**Predicting the Severity of Car Accidents using Predictive Model**

IBM Applied Data Science Capstone Project

Ahmad Khairi Bin Ahmad Khir

10 October 2020

**1. Introduction**

**1.1 Background**

Road accidents may not only bring losses to the driver but also to the road infrastructure present at the accident. The inflicted losses which are unbeneficial to both sides can be prevent using machine learning algorithm. Thus, a predictive model constructed using past historical data of car accidents can be used to predict the probability of the severity of an accident even before it is occurred.

**1.2 Problem Statement**

There are several factors that might affect the severity of an accident. Such data included the road condition, weather, light presence, influence of alcohol and speeding. Thus, this project main objective is to predict the severity of an accident based on these parameters.

**1.3 Interest**

The project can prove useful to the local authority to determine the possible location where accidents might happen so that they can take some initiatives to minimizes the destruction of road infrastructure. Others who might have interest in this project might be GPS service provider that can track the condition of the road, weather and also the driver condition through integration with the mobile application.

**2. Data Acquisition and Cleaning**

**2.1 Data Sources**

The datasets used in this project were published by the Seattle Department of Traffic (SDOT), Traffic Management Division which titled “Collision – All years” where all collisions recorded during 2004 to May 2020. The dataset can be downloaded through this [link](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv). The information on the attributes of this datasets were documented in this [metadata](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf) which describe each attribute and its datatype/length. One of the key attributes in this dataset is the severity level tabulated under the column ‘SEVERITYCODE’ which contains two values: ‘2’ for property damage collision and ‘1’ for injury collision. Other attribute such as the accident location, type of intersection, weather details, light exposure, road condition, time, coordinates and number of people involved. This dataset contains 194673 rows of collisions data included with their associated attributes.

**2.2 Data Cleaning**

The main problem existed in this dataset is it contain lot of missing values and duplicate of the same attributes. The attributes had many different values while carrying the same meaning which is redundant to our model predicting capability and will result in inefficiency.

The first step was to remove the unnecessary attributes which are not important to our model such as coordinates, date, pedestrian numbers from the dataset and form a new dataset. The rows that contain missing value from attributes such as road condition, light condition and junction type were removed while missing value in attributes such as alcoholic influence and speeding were renamed to default ‘none’. Finally, the cleaned dataset produce was trained and tested for the modelling stage.

**2.3 Feature Selection**

The attribute which are deemed as related to the severity level were kept as feature and dropping the rest of it. After cleaning of the original dataset, there were 171413 rows of data with 8 features which are reduced from the original dataset that contain 194673 rows of data with 38 features.

|  |  |
| --- | --- |
| Dropped Features | Reason |
| INCKEY, COLDETKEY, REPORTNO, STATUS, INTKEY, LOCATION, EXCEPTRSNCODE, EXCEPTRSNDESC, SEVERITYDESC, SDOT\_COLCODE, SDOT\_COLDESC, INATTENTIONIND, PEDROWNOTGRNT, SDOTCOLNUM, ST\_COLCODE, ST\_COLDESC, SEGLANEKEY, CROSSWALKKET, HITPARKEDCAR | Provides no impact on the model prediction capabilities as most of the features are description of other features or totally unrelated to the accident severity |
| ADDRTYPE, COLLISIONTYPE, INCDATE, INCDTTM, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, | Not impactful enough to be predictor of accident severity |

|  |  |
| --- | --- |
| Features | Value Counts |
| JUNCTIONTYPE – This feature detailed the relationship between the type of junction and the severity of the accident. |  |
| ROADCOND – The condition of the road whether the road is dry or wet in relationship with the accident severity. |  |
| WEATHER- The effect of different weather on the distribution of accidents happened. |  |
| LIGHCOND – The effect of different light condition (exposure) towards the distribution of accident severity. |  |
| SPEEDING – Factor of driver speeding or not in relation with the accident severity. |  |
| UNDERINFL – Factor of driver under the influence of alcohol in relation with the accident severity. |  |
| OBJECTID – For referencing purpose |  |

**2.4 Data Preparation**

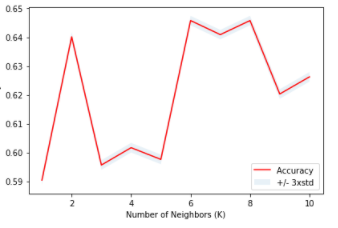
A new dataset was formed using the kept features. This dataset then was converted into a numerical information through one hot encoding to be use in modelling. The data then was normalized through scaled transformation to ensure every feature do not outsized each other when predicting the accident severity value.

**3. Predictive Modelling**

Considering the input data used in this project were categorical in nature, it is obvious that machine learning algorithm that can take in this type of input to be put to use. The supervised machine learning algorithm that were selected were the K-Nearest Neighbor, Decision Tree, Support Vector Machine and Logistic Regression. While each model uses the same training and test dataset, the accuracy of each models was evaluated using the Jaccard Similarity Score, F1 Score and Log Loss. Due to large size of the dataset, the split ratio of train and test were set to 50/50 for optimum accuracy and runtime of the model.

**3.1 K-Nearest Neighbor Model**

In K-Nearest Neighbor model, total of 10 ’k- value’ was used in a loop in order to find which produces the highest accuracy and from the script, it is found that k=6 produce the highest accuracy in predicting the accident severity.



With the value of k=6, the Jaccard Similarity Score and F1 Score were calculated and the result was shown below:



**3.2 Decision Tree Model**

In Decision Tree model, the number of the depth were loop in count of 15 to choose the best max depth for the model. From the script, it is found that max depth of 5 give out the highest accuracy on the accident severity prediction.

With the max depth = 4, the Jaccard Similarity Score and F1 Score were calculated and the result was shown below:



**3.3 Support Vector Machine Model**

The SVM model also were used to make the prediction on the accident severity. The Jaccard Similarity Score and F1 Score were calculated and shown below:



**3.4 Logistic Regression Model**

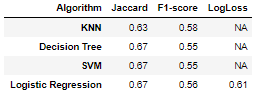
The prediction on the accident severity also make use of the Logistic Regression model. The Jaccard Similarity Score and F1 Score were calculated and shown below:



**3.5 Model Accuracy Result**

* By taking all the accuracy result from each model and tabulate it into single table and make comparison. Based on the table, the best performing model in term of both the Jaccard score and F1-score in correctly predict the severity value is the Logistic Regression model. Even though the KNN model boast the highest F1-score but the lower than other Jaccard score offset the capability of being a reliable model in predicting the value.

.



**4. Discussion**

Based on the result collected, Logistic Regression algorithm performs the best in terms of accuracy and also the fast computation time make it good at observing large dataset. The model ability to produce at most 67% accuracy can be apply in making prediction on the severity of an accident based not only with historical past data but also with real time data. The data on the condition of the weather and light exposure could be easily retrieved from weather monitoring station and the speed of the driver currently on can be easily calculated through the GPS tracking service and the road that had historical occurrences of accident. This live data can be fed to the machine learning model and provide feedback to the driver to alert them on possibility of dangerous driving and chances of getting into a bad accident if no precaution were taken.

**5. Conclusion**

The usage of machine learning model in predicting the severity of an accident can certainly help in reducing the frequency of accident per year. The dataset from the SDOT used in this project recorded of almost 200,000 collisions suggest that there is still long way to go in terms of road safety among the road user. With the help of machine learning, the result generate may prove useful for those that seeks to reduce the numbers of accident happening by recognizing the dangerous location and application developer to develop an application that integrate real time data with existing GPS service to alert the user of danger.

**6. Future Directions**

The existing model can include more fields of data to produce much more cohesive prediction such as inclusion of state of the car and road congestion index.