the information required to predict accident severity.

**Methodology:**

**Data Pre-processing:**

1. After looking at every data column, it was evident that only few columns can be used for model training. After dropping trivial columns from the data, only 'SEVERITYCODE', 'ADDRTYPE', 'WEATHER', 'ROADCOND', 'LIGHTCOND' were left.

'WEATHER', 'ROADCOND' and 'LIGHTCOND' stand for weather, road and light conditions. These values can be used to check the severity of the accident. Moreover, ADDRTYPE (address type) can be used to aware the driver whether he/she should be more aware around Alley, Block or Intersection.

1. After removing unnecessary columns down-sampling was done to make SEVERITYCODE 1 and 2 accident values balanced.
2. Next step was dealing with missing values and categorical variables.

* For missing values, all rows were removed and Unknown values were also removed.
* For categorical values, dummy variables were created and final column count was 29 with 105043 samples.

**Model training:**

For prediction of accident severity three different algorithms were used: Random Forest, K-Nearest and Logistic Regression.

1. Random Forest Classifier was used to train the data initially. However initial accuracy was just around 58%. Later, using Grid Search to find best parameters, accuracy was increased to nearly 72%.
2. Logistic regression gave just 58.18% accuracy and performed poorly.
3. For KNN, different values of nearest neighbours ranging from 3 to 20 were tried but accuracy was still short of 58%.

Therefore, Random forest with best parameters and accuracy of 71.76% was finalised.

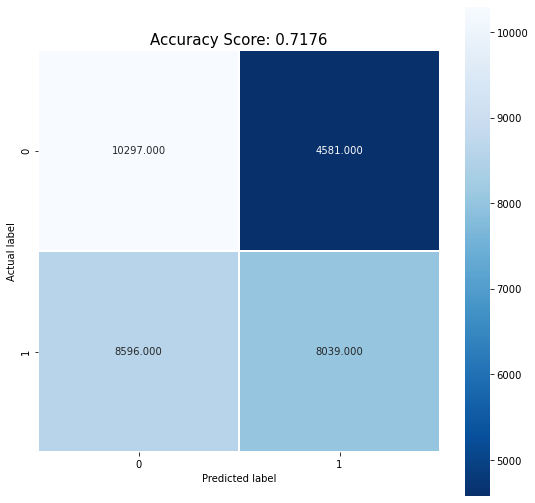
**Results:**

Final predictions results of accident severity were:

Average Error: 0.4185 degrees.

Accuracy = 71.76%.

Moreover, for model training of this algorithm, the parameters were 'n\_estimators': 200, 'min\_samples\_split': 2, 'min\_samples\_leaf': 4, 'max\_features': 'sqrt', 'max\_depth': 45, and 'bootstrap': True.



**Discussion:**

The final results of this model can be greatly improved by more data. In the dataset used, the data was unbalanced and most of the attributes could not be used for prediction of accident severity.

Our problem was to detect accident severity for person depending upon the weather, road and lighting conditions. Moreover, factors like Drug-influence and intersection could have been influential but could not be used due to problem requirement. More information on car type, season of year, car tyre types etc can be used for accurate accident severity prediction.

**Conclusion:**

To conclude, weather, road and lighting conditions can accurately predict the accident severity three-fourth of the time. However, more data and more attributes as discussed above can be very useful in accurate prediction.

Neural networks can also be used for better prediction of the accident severity.