**FINAL REPORT**

**CAR ACCIDENT PROJECT**

**IBM Data Science Capstone**

**1. Business Problem**

**1.1 Introduction**

When people are driving from one city to another city there is chance of you being struck in the traffic because of an accident. Now the problem is once a vehicle is met with an accident there will be lot of physical and monetary effort that will have to be put into transporting the victims to hospitals and reduce the damage caused to them. In addition to that there will also be a problem with the traffic jam which will create problems to a lot of people in many different ways such as delaying their travel times, increasing pollution and other lot of things. Now if we are having a way to predict the probability of accident taking place and its severity based on different conditions such as road condition, weather conditions and other important parameters. We can inform people and then they can take informed decisions on how to drive carefully. So that is the reason why we can provide a solution to this problem using the following project.

**1.2 Intended Audience**

Normal citizens, police department and civil engineering department scan use the project for their own benefits. For example, the data that is considered for the problem is of the Seattle city, so the people who are travelling in and out of Seattle city as well its residents can use this information for deciding how to drive carefully or let us avoid driving tighter based on the pre-existing conditions. Police, fire and a paramedical team can have a careful preplan for providing immediate support in case if the accidents occur. Civil infrastructure department can use the data to understand and improve the infrastructure and traffic flow patterns. Visualizations and data for the areas where maximum number os accidents are taking place due to various uncontrollable forces during road travel will also be made available to the users.

**2. DATA**

The data that we are going to use for creating this project is downloaded from the link provided in the course itself. Upon inspecting the data, we can see the the sources of the data is Seattle state department of transport. This data has been provided to us in the form of comma separated values or .csv file and it contains 36 features and 194673 rows.

We will clean and remove the unnecessary features and apply the ml algorithms over that data.

This dataset has an attribute that classifies the accident into 5 different categories labelled as a number from 0 to 3. SEVERITYCODE is the target feature that we have to find about. A code that corresponds to the severity of the collision:

• 3- Fatality

• 2b- Serious Injury

• 2- Injury

• 1- Property Damage

• 0- Unknown

We will be extracting the feature data for the major features such as weather, road and light conditions, time of day and location and use this particular feature for modelling the ml problem. We will clean this data, remove null(nan) values and categories that have a very small sample size to avoid the problem of biasing appearing in the problem. Once data extracted and cleaning is completed, we will move ahead with data visualizations through which we can perform the exploratory data analysis part and future understand and processes the available data. Once the final data is present we will apply the ml algorithms of KNN, Decision Tree or Logistic Regression and finally plot the results. We will be finding the parameters of Jaccard, f1 score and log loss for all the 3 algorithms.

**2.1 Data Preparation**

Out of all the available attributes, we are considering to use the features such as roadcond, weather, lightcond, location, x, y, incdttm, incdate and severitycode. These will be the main attributes that we will be using throughout the project. We will remove all the extra features which are not required for the analysis.

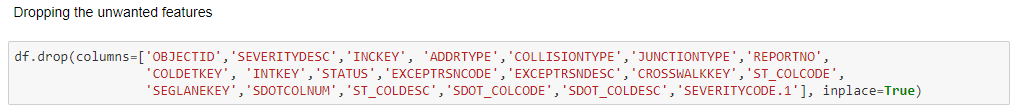


Fig.1. code for dropping the unwanted features from the data set

Then once the unwanted features are removed from the data set, the reaming data will be the main data set that we will be using for the analysis. Now, we will be dropping the null values from all the features that are available.

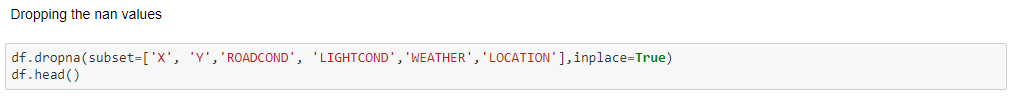


Fig.2. code for dropping the null values from the data set

The next step for us is to clean the data by either eliminating the unknown values or replacing the existing values with the required values and value types.

**2.2 Data Cleaning**

First we will try to homogenizing the data by converting Yes/Y to 1, No/N to 0 and NaN values to 0 for the attribute that are containing this values using the replace function.



Fig.3 data homogenization

Now we will be standardizing the date and time formats across all the data that is available.

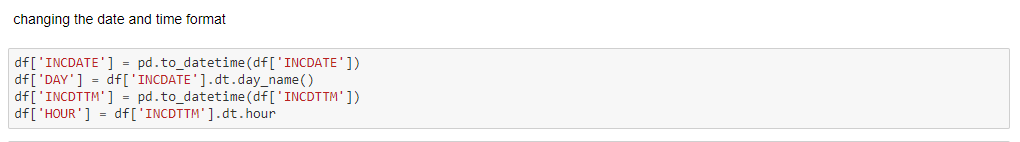


Fig.4 changing the data and time formatting

Now we need to observe the required features and try to identify discrepancies that may be present in the data. First using the count command, I will be trying to find the how many numbers of instances for each type of attribute for a specific feature are present in the data that we contain.

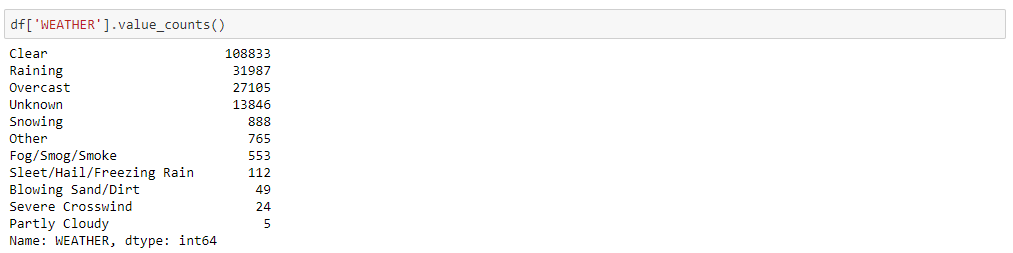


Fig.5 Finding number of instances of each attribute for a specific feature of weather

It can be easily observed that few attributes or let us say categories have very high number of values when compared to other categories. So we need to remove the categories that are containing very less number of instances or data because it will be causing the problem of biasing and also lead to other unintended inaccuracies. Now, we will drop all the attributes with leass number of attributes.

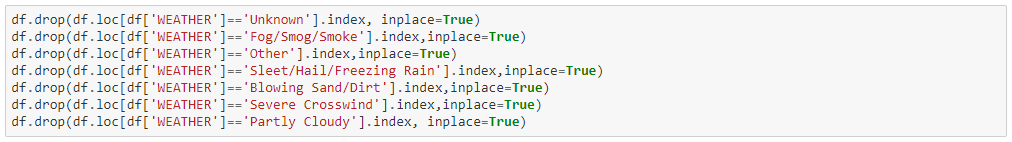


Fig.6 dropping the unwanted categories from the data

Now once again we will be try to observe the data by using the same command of count and try to establish the number of counts.

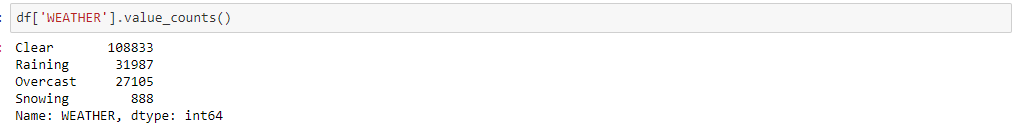


Fig.7 Finding number of instances of each attribute for a specific feature of weather after dropping unwanted attributes

This types of removal of categories was performed on the other attributes such using similar piece of code on the features out on ROADCOND, WEATHER and LIGHTCOND. Now we are create a copy of the main dataset with the required features. We will have to create this copy so as to not to disturb the main raw data which might be required in a unmodified form in case of unexpected conditions. The new data frame has the following attributes: severitycode, x, y, roadcond, weather, lightcond, day, hour and location. In which severity code is the target variable and rest of them are the predictor variables or features.

**3. METHODOLOGIES**

**3.1 Exploratory Data Analysis**

In this step we are going to use visualization capabilities from the matplotlib and seaborn modules present in python to understand the data better and plan the rest of the project that is going to be done.

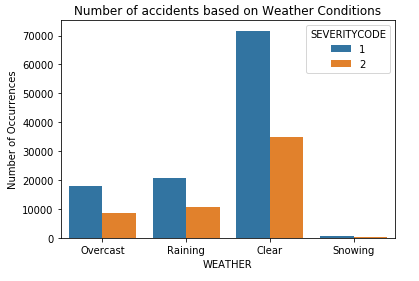


Fig.8 Number of accidents based on weather conditions

From the figure.8 we can easily identify that more number of accidents that more number of accidents are occurring in the clear weather than in the rest of the weathers and even in that also the cases of people getting injured is on the higher side in comparison with the fact of property being damaged. The number of accidents are lower in the snowing weather. Like two other plots were created to plot the number of accidents against the lighting conditions and road conditions. From those two plots it was observed that that more number of accidents that more number of accidents are occurring in the daylight conditions and dry road conditions than in the rest of the weathers and even in these cases also the number of instances of people getting injured is on the higher side in comparison with the fact of property being damaged.

In addition to these plots another 3 plots were created to understand on which days, what particular hour of the day and the top 10 areas in the city where the most accidents are taking place. These plots are shown in the figures.9,10 and 11.

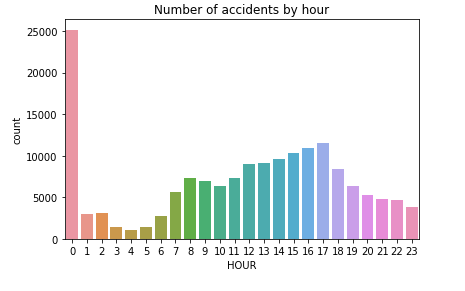


Figure.9 shows how many numbers of accidents take place at a particular hour of a day. From this it can be easily seen that accidents are more at the midnight in comparison with the other times of the day.

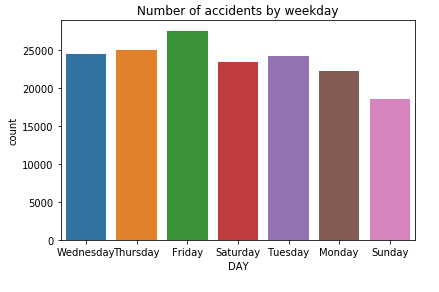


Figure.10 shows how many numbers of accidents take place on which days of the week. From this it can be easily seen that accidents are more on Friday in comparison with the other days of the week.

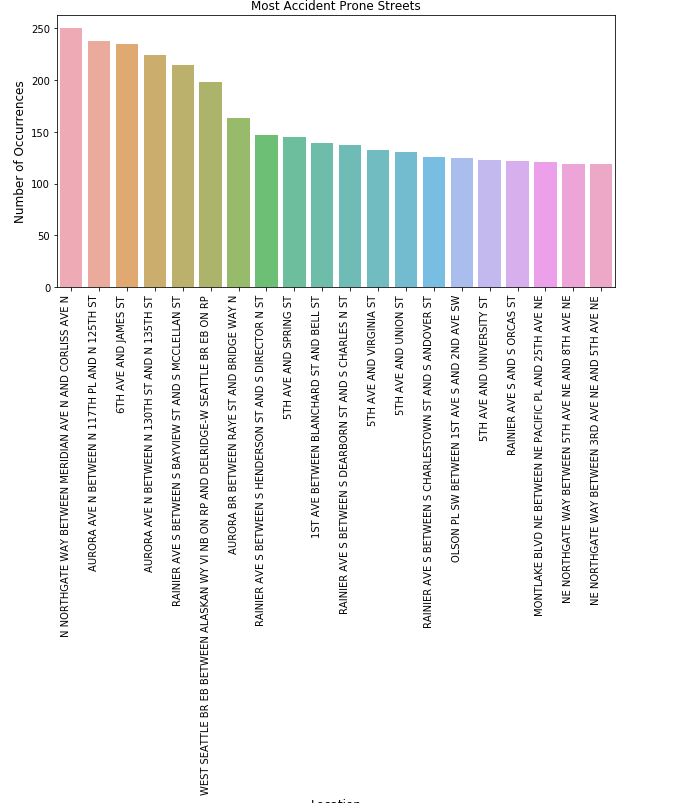


Figure.11 shows how the top 20 areas based on the number of accidents taking place. From this it can be easily seen that accidents are the most at the meridian gate area of the Seattle.

**3.2 ML Modelling**

In order to apply the ML algorithms on the data, the data should be numerical data type. So now we need to use the process of label encoding the change each categorical data into numbers. For the modelling purpose we will make a copy of the dataset. Now we will be removing the category features that are available in the data, as new variables have been created. Now, the data that is finalized is available for applying the ML algorithms. But, if observed the data is unbalanced and this unbalance has to be removed for avoiding to create a biased model.

But before doing predictive modelling using various machine learning algorithms, data imbalance must be ruled out in order to create an unbiased model. We check the value counts for each of the two classification categories for accident severity and observe that the severity index 1 has a sample size twice that of 2.

For balancing the dataset, the process of down sampling can be used. This particular down ranking has be performed using the resample module from the sci-kit learn library. Figure.12 shows the code and the method used for the process of balancing. This process of balancing ensures that the recall and the precision parameter for the model are not skewed.

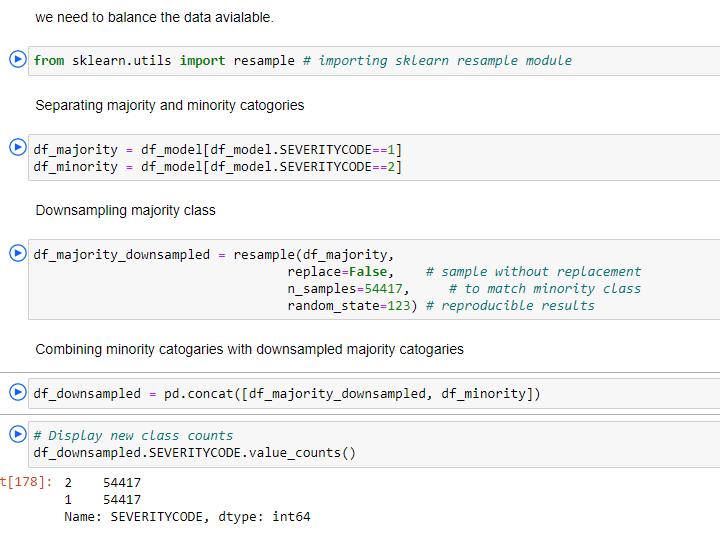


Figure.12 Code for balancing the samples

In the modelling phase of the problem, the first and foremost thing that needs to be performed is the splitting of the total data into training, and test sets. This was done using the scikit learn module.

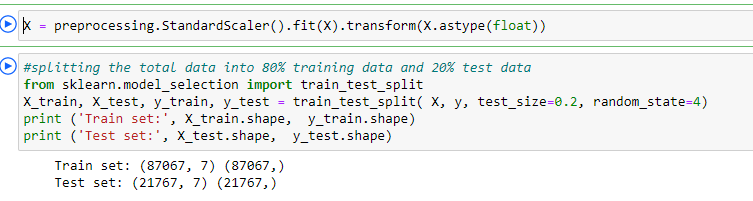


Figure.13 Code for splitting the data

The first type of algorithm that was applied is KNN and the accuracy was calculated for each value of K starting from 1 to 20. The maximum amount of accuracy was obtained for the test set for a k value of 14. Then the decision tree model was to perform analysis on the data. The test size used was 30% and a random state of 3 to split the data was used. The max the depth was defined to be 7 and after calculating trail accuracies with other values ranging between 1 and 20. F1 score and Jaccard score for this model is also evaluated. Finally, the logisitic regression model was implemented and the probability for the values that will be predicted was calculated along with the metric evaluation.

**4. RESULTS AND EVALUATION**

After observing the results, it was observed that the decision tree was the best among the three for application of this particular problem. If we compare the jaccard and the f1-score for all the 3 models, the decision tree model is having the highest value of 0.532174. It indicates that road, weather and light conditions together with location and time of day can be used to predict severity of accident with an accuracy of 0.5321. The accuracy of the model can be future increased by increasing the depth of levels. In addition to that we can also use the logistic regression for this data that is provided. Figure.14, shows the Jaccard, f1-scoere and the Logloss values for the three models.

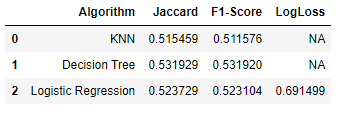


Figure.14, Jaccard, f1-scoere and the logloss values for the three models.

**5. DISCUSSION**

A visualization was created for the observing the effect of the various conditions on the severity of the accidents. From the EDA phase of the project it can be seen that dry, clear skies and daylight conditions which are considered to be normal have caused the greatest number of accidents. So, from this we can come to a conclusion that driver errors are most profound reason for the accidents to occur. But anyway, we tried to predict the severity of the accident if in case all of these conditions or some of these conditions come into play. Analysis for the same could be carried out to understand effect that each attribute has separately on the model and how does all of these attributes considered together affect the model. In order to improve the model we can even add more number of parameters in to the mix such as the day of the week, location and hour of the day also where the accidents are happing in combination with the above considered factors.

**6. CONCLUSION**

The date provided was used to model and predict the severity of the accident based on different conditions such as lighting, road and weather conditions. We used the pre-existing machine learning algorithms of decision tree, KNN and logistic regression for the data and found the metrics of jaccard, F1-score and logloss. The models are able to predict the accident severity in case an accident happens based on the combination of above-mentioned features. This model can also be future improved using techniques available in data science. If this particular model can be implemented into main stream properly, we can save many lives and also will be able to reduce the damage to both public as well as private property.