A REPORT ON CAPSTONE PROJECT

PREDICTING THE SEVERITY OF CAR ACCIDENTS

1. Introduction

1.1 Background

In most cases the major uncontrollable factors for an accident to occur are weather, light and road conditions. Accidents can be mitigated by studying the underlying patterns in the data and giving this invaluable information as a warning and or consideration to the local government, traffic police and the respective drivers of the vehicles. Also this information extracted from the data can also be shared with city planning and urban development construction department who will be taking into consideration the future designs and constructions of the motorways in order to reduce the uncertainty of accidents moreover from the financial point of view the cost and labor involve in these sorts of projects can be managed in an efficient way by the policymakers.

1.2 Problem Statement

Data that commits to determining the severity of accident that includes addresstype, junction type, weather condition, Road and Light conditions. This project aims to predict Severity of accident based on these metrics**.**

1.3 Interest

The target audience of the project is local government, police, rescue groups, car insurance companies and last but not the least the drivers themselves. The model and its results are going to provide some key insights for the target audience that will enables these authorities to leverage on the model in making data-driven decision and reduce the number of accidents and injuries in their localities

2. Data Acquisition and Cleaning

2.1 Data Source

Data provided by the Seattle Department of Transportation (SDOT) on vehicle accidents along with its severity is used to derive insights and patterns on how and when these accidents have taken place with the environmental factors like weather, road conditions etc. The dataset consists of 40 columns having different features of data like, collision severity, road conditions, number of people involved, location of collision, weather etc.

Source:
https://www.coursera.org/learn/applied-data-science-capstone/supplement/Nh5uS/downloading-example-dataset

2.2 Data Cleaning

I have dropped certain columns based on problem statement and understanding of the data. Handling Null/Missing values: There are some null values for ADDRTYPE, JUNCTIONTYPE, WEATHER, ROADCOND and LIGHTCOND attributes which are replaced with value 'others'

2.3 Feature selection

Feature selection/engineering is the process of using domain knowledge to extract features from raw data via data mining techniques. These features can be used to improve the performance of machine learning algorithms.

Based on the problem statement, our predictor or target variable will be 'SEVERITYCODE' because it is used to measure the severity of an accident from 0 to 5 within the dataset. Attributes used to weigh the severity of an accident are 'ADDRTYPE', JUNCTIONTYPE', 'WEATHER', 'ROADCOND' and 'LIGHTCOND'. Regarding exploratory data analysis I thought to choose some of numerical variables such as (PERSONCOUNT, VEHCOUNT, PEDCOUNT, PEDCYLCOUNT)

3. Exploratory Data Analysis

3.1 Getting to know the Target variable

As per the Data collision csv file there are only two types of accidents listed that are severity 1(Prop damage) and severity 2(Injury). So, based on the available data a visual representation of Verified count of accidents by 'SeverityCode' using bar chart is shown below

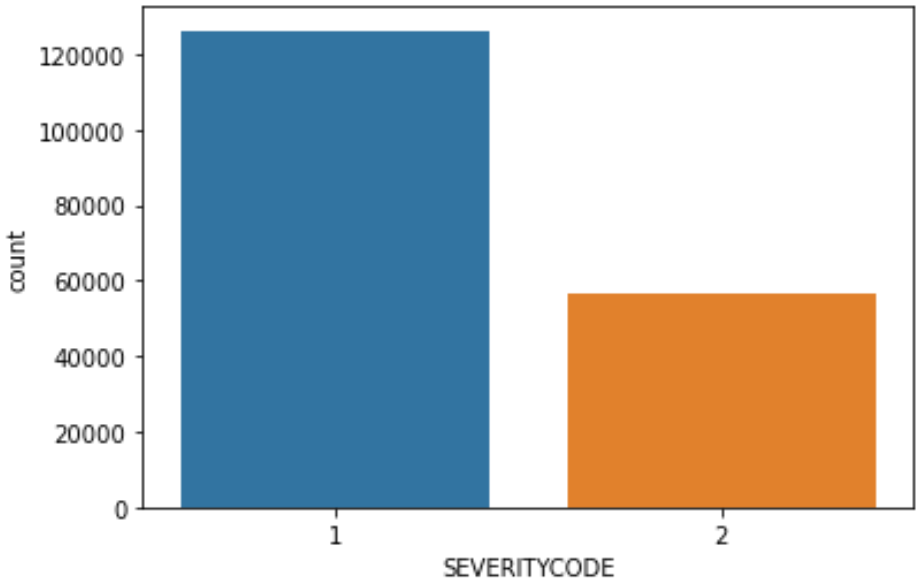

Figure 1

3.2 Analysis of Categorial variables

As from the below figure we can say that majority of the accidents took place in a clear weather condition which is a bit surprising along with dry road condition and in daylight conditions. There are over 10k incidents where road, light and weather conditions were unknown. *(Refer to cell : 45 in the notebook for exact figures)*

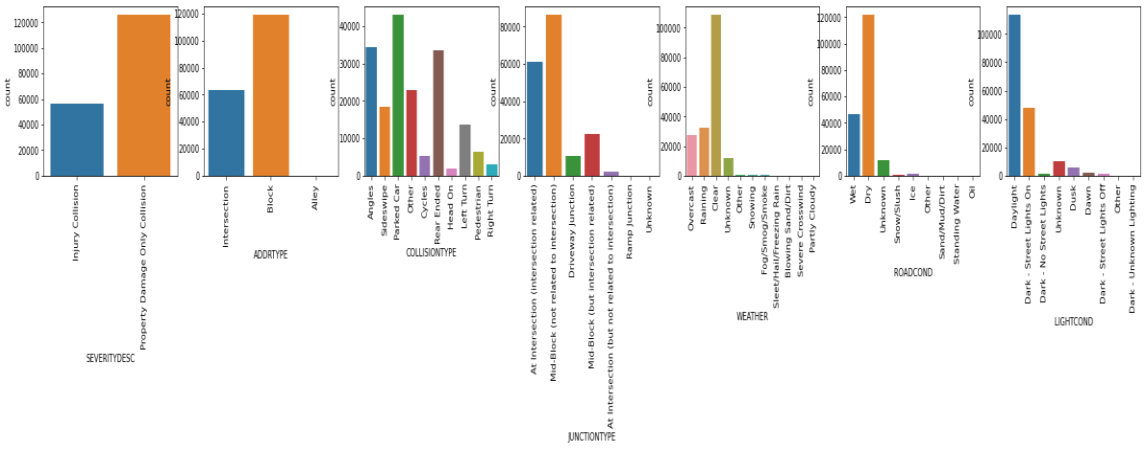

Figure 2

Then I have plotted a heat map to highlight the correlation between different variables, it's a graphical representation of the data helpful to see the interaction among the different variables as we can see from the figure 3 there is a positive correlation between PERSONCOUNT and VECHCOUNT of 0.4 and among the rest of the features we cannot see any significant correlation

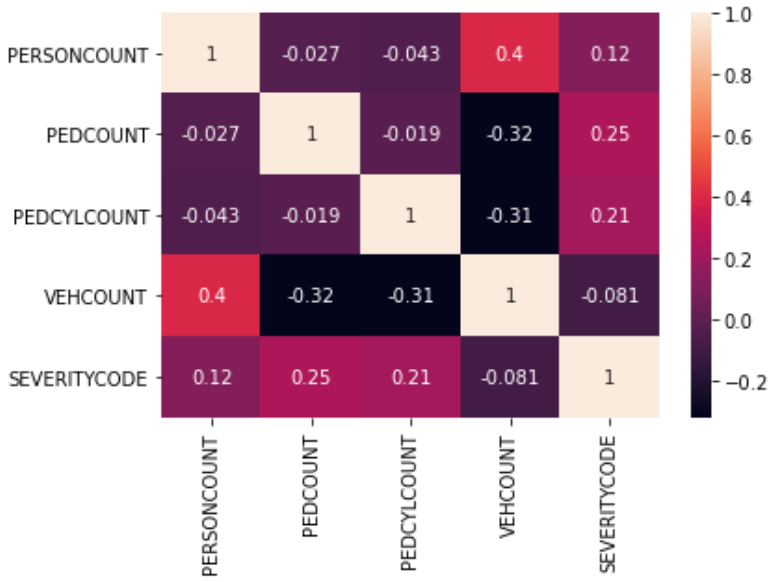

Figure 3

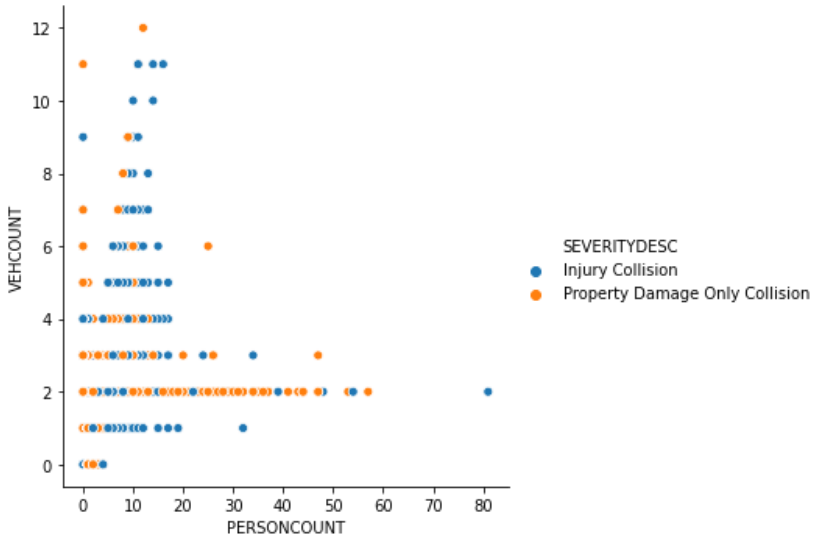

Figure 4

As we can see from the above figure that in majority of the cases the number of vehicles involved were 2 and the severity of accident was mostly related to property damage even the persons involved in the car accident were mostly two persons. *(Refer cell 45 for exact numbers)*

4. Predictive Modeling

Predictive modeling uses statistics to predict outcomes. Most often the event one wants to predict is in the future, but predictive modelling can be applied to any type of unknown event, regardless of when it occurred. Based on exploratory data analysis, we can see that the categorical variables

have key impact in predicting severity in accidents hence I have selected these most important features to predict the severity of accidents in Seattle. So, in our case I have used three predictive modeling techniques which I have learnt during the course.

4.1 K-Nearest Neighbor

K-Nearest Neighbors is an algorithm for supervised learning. Where the data is 'trained' with data points corresponding to their classification. Once a point is to be predicted, it takes into account the 'K' nearest points to it to determine its classification.

After standardization of the training data, which dramatically increases the accuracy the data has been ready for building machine learning models.

Here while implementing KNN algorithm, I have taken it for set of 25 values the model will be trained on training set of data, and then predicting the values based on test data set, At last the model accuracy will be calculated by comparing actual values with predicted values of test data set.

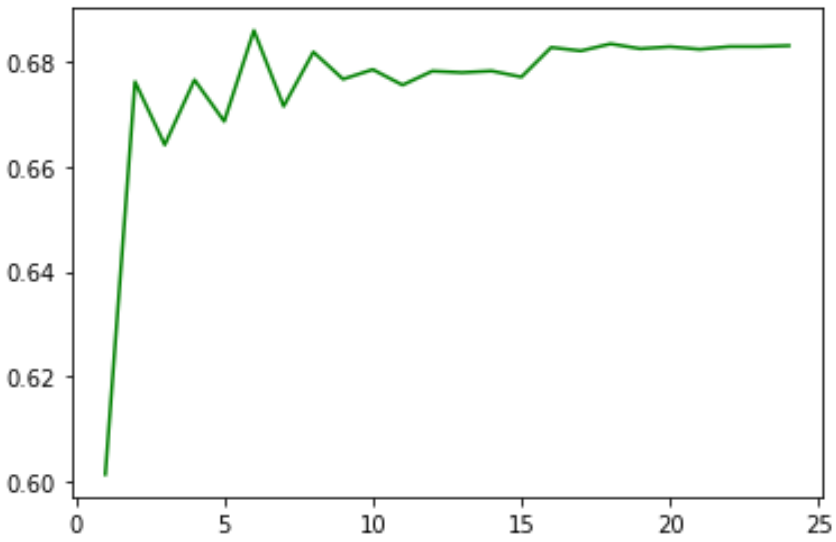

Figure 5

While evaluating model, plotting all 25 accuracy values, so the highest value for k is the best accuracy KNN Model. In this case it is for k=6 we have got maximum accuracy of 0.68

4.2 Decision Tree

Another predictive modeling approach I have used was to incorporate Decision trees into out modeling. Importing the DecisionTreeClassifier module sklearn library and training the model with training data set and making a prediction on the test dataset we have reached an accuracy of 0.69

4.3 Logistic Regression

Logistic Regression is a variation of Linear Regression, useful when the observed dependent variable, y, is categorical. It produces a formula that predicts the probability of the class label as a function of the independent variables.

Logistic regression fits a special s-shaped curve by taking the linear regression and transforming the numeric estimate into a probability with the following function, which is called sigmoid function σ:

After applying the LogisticRegression algorithm the results for this model are shown in the below table.

Model	**Jaccard Score**	**F1 Score**	**Logloss**
Logistic Regression	0.684	0.811	0.59

We have obtained a decent Jaccard score that signifies similarity between the predicted and actual results

F1 score is calculated based on the precision and recall of each class and with this model we have obtained a pretty good F1 score

Log loss measures the performance of a model where the predicted outcome is a probability value between 0 and 1. And our result was not good enough as the value 0.59 is high (lower the better)

5. Results and Evaluation

Once we analyzed and cleaned the data, it was then fed through three ML models; K-Nearest Neighbor, Decision Tree and Logistic Regression. The first two algorithms are ideal for this project, logistic regression did not give any fruitful results with a significant log loss perhaps it requires more feature selection and parameter tuning.

Choosing different k values helped to improve our accuracy to be the best possible.

The results of the all the three model evaluations are summarized as below:

Model	**Accuracy**
KNN	0.68
Decision Tree	0.69

Model	**Jaccard Score**	**F1 Score**	**Logloss**
Logistic Regression	0.684	0.811	0.59

6. Conclusion

I have got a decent accuracy value for all classification algorithms. So, the best classifier of this problem is Decision Tree, KNN and Logistic based on their accuracy value.By revealing hidden patterns in predicting severity in accidents based on the features Weather, Road and Light conditions, addresstype, junctiontype have significant impact on whether to travel or not which often result in injury and property damage.