Car Accident Severity Prediction

1. Background

Nowadays, car crashes have risen to the 8th leading cause of death for people globally. About 1.35 million people are cut short as a result of road traffic accident every year, — 3,700 deaths a day. Between 20 and 50 million more people sustain serious injuries and living with long-term adverse health consequences. Road traffic injuries cause considerable economic losses to individuals, their families, and nations as a whole[1]. Globally, road traffic crashes are a leading cause of death among young people, and the main cause of death among those aged 15–29 years. Road traffic injuries are currently estimated to be the eighth leading cause of death across all age groups globally, and are predicted to become the seventh leading cause of death by 2030.

It was reported that the cost of medical care and productivity losses associated with motor vehicle crash injuries was over 99 billion, or nearly $500, for each licensed driver in the United States (Centers for Disease Control and Prevention, 2010)[2]. These losses arise from the cost of treatment, as well as lost productivity for those killed or disabled by their injuries and for family members who need to take time off work or school to care for the injured. Road traffic accident cost most countries 3% of their gross domestic product [3]. Traffic accidents severity prediction is an important step in accident management. It provides important information for emergency personnel to assess the severity of accidents, assess the potential impact of the accident, and implement effective accident management procedures.

2. Problem

Car accident severity data utilized in this report includes two dependent variables: SEVERITYCODE1 and SEVERITYCODE2. SEVERITYCODE1 represents “Very Low Probablility - Chance or Property Damage”, and SEVERITYCODE 2 represents “Low Probability - Chance of Injury”. The problem becomes the prediction of a car accident severity that can be classified into either SEVERITYCODE1 or SEVERITYCODE2 by utilizing a proper machine learning model.

3. Approach

- Data cleaning

Original data will be re-arranged and all missing/unknown values will be dropped

- Data visualization

Pie chart, bar chart, histogram chart and scatter plot etc. will be plotted to demonstrate a much clearer picture of the dataset

- Data balance

In order to avoid the biased machine learning model, a data balance was performed and the extra data from SEVERITYCODE 1 was randomly chosen and dropped.

- Data modeling

K Nearest Neighbors, Decision Tree, Selected Vector Machine and Logistic Regression algorithms are employed for data modeling

- Model evaluation

Accuracy score, Jaccard, F1-score and LogLoss are calculated to evaluate the models

4. Results & Discussion

4.1 Data cleaning

Data analysis for this project starts with the overview of the dataset and then a data cleaning process was carried out. Original data has 194673 rows and 38 features, among which only strongly correlated attributes were selected such as Collision type, Weather type, Road condition, Light condition and weekend/non-weekend, etc. for model training and prediction. Interestingly, Speeding seems a good feature which usually correlates to car accident severity but the amount of missing values in this feature is so large (185340) so the entire independent variable is dropped. During the data cleaning process, all the missing/unknown numbers were dropped, which reduced the rows from 194673 to 167980. Also, the format of date was converted to a numerical form for dayofweek/weekend which is easier for analysis later.

4.2 Data visualization

4.2.1 Collision types in car accident

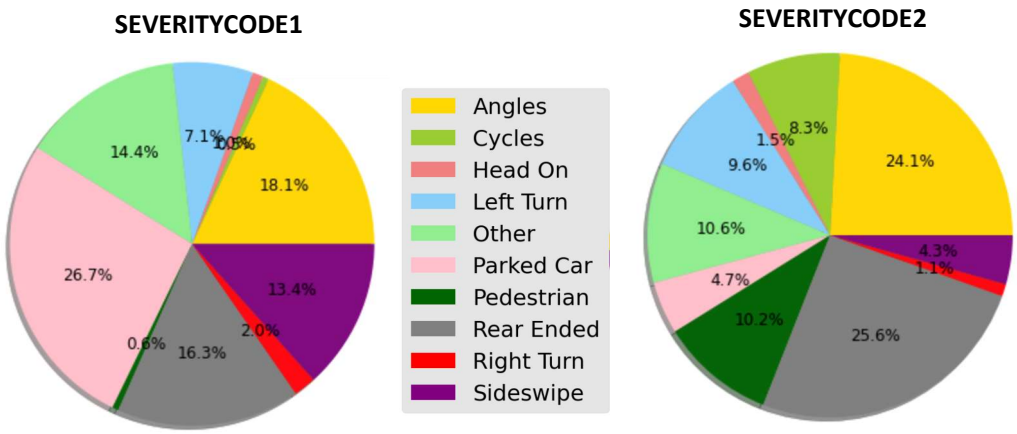

Figure 1. Collision type vs. car accident severity

The collision types are usually directly correlated with the car accident severity. As shown in Fig. 1, the percentages of collision types vary widely in car accident severitycode1 and 2. Most car accidents that result in SEVERITYCODE 1,2 are from parked cars and Rear Ended cars, with 26.7% and 25.6%, respectively. This statistics match well with the common sense that most parked car accidents are less likely to cause severe injuries compared to those caused by rear ended ones. Also, only 0.6% of SEVERITYCODE 1 car accidents involve pedestrians and 10.2% for SEVERITYCODE 2, which means more severe car accidents usually involve pedestrians. From the histogram shown in Fig. 2, it's clear that SEVERITYCODE1 and SEVERITYCODE2 demonstrated different frequency distribution pattern with different collision types, which also can be testified by the box plot, as shown in Fig. 3. The distribution of SEVERITYCODE1 is more continuous from 0 to 30000 while there are two outliners for SEVERITYCODE2 and most of

SEVERITYCODE2 drop below 10000. The median counts number for SEVERITYCODE 1 and 2 are ~12000 and ~5000, respectively, which can also be effected by the unbalanced dataset.

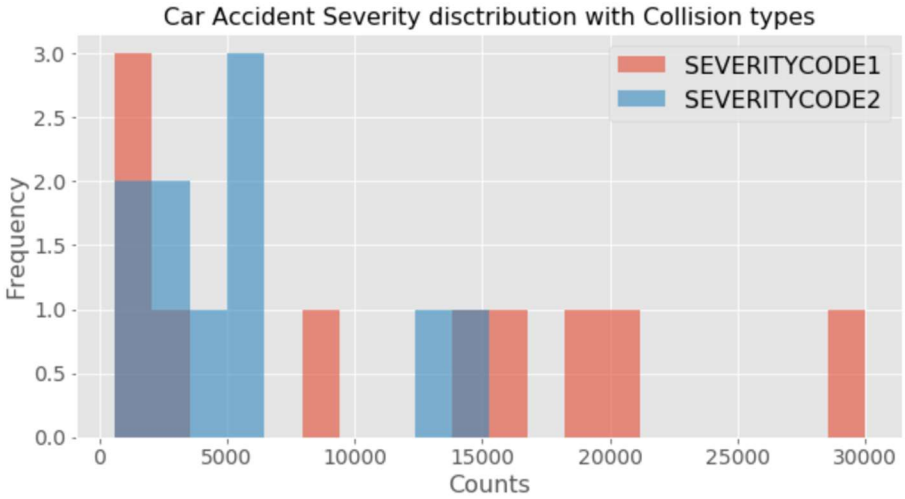

Figure 2. Histogram of SEVERITYCODE1 and 2 with collision types

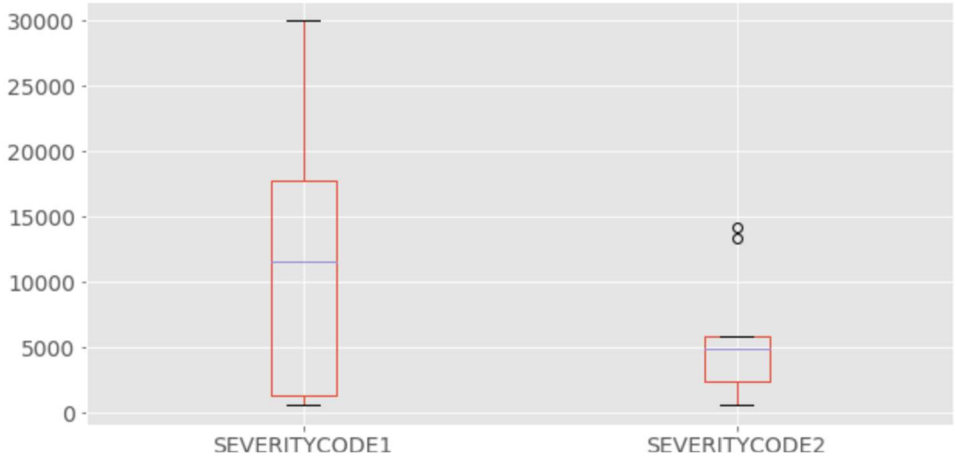

Figure 3. Box plot of SEVERITYCODE1 and 2 with collision types

4.2.2 Number of car involved vs. car accident severity

Both SEVERITYCODE1 and 2 car accidents show similar distribution pattern with the number of car involved, where most SEVERITYCODE1 and 2 accidents involved two cars. Interestingly, the absolute counts of SEVERITYCODE 2 car accidents that involve 1 car is more than that of SEVERITYCODE 1 even the dataset is not balanced where the quantity of SEVERITYCODE 1 accidents is double of that of SEVERITYCODE 2. This result showed that it's more likely that more severe accidents may happen when there is only one car involved in car accidents.

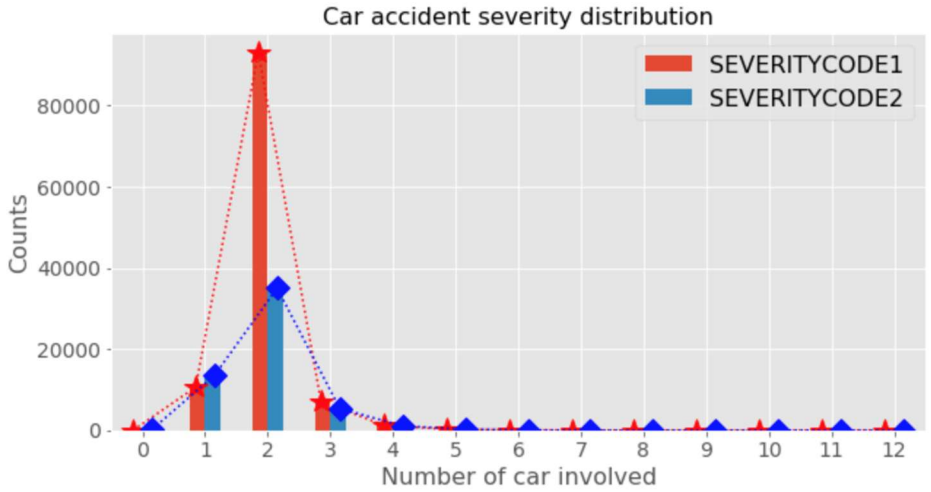

Figure 4. Counts of different severity car accident with the number of car involved

4.2.3 Day of week vs. car accident severity

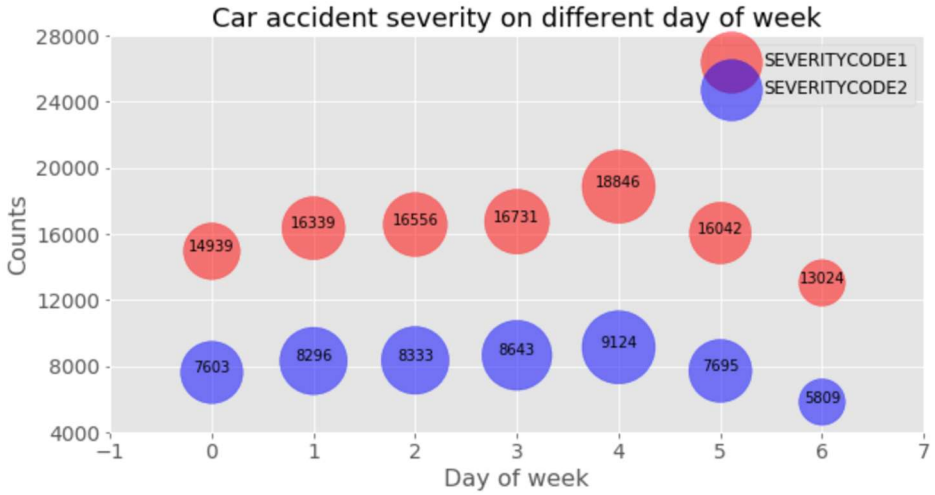

Figure. 5 Counts of car accident with different severity on different day of week

The scatter plot of Fig. 5 shows the counts of SEVERITYCODE 1 and 2 at different day in a week. From Fig.5, the counts for car accidents associated with both SEVERITYCODE 1 and 2 showed the same tendency pattern on different day of week regardless of the imbalanced dataset. From Monday to Thursday, both SEVERITYCODE 1 and 2 showed a slight increase and there is a sharp increase from Thursday to Friday. However, the counts for both SEVERITYCODE 1 and 2 showed a dramatic drop from Friday to Sunday. These results indicate the facts that there is more traffic on Friday since it's the day before weekend while there is much less traffic on road during weekend because of the significantly reduced daily work-home commute.

4.2.4 Weather, light and road conditions vs. car accident severity

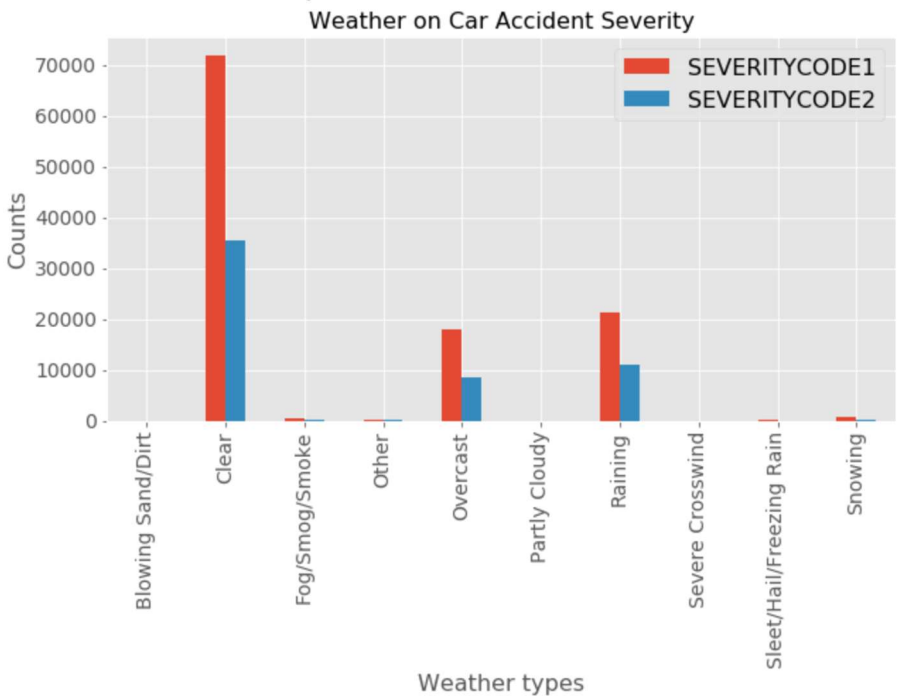

Figure 6. Counts of car accident with different severity on different weather conditions

Normally, weather, light and road conditions are considered as critical factors to determine the severity of a car accident so these three independent variables are included in our machine learning models. The

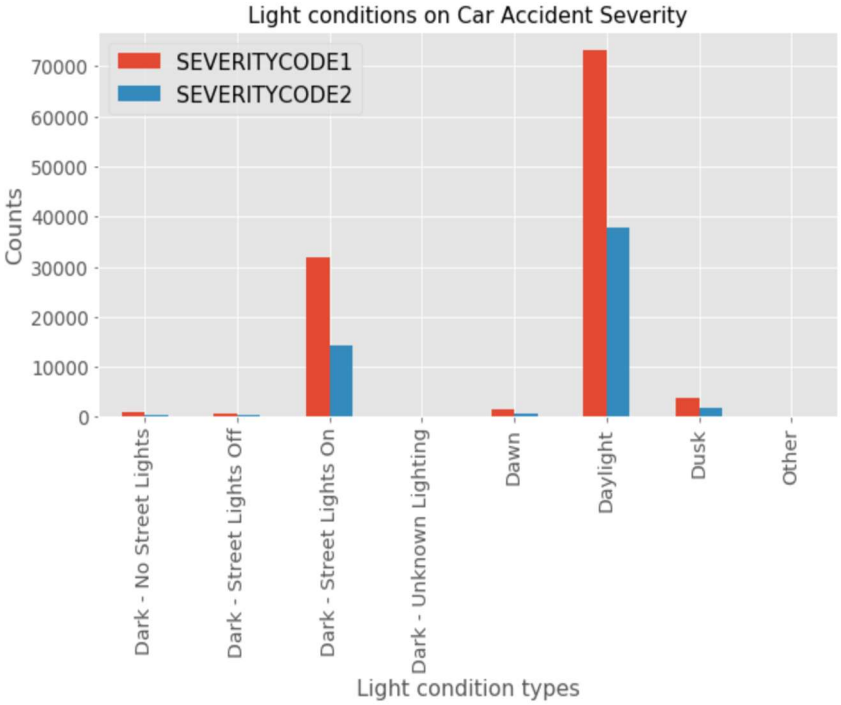

Figure 7. Counts of car accident with different severity on different light conditions

bar charts of Fig 6-Fig 8 show the correlations of the counts of car accidents associated with both SEVERITYCODE1 and SEVERITYCODE2 between the weather, light and road conditions.

The counts of both SEVERITYCODE1 and 2 accidents showed similar distribution pattern under different weather types as showing in Fig. 6. The highest counts are under Clear conditions and followed by Raining and Overcast, which should correspond to the number of days that these weather conditions appear during the period of the data collection. In Fig 7, expectedly, the counts of both SEVERITYCODE1 and 2 accidents under Daylight condition are significantly higher than those under other weather conditions such as Dark, Dawn or Dusk because the majority of the traffic are under Daylight condition. Similarly, for the dark conditions, the counts of both SEVERITYCODE1 and 2 accidents under Dark- Street Lights On are much higher than those under other dark conditions. Finally, as shown in Fig 8, there are much more car accidents associated with SEVERITYCODE1 and 2 recorded under Dry and Wet road conditions than any other types such as Ice, Oil and Standing Water. etc.

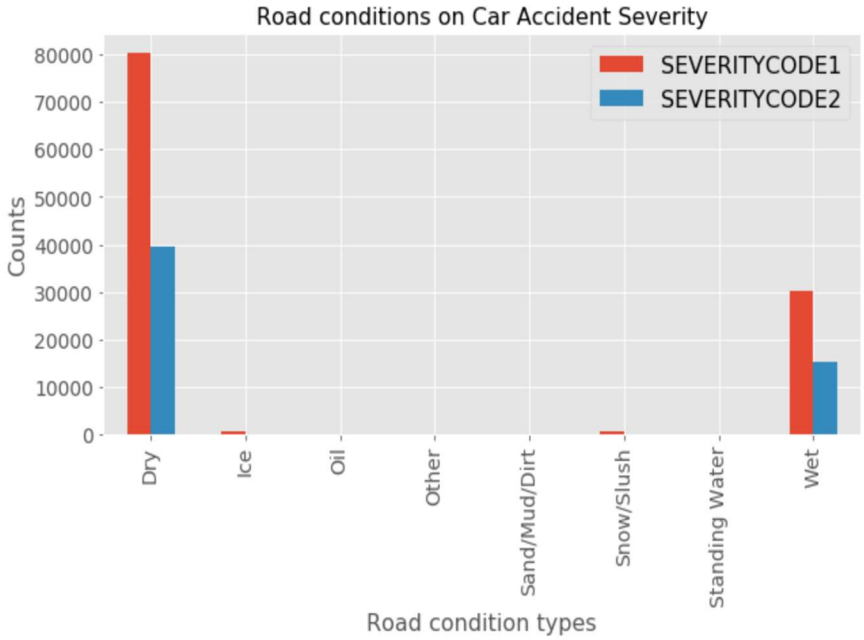

Figure 8. Counts of car accident with different severity on different road conditions

4.3 Data Balance, data type conversion & feature standardization

Before data balance, the data quantities of dependent variable SEVERITYCODE1(11247 rows) is almost double of that of SEVERITYCODE2 (55503 rows), which will lead to a biased model for further prediction. The imbalanced dataset is also reflected by the data visualization presented above, where most of the SEVERITYCODE1 data showed double quantities of that of SEVERITYCODE2. For instance, in light/road/weather condition bar charts, both SEVERITYCODE1 and SEVERITYCODE2 showed similar distribution pattern with the same independent features but the absolute count of SEVERITYCODE1 is almost double of that of SEVERITYCODE2. In order to avoid the biased prediction caused by the imbalanced dataset, data balance is applied to the raw dataset and the randomly selected data from SEVERITYCODE1 results were dropped. After data balance, the data quantities of dependent variable

SEVERITYCODE1(55515 rows) is almost equal to that of SEVERITYCODE2(55503 rows). Followed by data balance, all selected features with categorized characteristics such as the weather types, light conditions and road conditions were converted to numerical variables. Additionally, before modeling, data standardization was performed to normalize the original data.

4.4 Data modeling & evaluation

As discussed previously, different machine learning models and their evaluations were employed. Main algorithms utilized to answer the classification question in this scenario are K-Nearest Neighbors, Decision Trees, Selected Vector Machine and Logistic Regression. Before modelling, dataset was split into train and test category, where 80% of the data was used for training and the rest was used for test.

For K-Nearest Neighbors, the accuracy score of this model with the number of neighbors (K) was plotted to target the best KNN model with an optimized K number. From Fig 9, it's clear that the model shows a highest accuracy score of 0.667 when the K number is 19. Similarly, as showing in Fig 10, SVM models were deployed with different Kernel functions of "sigmoid", "poly", "rbf" and "linear", among which, the SVM model with "rbf" function demonstrated the highest accuracy of 0.63. Interestingly, in Fig 11, the decision tree model with a maximum depth of 5 showed the highest accuracy score of 0.685 while the logistic regression model showed almost the identical accuracy score regardless of C values or functions applied as shown in Fig 12.

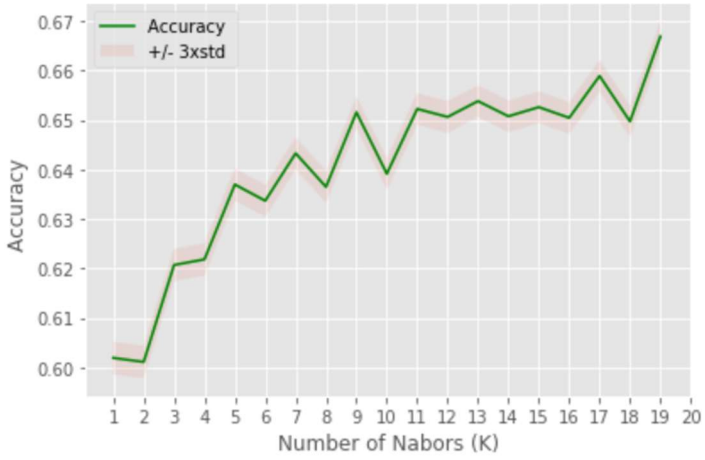

Figure 9. Accuracy score of KNN models with different K values

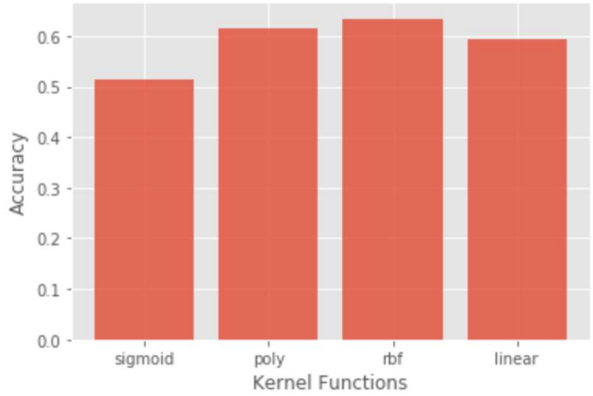

Figure 10. Accuracy score of SVM models with different Kernel Functions

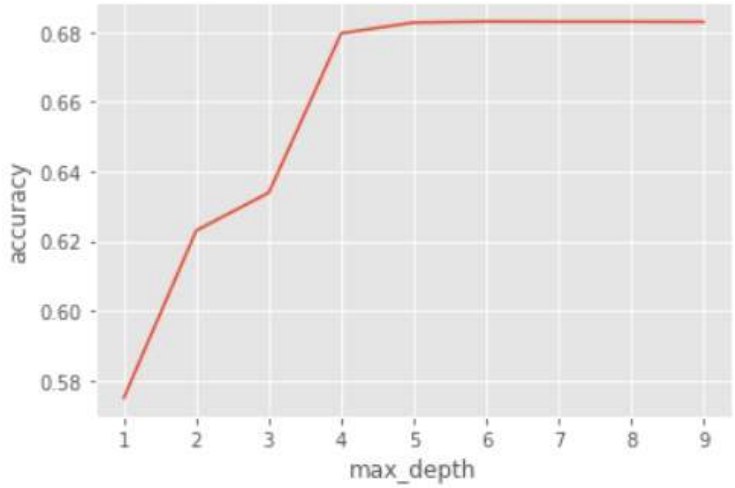

Figure 11. Accuracy score of decision tree models with different max_depth

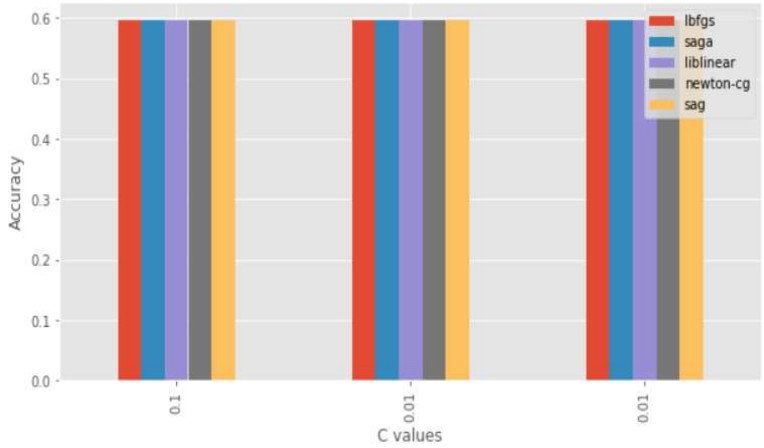

Figure 12. Accuracy score of LogisticRegression models with different C values

Model evaluation was conducted following modeling section. Jaccard, F1-score were calculated for the best performed models from each algorithm and Logloss was calculated for the LogisticRegression model. From these results, it's clear that the Decision Tree model demonstrates the best performance with the highest Jaccard and F1 score of 0.68 and 0.68, respectively, while the SVM model showed the worst performance.

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.67	0.64	NA
Decision Tree	0.68	0.68	NA
SVM	0.59	0.61	NA
LogisticRegression	0.60	0.61	0.67

Figure 13. Summary of evaluation on different models

5. Conclusions

Car accident severity prediction was studied from the original dataset after conducting data cleaning, data visualization and data balance, etc. For this classification question, KNN, Decision Tree, SVM and LogisticRegression were applied to the cleaned data and the models were evaluated. From the evaluation results, it's clear that the Decision Tree model with a max_depth of 5 presents the best prediction performance with both Jaccard and F1-score of 0.68.

6. References

[1] M. Zheng et al., "Traffic Accident's Severity Prediction: A Deep-Learning Approach-Based CNN Network," in IEEE Access, vol. 7, pp. 39897-39910, 2019

[2] F. Zong, H. Xu and H. Zhang, "Prediction for traffic accident severity: Comparing the Bayesian network and regression models", *Math. Problems Eng.*, vol. 2013, no. 3, 2013

[3] Road Traffic Injuries, Jul. 2018, [online] Available: http://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.