Predicting Severity of Traffic Accidents in Seattle

Zi Yi Mok

September 29, 2020

1. Introduction

1.1 Background

Traffic accidents is one of the world's leading cause of death and injuries, on top of causing significant monetary losses due to damages to vehicles and public property. When traffic accidents occur, the severity of an accident usually determines the amount of time needed to clear the scene. Traffic accidents cause traffic jam. When commuters are stuck in traffic, man hours are lost, resulting in loss productivity and ultimately money. When no injuries are involved in an accident, the vehicles are towed away and traffic returns to normal in a reasonable amount of time. Things are not as simple when injuries are involved. Civil defence or firefighters may be needed to extract victims from the vehicles while paramedics are required on scene to treat the injuries before transporting the victims to the hospital, hence more time is taken to return traffic to normal. The severity of a traffic accident, injury versus no injury, thus determines how long it is needed for traffic to return to normal.

1.2 Problem

In a traffic accident, severity of a traffic accident can be influenced by the type of collision, location of the accident, number of people involved, number of vehicles involved etc. External factors such as weather, road conditions and light conditions are factors that contribute to the severity. This project aims to predict if a traffic accident is severe or not, where severity is determined if the accident involves at least an injury.

1.3 Interest

Local governments and city councils would be interested in this study. Firstly, by understanding the factors involved such as road conditions, severe accidents can be prevented. Secondly, when a traffic accident occurs, the city can optimize resources to be deployed on scene based on the predicted severity of the accident. Thirdly, the predicted severity of an accident can also be used as an indicator to gauge the time needed for traffic to return to normal, to plan for road diversions to ease traffic.

2. Data Acquisition and Cleaning

2.1 Data Source

Data used in this study is the collisions dataset provided by the Traffic Records Group, Seattle Department of Transportation (SDOT). The dataset includes all collisions provided by Seattle Police Department from year 2004 to May 2020. Due to system limitation where processing power is unable to process all data, only records from years 2017, 2018 and 2019 were used for this study.

2.2 Data Cleaning

It was fortunate that the data was in a nicely structured dataset. However, it is not without problems.

Before checking for missing values, some features deemed not necessary for predicting severity of traffic accident were dropped. They include unique identifier provided by SDOT, unique identifier of the incidents, unique identifier of the collisions, report number and other identifiers. As all records in this dataset are in Seattle, the coordinates feature was also dropped. Although the exact location where the accident occurred (intersection key, crosswalk key, lane segment) would be a great feature, there were too many unique cases under this feature, and hence was also dropped from the study.

Date and time variables were also provided in the dataset, however, while the dates are all complete, the time feature is not. It is generally perceived that time of the day could be an important factor in predicting traffic accidents and hence its severity, there are several other features provided in the dataset that imply the conditions at that moment, such as light conditions, road conditions and weather.

Several features provided in the dataset also points to the same thing. They include a code given to the collision by SDOT and the description of the code. The description is dropped from the study. There is also a code given to the collision by the state provided together with another variable of its description. The description is also dropped.

In terms of missing values, the variables for weather, road conditions and light conditions have significant numbers of missing values. We consider these as crucial predictors of traffic accident severity and hence decided to delete records with missing values for these 3 variables, instead of replacing them.

There were three indicator variables with missing values, i.e. whether or not the pedestrian right of way was not granted, Whether or not speeding was a factor in the collision, and whether or not collision was due to inattention. These variables only have entries for true cases and the missing values were replaced with false indicators.

Another indicator variable showing whether or not a driver involved was under the influence of drugs or alcohol has 4 unique entries, namely ‘Y’, ‘N’, ‘1’ and ‘0’. ‘Y’ and ‘1’ were grouped as positive cases while ‘N’ and ‘0’ were grouped as negative cases.

2.3 Feature Selection

The correlation among variables were studied and there were no 2 variables found to be highly correlated (Pearson correlation >0.9) and hence no features were dropped based on this method.

The final dataset contains 28,218 records with the following 17 predictors.

Feature	Description
ADDRTYPE	Collision address type: • Alley • Block • Intersection
COLLISIONTYPE	Collision type
PERSONCOUNT	The total number of people involved in the collision
PEDCOUNT	The number of pedestrians involved in the collision
PEDCYLCOUNT	The number of bicycles involved in the collision
VEHCOUNT	The number of vehicles involved in the collision
JUNCTIONTYPE	Category of junction at which collision took place
SDOT_COLCODE	A code given to the collision by SDOT
INATTENTIONIND	Whether or not collision was due to inattention
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol
WEATHER	A description of the weather conditions during the time of the collision
ROADCOND	The condition of the road during the collision
LIGHTCOND	The light conditions during the collision
PEDROWNOTGRNT	Whether or not the pedestrian right of way was not granted
SPEEDING	Whether or not speeding was a factor in the collision
ST_COLCODE	A code provided by the state that describes the collision
HITPARKEDCAR	Whether or not the collision involved hitting a parked car

Table 1

Among the 28,218 cases, 9,539 are considered severe cases while 18,679 are considered non-severe cases.

3. Exploratory Data Analysis

3.1 Relationship between target variable SEVERITYCODE and categorial variables

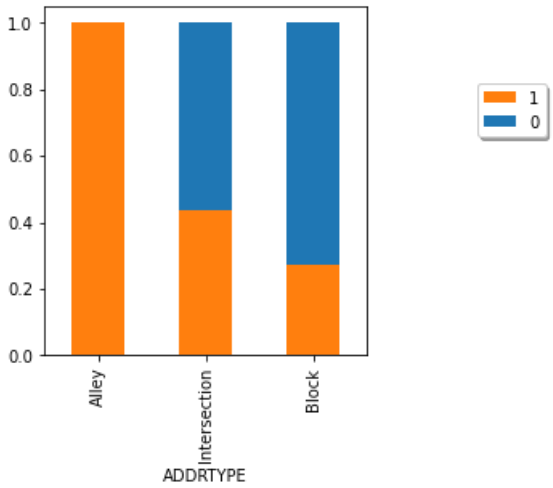

Figure 1

Figure 1 above shows the percentage of severe cases by address types of the collision. It can be seen that all collisions the happened in alleys were severe cases. It is however noted that only there were only 3 collisions in alleys. Between intersections and blocks, collisions at intersections have higher chances of being severe cases.

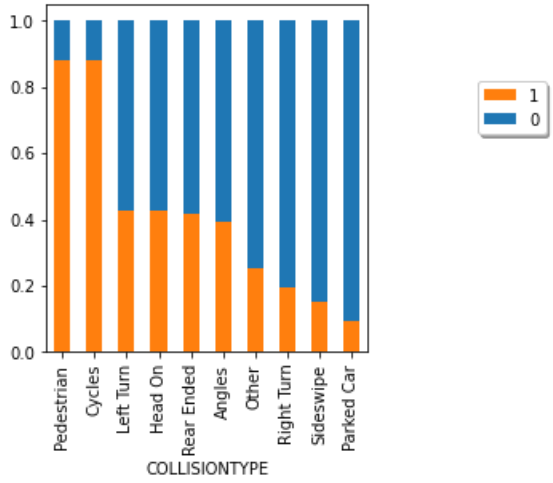

Figure 2

Figure 2 shows the percentage of severe cases by collision types. Collisions that involved pedestrians and cycles have far more percentages of severe cases.

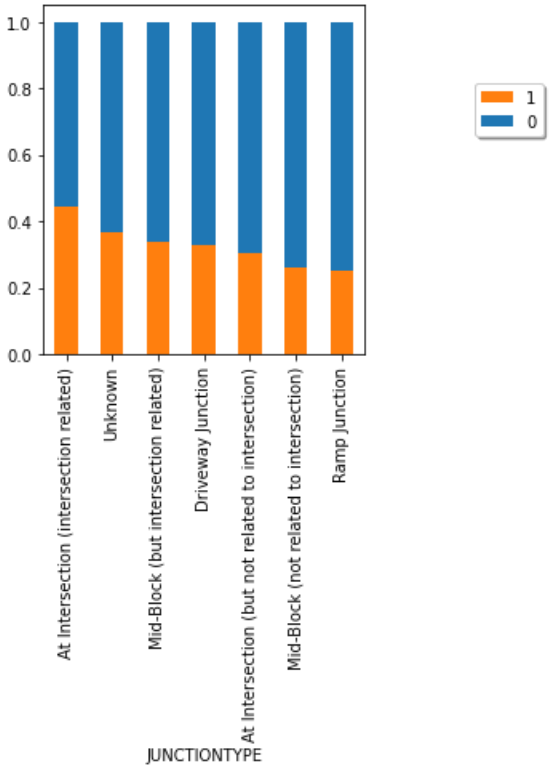

Figure 3

Figure 3 shows collisions at intersections have higher percentages of severe cases.

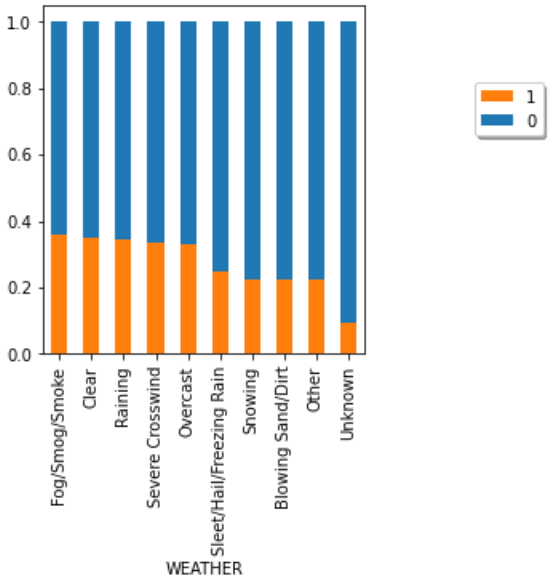

Figure 4

Figure 4 shows that cases tend to be more severe when the weather is foggy, smoggy or smoky. However, it is surprising that collisions during clear weather have higher percentages of severe cases than those during rain or snow. It could be that people tend to drive more carefully during rain or snow, while they tend to let their guard down during clear weather.

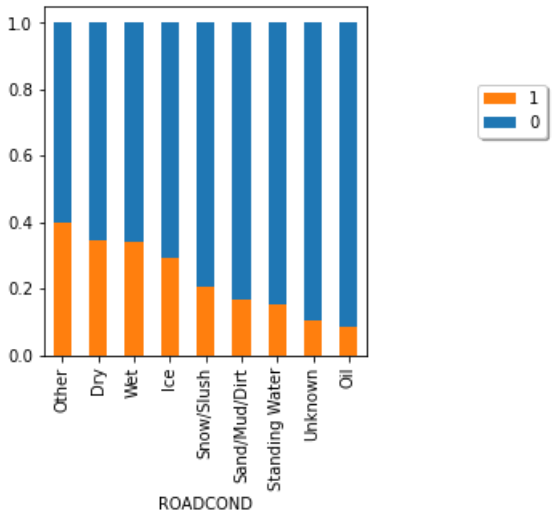

Figure 5

Figure 5 shows another surprising relationship. Accidents tend to be more severe when road conditions are dry instead of when road conditions are wet, icy or oily. Once again, it could be that people let their guard down when conditions are favourable.

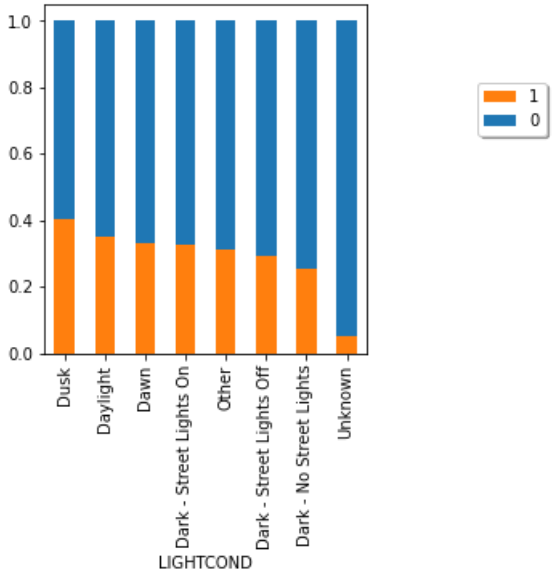

Figure 6

Figure 6 shows the percentage of severe cases by light condition. It is seen that there is a higher percentage of severe cases during dusk.

3.2 Correlation between numerical variables

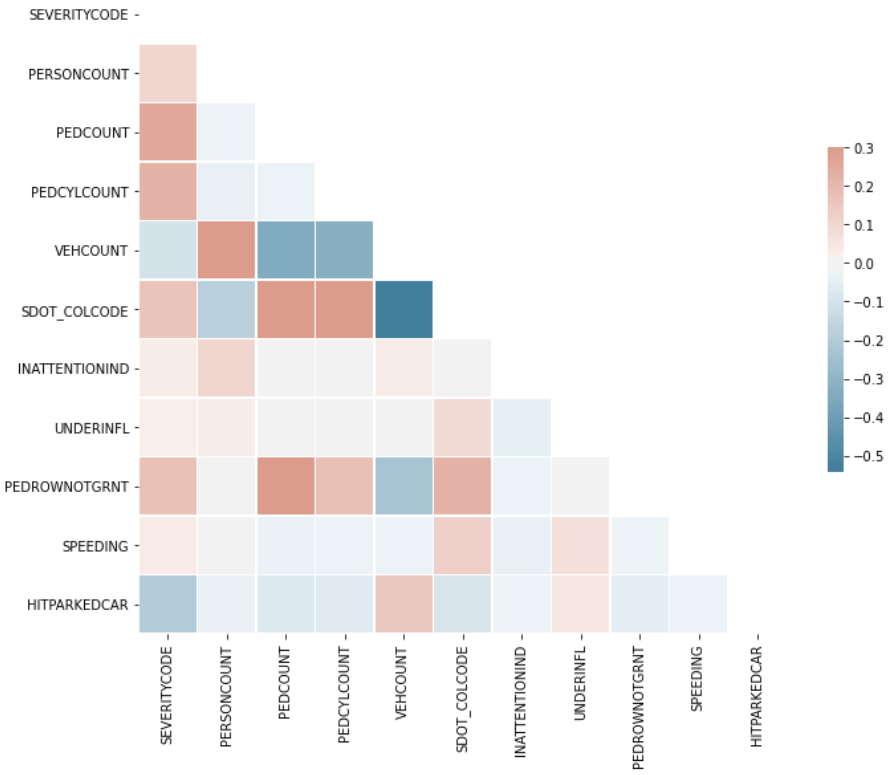

Figure 7

Figure 7 shows the correlation among all numerical variables. There are no variables that are highly correlated, hence no multicollinearity.

4. Predictive Modelling

This problem is a classification problem, where the target variable SEVERITYCODE is a binary variable. 4 models were trained and evaluated based on the testing set. The 4 models were K Nearest Neighbour Model (KNN), Decision Tree Model (Tree), Support Vector Machine Model (SVM) and Logistic Regression Model (LR). The results are shown in the next section.

5. Results

The model prediction on the test set was compared against the actual severity codes and evaluated based on accuracy, precision, recall, F1-score, Jaccard score and Area under the Curve (AUC).

	KNN	Tree	SVM	LR
Accuracy	0.72	0.72	0.72	**0.73**
Precision	0.64	**0.85**	0.79	0.73
Recall	**0.37**	0.21	0.24	0.32
F1-Score	0.69	0.66	0.67	**0.70**
Jaccard Score	**0.31**	0.20	0.23	0.29
AUC	0.72	0.75	0.73	**0.77**

Table 2

Table 2 above summarizes the evaluation outcome of the models. Models with the best score in the respective metrics are in bold. KNN performs the best in terms of recall and jaccard score. Tree performs the best in terms of precision. LR performs the best in accuracy, F1-socre and AUC.

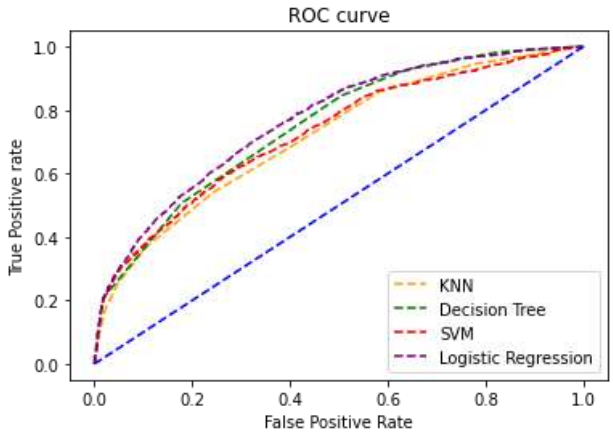

Figure 8

Figure 8 shows the ROC curve comparing all the 4 models. Consistent with AUC metric, the Logistic Regression model performs better than the other 3 models.

Based on the above metrics, it is recommended that the Logistic Regression model is the best classifier model to predict severity of traffic accident cases in Seattle.

6. Conclusion

In this study, the severity of traffic accidents and factors that influence it were analysed. A classification model was built to predict if a traffic accident involves an injury or not based on several factors such as weather conditions, location of collision, angle of collision and etc. The model that was selected is built based on logistic regression. The model could be useful in preventing severe accidents, optimizing resources to be deployed on scene based on the predicted severity of the accident, and gauging the time needed for traffic to return to normal in order to plan for road diversions to ease traffic.

It is however noted that prediction modelling is a reiterative process. There is still room for improvement to achieve higher accuracy in predicting severity of traffic accidents.