**IBM DATA SCIENCE**

**CAPSTONE PROJECT**

**CAR ACCIDENT SEVERITY PREDICTOR**

**Scope of Work**

The Seattle government is going to prevent avoidable car accidents by employing methods that alert drivers, health system, and police to remind them to be more careful in critical situations.

In most cases, not paying enough attention during driving, abusing drugs and alcohol or driving at very high speed are the main causes of occurring accidents that can be prevented by enacting harsher regulations. Besides the aforementioned reasons, 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 last but not least, car insurance institutes. The model and its results are going to provide some advice for the target audience to make insightful decisions for reducing the number of accidents and injuries for the city.

**Data**

The data was collected by the Seattle Police Department and Accident Traffic Records Department from 2004 to present.

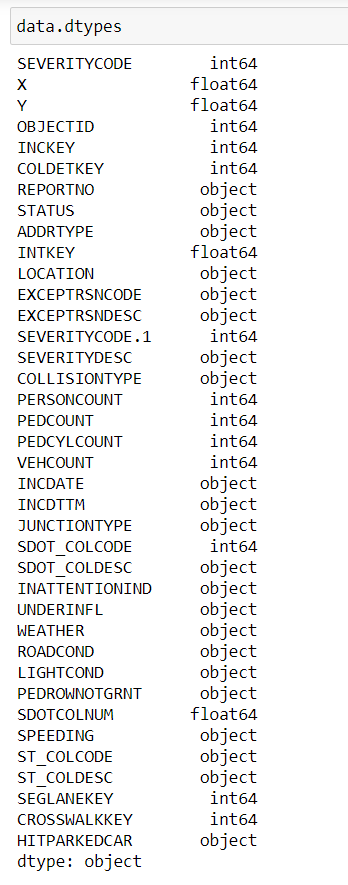
The data consists of 37 independent variables describing the details of each accident including the weather conditions, collision type, date/time of accident and location (latitude and longitude) and 194,673 rows.

The dependent variable, “SEVERITYCODE”, contains numbers that correspond to 2 different levels of severity caused by an accident.

Severity codes are as follows:

1: Very Low Probability — Chance or Property Damage

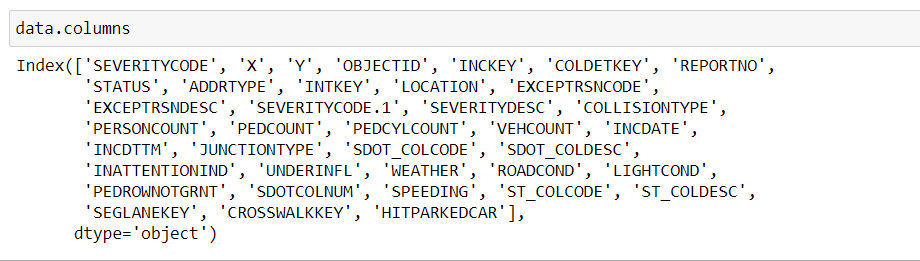
2: Low Probability — Chance of Injury



**Data Pre-processing**

**1. Dataset includes unnecessary/redundant columns:**

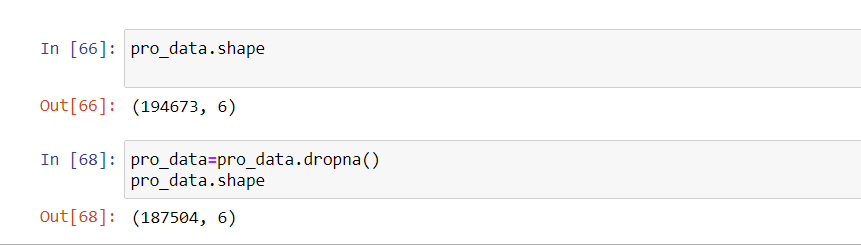
The accident dataset includes many columns of metadata (such as incident report numbers) and columns which duplicate information which is already included in other columns (such as a text field “SEVERITYDESC” which provides a written definition of the accompanying accident severity code, the target variable). Columns which include unnecessary/redundant information were removed from the dataset.



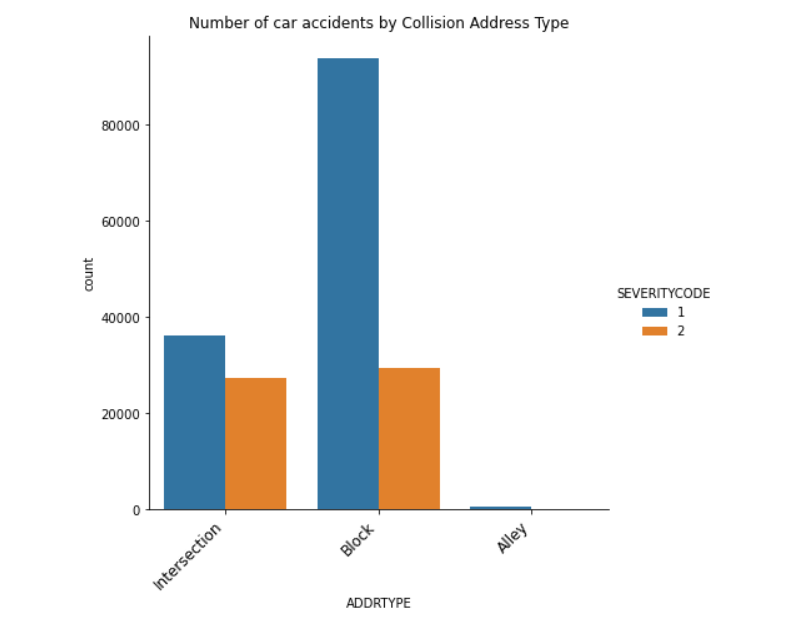
After analysing the data set, I have decided to focus on only six features Severity Code, Address Type, Collision Type, Weather, Road Conditions and Light Conditions.

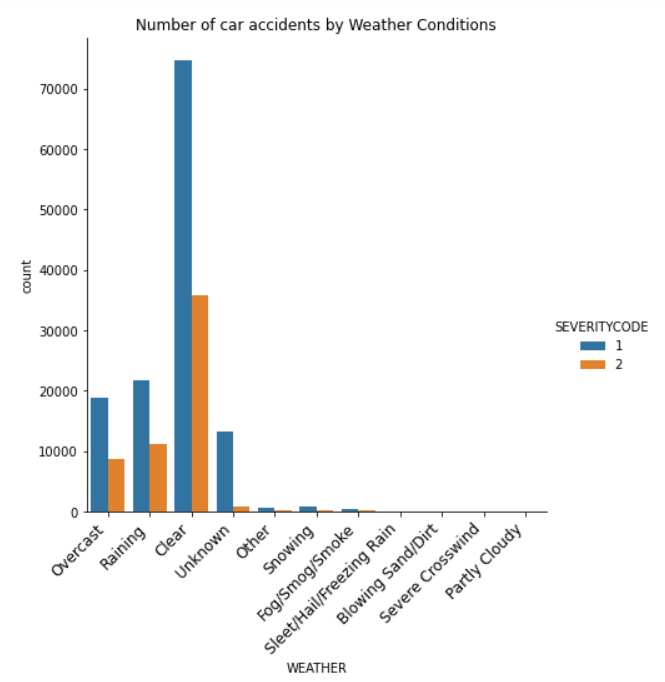
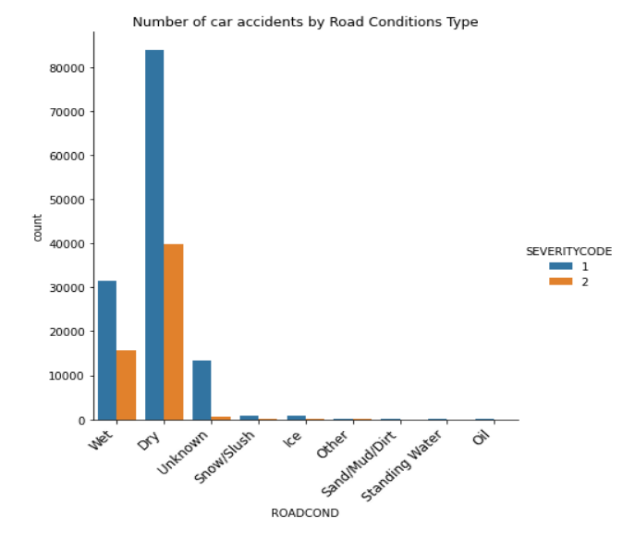
**2.** **Data incompleteness:**

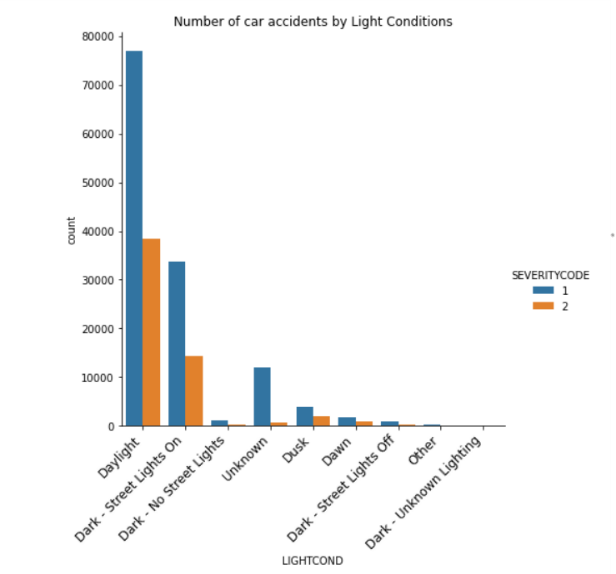
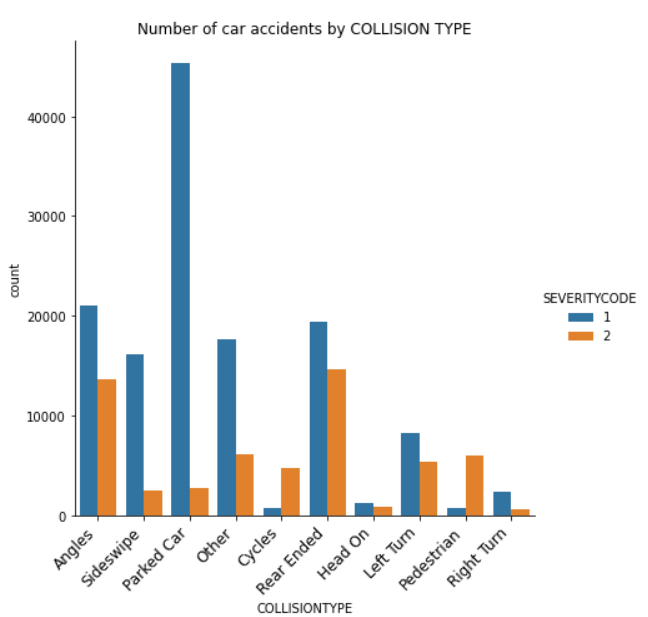
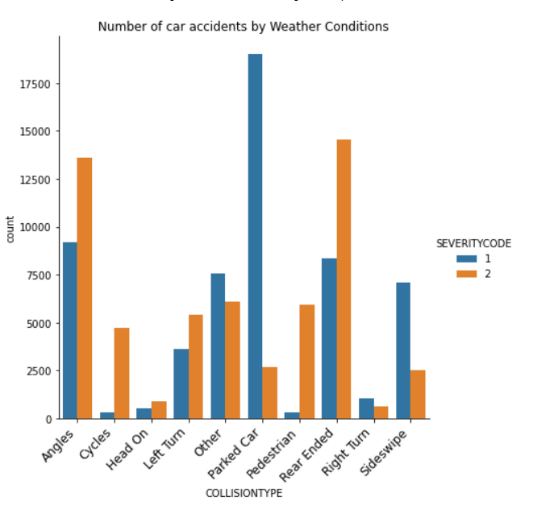
**A**round 4% of the accidents in the dataset are missing one or more key features, including in some cases the target variable (accident severity code) and in others, are missing information about weather or road conditions. As the purpose of building the model is to see how these various features interact and influence the overall accident severity, data entries which are missing one or more of these key features are not useful, and were removed from the dataset.



After processing the missing values I visualised the chosen attributes to get a better understanding of the data.

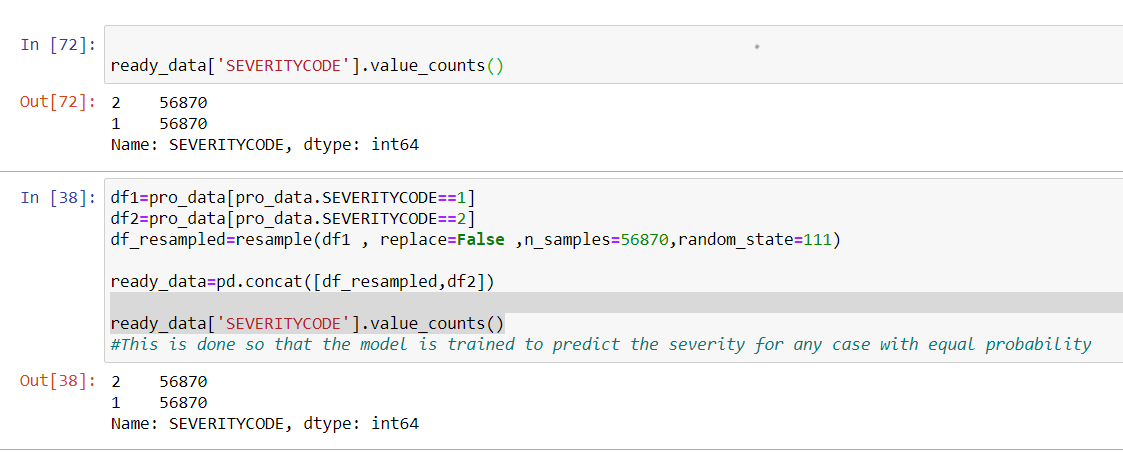




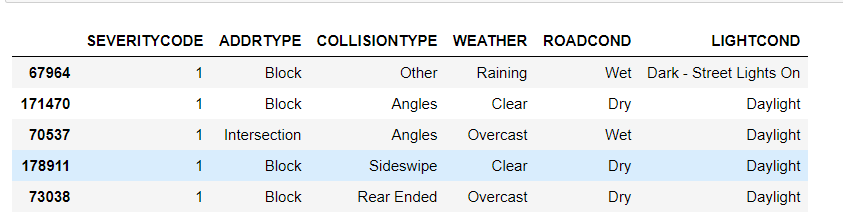


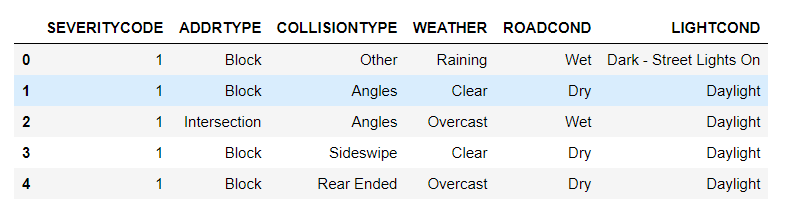
**3. Data is imbalanced:**

After visualising and analysing the data, I checked difference in values for different severity codes and the results show, the target feature that is imbalanced and needs to be balanced. Furthermore, the attributes Road Condition, Weather and Light Condition had values as ‘Unknown’ and ‘Other’ which basically meant the same so I changed the values for the ‘Other’ to ‘Unknown’. And then resampled the data set so that the number of rows for Severity Code ’1’ and number of rows for Severity Code ’2’ are equal.

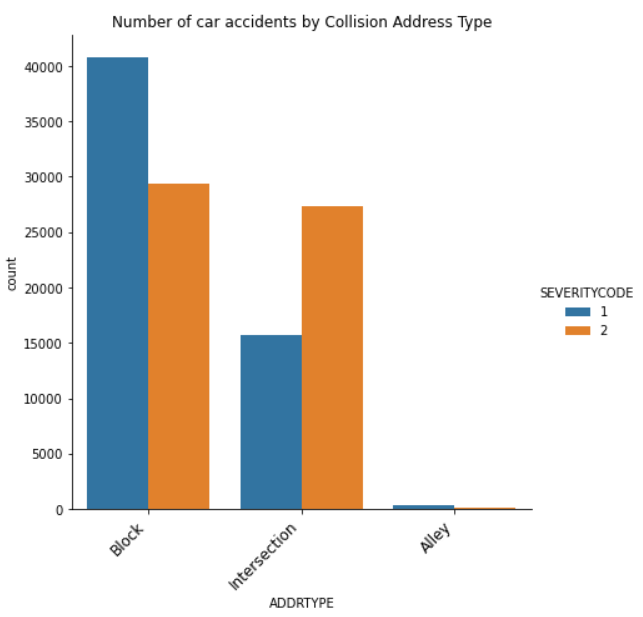


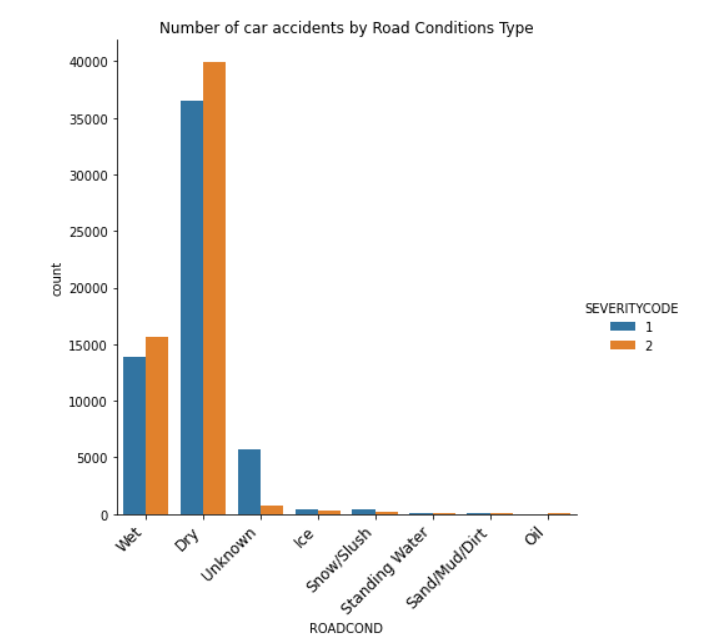
Doing so resulted in the dataframe index to become irregular,so the dataframe index had to be reset. For this I used reset\_index() function reseting the index starting from 0.

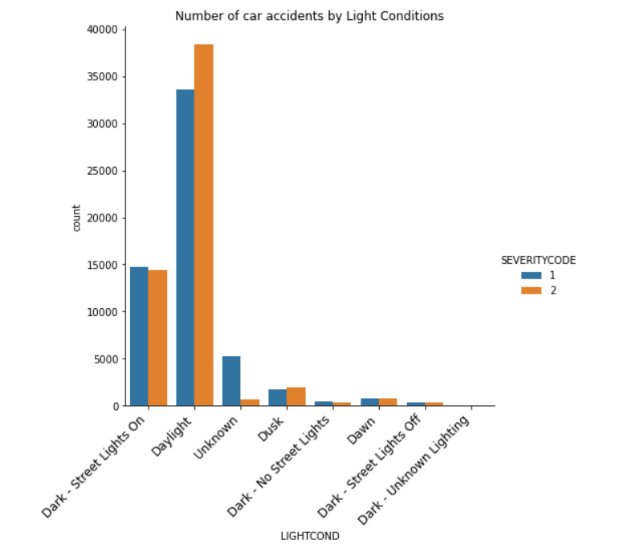


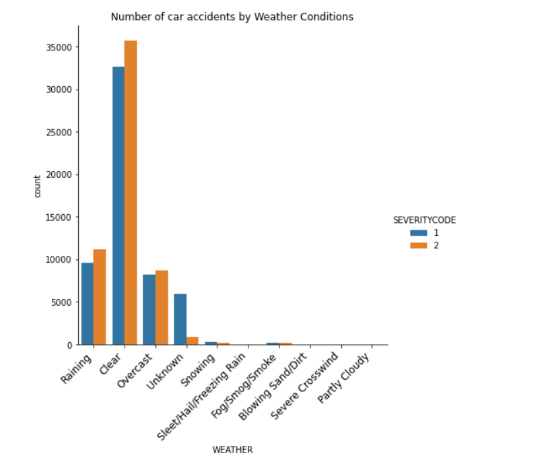


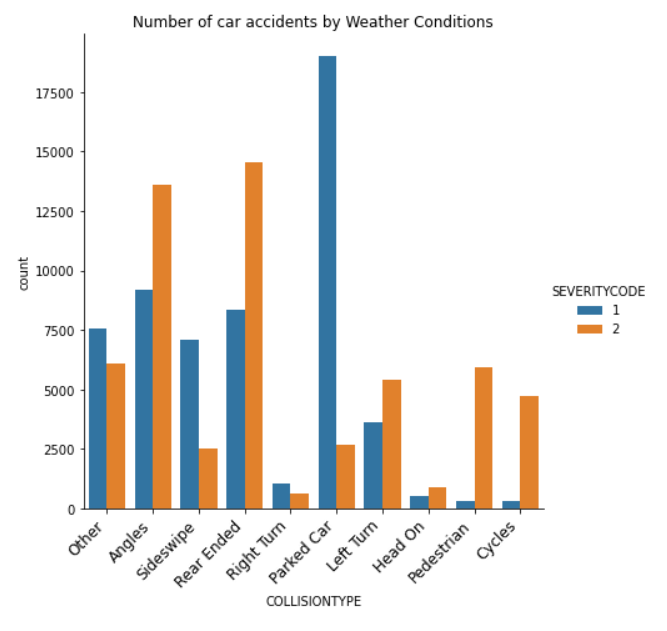
Before building a classification model, I visualised the attributes to see the data distribution for different severity codes.





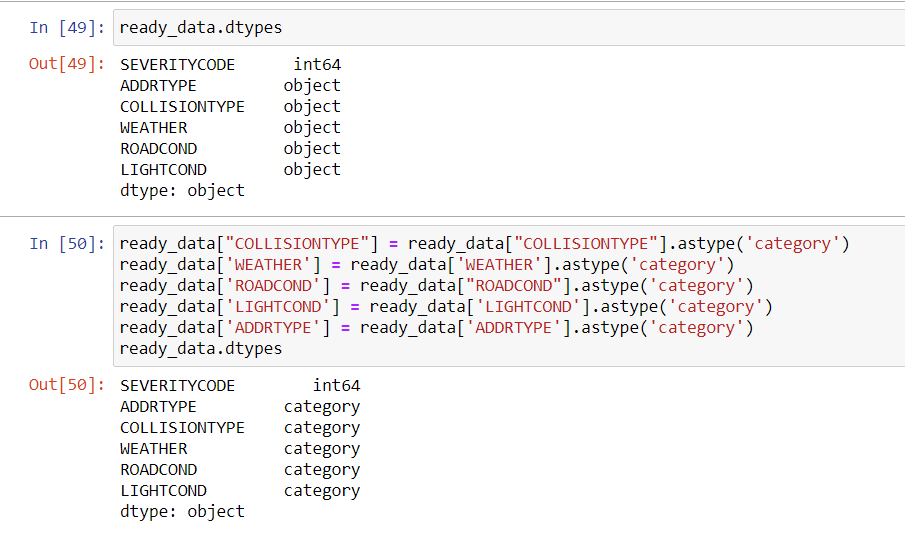




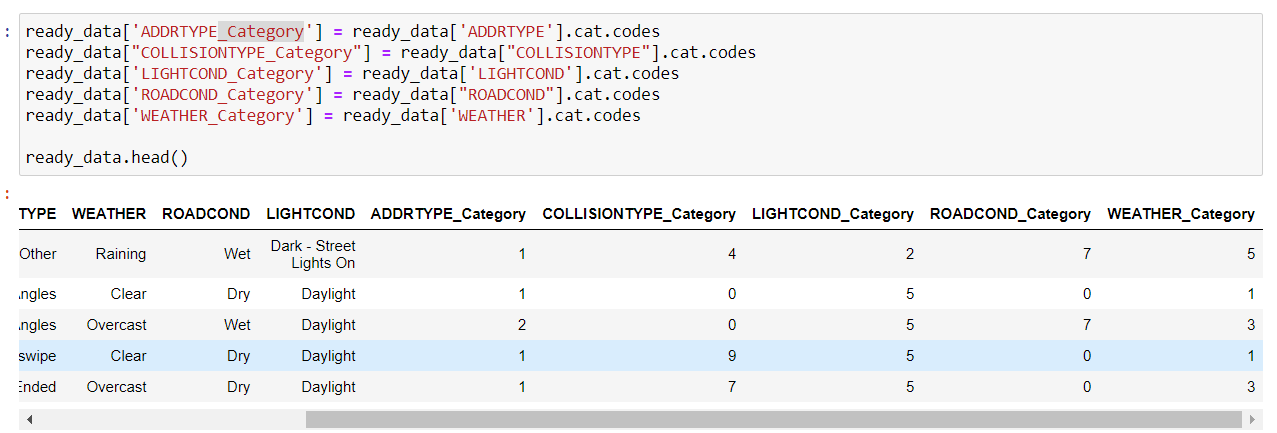


**4. Presence of categorical/non-numeric data:**

Machine Learning models require numerical data, and cannot handle alphanumeric strings. For example, each entry in the “WEATHER” column contains a text string which takes one of eleven values (e.g. “Clear”, “Rain”, “Snow”, etc) which describes the prevailing weather conditions at the time of the accident.



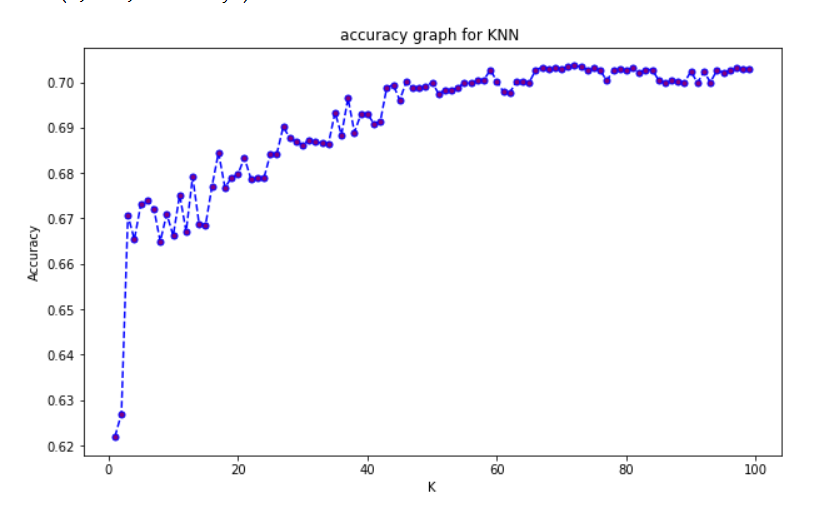
For this using astype() first, I converted the data from object to category data type and the encoded them in numerical values using Label Encoder.

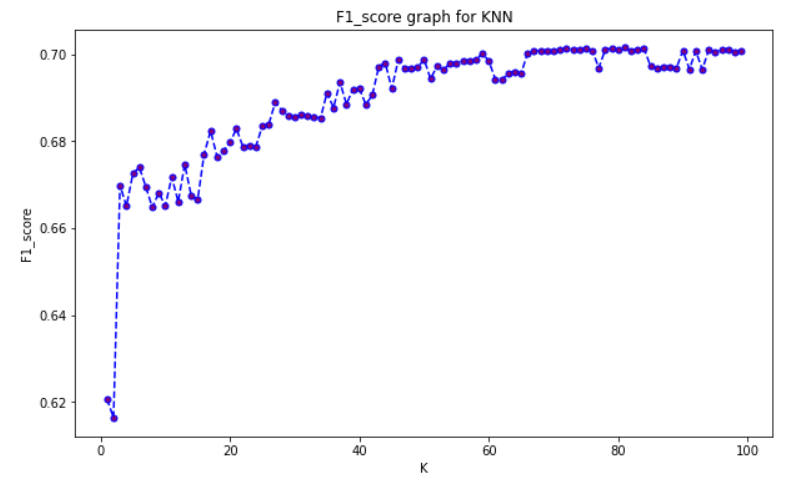
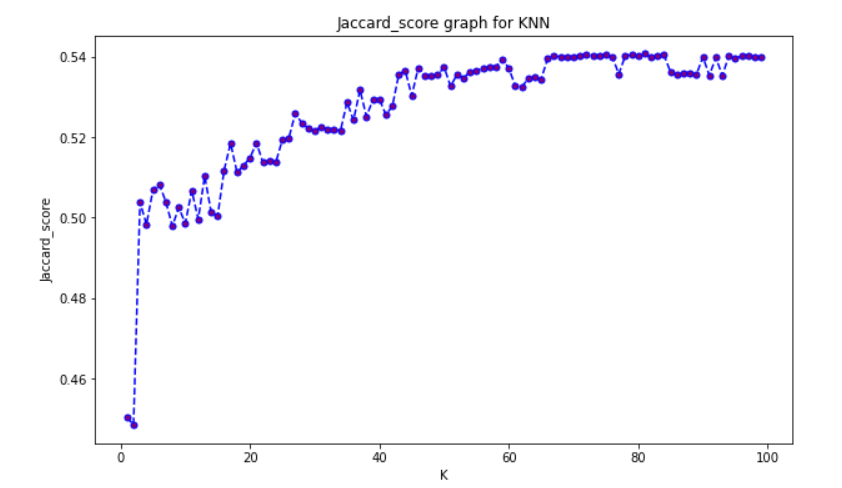


**Model Development**

In order to develop a model for predicting accident severity, the re-sampled, cleaned dataset was split in to testing and training sub-samples (containing 21% and 79% of the samples, respectively) using the scikit learn “train\_test\_split” method. In total, three models were trained and evaluated.

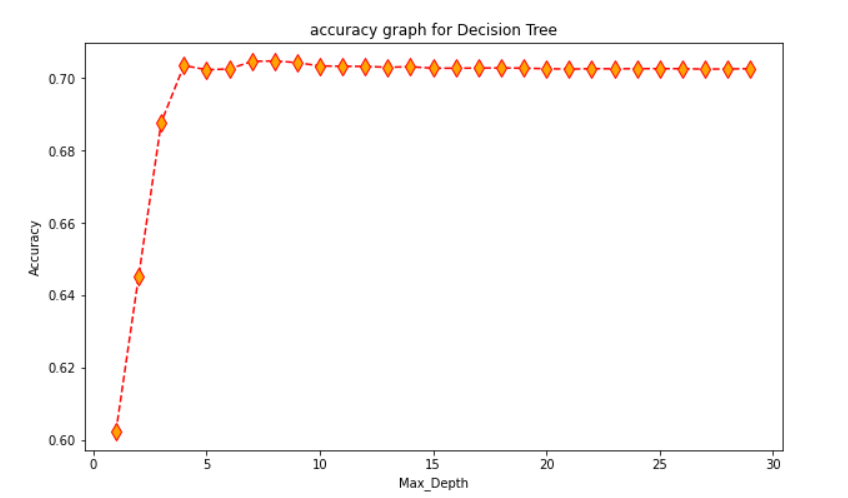
**1. k-Nearest Neighbour Model(kNN model)**

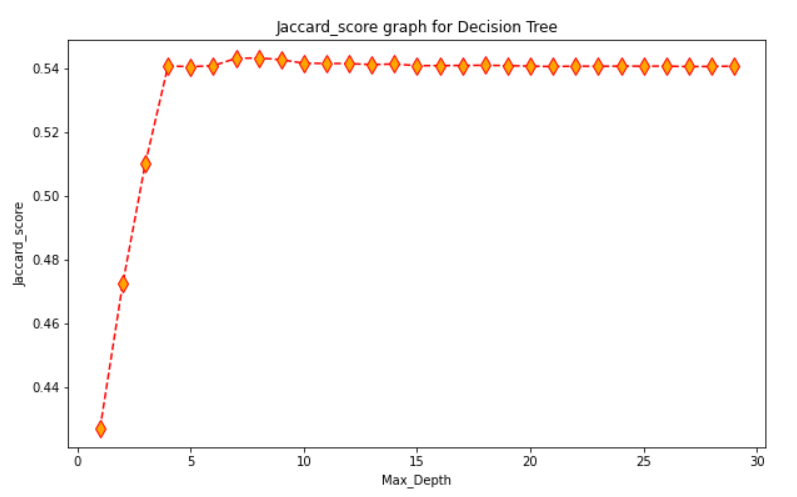
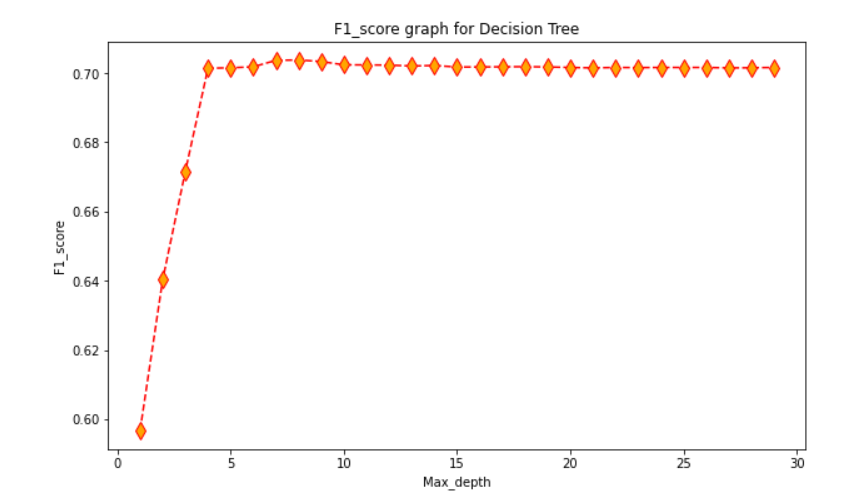
****kNN models seek to categorise the outcome of an unknown data sample based on its proximity in the multi-dimensional hyperspace of the feature set to its “k” nearest neighbours, which have known outcomes. Establishing the value of “k” which optimises the model’s accuracy (between 1 and the total number of samples in the dataset) is an empirical undertaking: if too-few neighbouring datapoints are used, the model is susceptible to being dominated by noise, however if too many neighbours are included in the classification, the model risks losing all diagnostic power completely. kNN models were built for k=1–100 using the kNeighborsClassifier function from scikit learn. The model is optimised at k=80

****

**2. Decision Tree:**

Decision tree models identify the key features on which the data can be partitioned (and the thresholds at which to partition the data) in the hope of arriving, after some iterations, at “leaves” which contain only accidents belonging to one target variable value (in this case, accident severity code).

Decision Trees were made for maxdepth in range for (1,30) to find the depth with maximum accuracy which came out to by at max\_depth=7.

****

**3.** **Logistic Regression Model:**

By converting the accident severity to a binary variable (0 for no/minor injuries, 1 for major injuries/fatalities) we can employ Logistic Regression techniques to attempt to classify accident outcomes based on the properties in the feature set. A Logistic Regression model was trained using an inverse-regularisation strength C=0.01, and tested on the testing subset.

**Model Evaluation**

The models were evaluated on 3 model evaluation scores:

* Jaccard\_similarity score: This score represents the ratio of intesction and union of predicted result and actual result.
* Accuracy score: It is the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions
* F1\_score: It is the Harmonic Sum of precision and recal (where the precision is the number of correctly identified positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number of correctly identified positive results divided by the number of all samples that should have been identified as positive).

|  |  |  |  |
| --- | --- | --- | --- |
| Model/Score | Accuracy | F1\_Score | Jaccard Score |
| KNN | 0.702587 | 0.701023 | 0.540100 |
| Decision Tree | 0.704639 | 0.703682 | 0.543092 |
| Logical Regression | 0.605417 | 0.601163 | 0.431000 |

From the above table we can see that the best performing Model is the Decision Tree which is closely followed by the K Nearest Neighbour model and at last it is the Logical Regression.

**Conclusions and Future Work**

This work highlights that machine learning techniques can be used to probe historical data in order to make reliable predictions about the outcome of road traffic accidents, given information which is available at the time when an accident is reported. The accuracy, F1 scores for the models developed are appreciable, and it can hence be relied upon to help the govt. machineries to reduce the number and severity of accidents. Although the Jaccard similarity score is low this can be changed by extending this project and adding more features and data to train the data set and increasing the accuracy and efficiency of the model.

The prediction of car accident severity is not completely finished. Based on the results, the dataset is under fitted, which means I will need to collect more data model. In addition, the dataset only contains binary data for severity, hence this model has a lot of scope of improvement. Also this project can further be extended and applied to accident databases of other regions/cities. By doing so, city planners can gain insight into the road conditions/features which are associated with high accident severity, and use this insight to improve road design. Additionally, by predicting accident severity as functions of weather, date, location and road conditions, this model may be able to help aid the decision making of emergency services call handlers, by allowing them to prioritise resources toward collisions with a greater likelihood of severe consequences.

**References**

**1.** IBM Data Science Professional Certificate course (<https://www.coursera.org/professional-certificates/ibm-data-science>)

**2.** <https://pbpython.com/categorical-encoding.html>