Analysis of severity of traffic accidents

Coursera Applied Data Science Capstone

October 25, 2020

Contents

1 Introduction

Driving is an essential activity for a lot of Americans. People may drive to work, or drive to buy grocery, or drive to visit friends and relatives and so on. However, traffic accidents may happen occasionally. Sometimes traffic accidents may lead to property damages. But sometimes traffic accidents may lead to bodily injuries and even fatalities.

While the severities of traffic accidents vary a lot, it is helpful to identify some risk factors that may make traffic accidents relatively more severe. If the risk factors can be identified, it will be helpful to many people. The risk factors can be served as early warning signals to remind drivers that they should pay even more attention when they drive. If they drive more carefully, it can help the community as a whole to significantly reduce the traffic accidents and the severity of losses.

2 Data

In order to identify the potential risk factors that are correlated to severities of traffic accidents, I am going to use the collisions data published by the City of Seattle. There is an open data platform and the data is updated in a regular basis. Also, the source is the government so it can be considered as pretty reliable.

According to the attribute information provided, the data set covers all types of collisions from 2004 until now. The data set provides a variety of information, including severity of the traffic accidents, location of the traffic accidents, light conditions during the accidents, whether the driver involved was under the influence of drugs or alcohol etc.

3 Exploratory analysis / inferential statistical testing

Based on the data set provided, there are about 200,000 rows and 38 columns. It implies that there are about 200,000 traffic accidents records. Also, there are 38 attributes or features that can be used to do the prediction or modeling.

3.1 Target variable

As the purpose of the project is to determine the severity of the traffic accidents, the target variable is the “SEVERITYCODE”. Based on the data set provided, it is located in the first column and it is a binary variable. The values of this variable can be “1”, meaning property damages are caused by the traffic accidents. Another value of this variable can be “2”, which means bodily injuries are involved in the accidents. So it can be considered that accidents with “2” in the “SEVERITYCODE” are generally more severe.

Given it is the target variable that we would like to predict, it is important to perform inspection of data and assess whether there is material data quality issue. If there was material data issue, data cleaning would be considered to enhance the overall data quality as appropriate.

All records either have “1” or “2” in the “SEVERITYCODE” column. There is no record with weird value like NA, blanks, or digits not consistent with the attribute information document. Although it is stated that “SEVERITYCODE” may contain “2b” and “3” and these are not found in the current data set provided, I presume that the data set provided has been modified in a way to broadly split all records into 2 categories (i.e. property damages only, versus, people are injured). This modification can be regarded as a reasonable simplification, without defeating the original purpose of identifying the risk factors that may lead to more severe accidents.

The distribution is shown in the table below.

SEVERITYCODE	Count	% of Count
1	136,485	70.1%
2	58,188	29.9%
Total	**194,673**	**100.0%**

Broadly speaking, about 70% of the records involve property damages only, while the remaining 30% involve bodily injury. While there are more accidents with “1” in the “SEVERITYCODE”, both values in column “SEVERITYCODE” seem to have ample data for analysis.

3.2 Features

After the target variable has been identified, the next step is to identify the features or attributes that may potentially cause more severe traffic accidents.

3.2.1 Location of traffic accidents

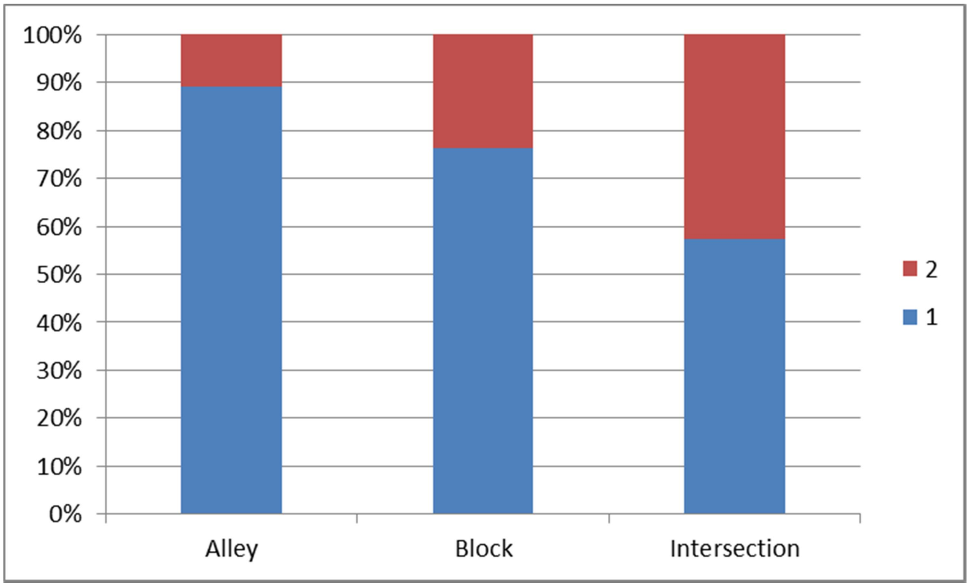
The first variable that may potentially be useful is the location of the accidents, which is the column with header “ADDRTYPE”. If there is a significant correlation between the location and the severity of the accidents, it means that the drivers should pay more attention when they are driving through such locations.

The distribution is shown in the table below.

ADDRTYPE	1	2	Grand Total
Alley	669	82	751
Block	96,830	30,096	126,926
Intersection	37,251	27,819	65,070
(blank)	1,735	191	1,926
Total	**136,485**	**58,188**	**194,673**

There are 4 possible values in the “ADDRTYPE” column, including Alley, Block, Intersection and blank. The records with blank value do not contain much information. Furthermore, there are relatively few records with blank in the “ADDRTYPE” column. Excluding these records is not expected to affect the modeling results significantly.

Analysis has been done to understand the relationship between the location and the severity of the accidents. As shown in the chart below, accidents at the intersection seem to be more severe than alley and block in general.



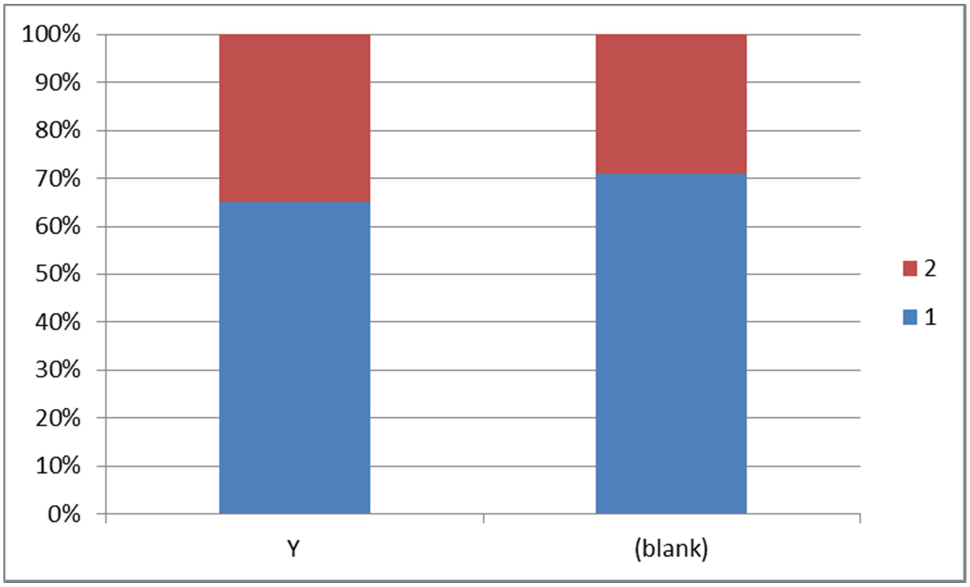
3.2.2 Whether the collision was due to inattention

There is a column with header “INATTENTIONIND”. According to the attribute information document, it is an indicator whether or not collision was due to inattention.

The distribution of “INATTENTIONIND” is shown in the table below. It is presumed that “Y” implies that the collision was due to inattention, while blank implies that the collision was not due to inattention. There does not seem to be any material data quality issue.

INATTENTIONIND	1	2	Grand Total
Y	19,408	10,397	29,805
(blank)	117,077	47,791	164,868
Total	**136,485**	**58,188**	**194,673**

The relationship between “INATTENTIONIND” and the severity of the accidents are illustrated in the chart. It may indicate that if the accident is caused by inattention, the severities of the accidents seem to be higher.



3.2.3 Whether the driver involved was under the influence of drugs or alcohol

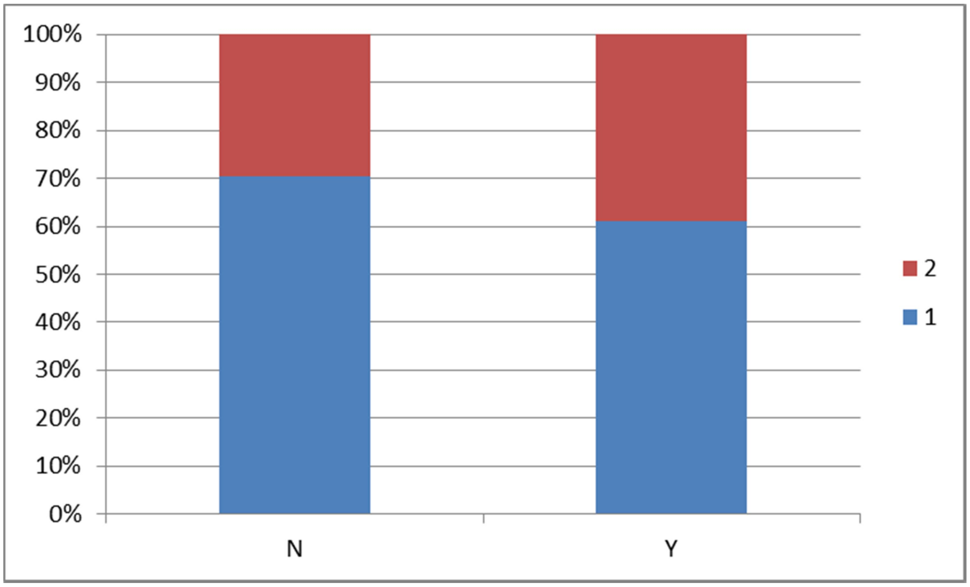
There is a column with header "UNDERINFL". According to the attribute information document, it is an indicator whether or not a driver involved was under the influence of drugs or alcohol.

The distribution of "UNDERINFL" is shown in the table below.

UNDERINFL	1	2	Total
0	57,693	22,701	80,394
1	2,372	1,623	3,995
N	69,378	30,896	100,274
Y	3,187	1,939	5,126
(blank)	3,855	1,029	4,884
Total	**136,485**	**58,188**	**194,673**

It seems that there are two data issues. Firstly, it is supposed to be an indicator so that this field should be binary and thus it should not contain 5 types of variables (i.e. "0","1","N","Y", blank). For blank, it is unclear whether it means the driver is not under influence of drugs or alcohol, or, it means that the information is unavailable. Therefore, it is appropriate to exclude those records with blank value. Excluding the blank records, it is noted that there are relatively more records with "0" and "N", and much fewer records with "1" and "Y". It is likely that the meaning of "0" and "N" is the same. Similarly, it is likely that "1" and "Y" carry the same meaning. Therefore, it is reasonable to perform data cleaning to treat "0" and "N" the same, and similarly treat "1" and "Y" the same.

After the data cleaning, it signals that if the driver was under influence of drugs or alcohol, then it was more likely that the accidents would be severe.



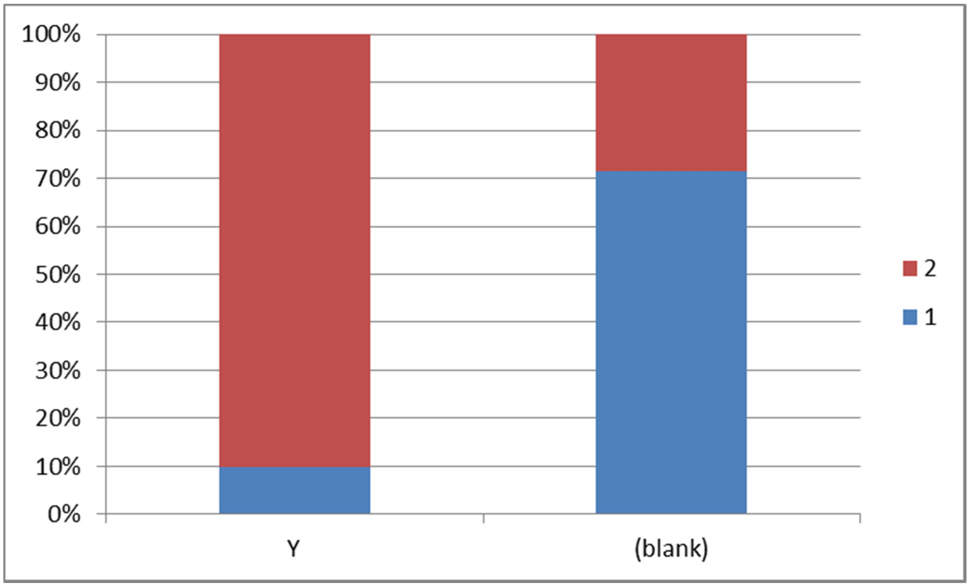
3.2.4 Whether the pedestrian right of way was not granted

There is a column with header “PEDROWNOTGRNT”. According to the attribute information document, it is an indicator whether or not the pedestrian right of way was not granted.

The distribution of “PEDROWNOTGRNT” is shown in the table below. There does not seem to be any material data quality issue.

PEDROWNOTGRNT	1	2	Grand Total
Y	460	4,207	4,667
(blank)	136,025	53,981	190,006
Total	**136,485**	**58,188**	**194,673**

The relationship between “PEDROWNOTGRNT” and the severity of the accidents are illustrated in the chart. It may indicate that if the pedestrian right of way was not granted, it was very likely that the severities of the accidents would be high.



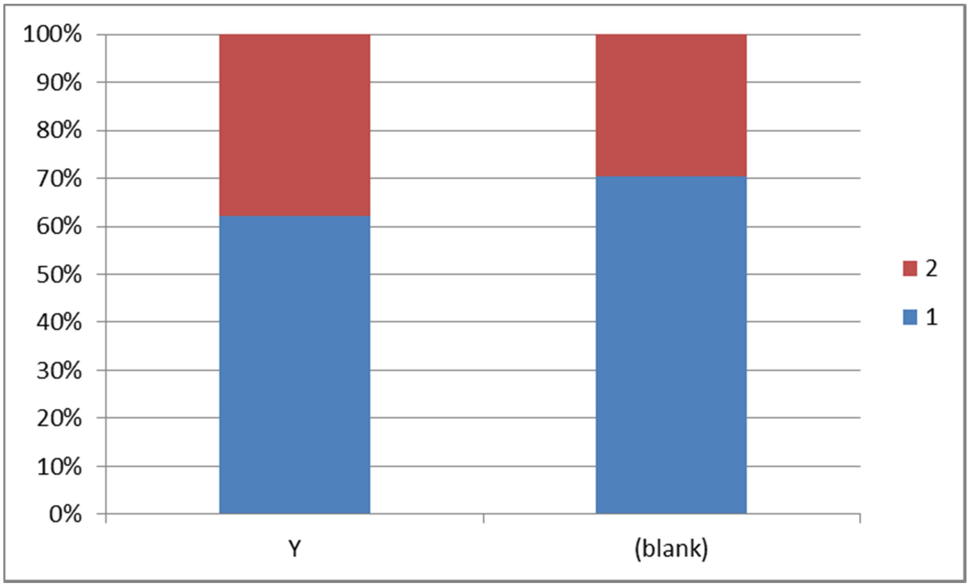
3.2.5 Whether speeding was a factor in the collision

There is a column with header “SPEEDING”. According to the attribute information document, it is an indicator whether speeding was a factor in the collision.

The distribution of “SPEEDING” is shown in the table below. There does not seem to be any material data quality issue.

SPEEDING	1	2	Grand Total
Y	5,802	3,531	9,333
(blank)	130,683	54,657	185,340
Total	**136,485**	**58,188**	**194,673**

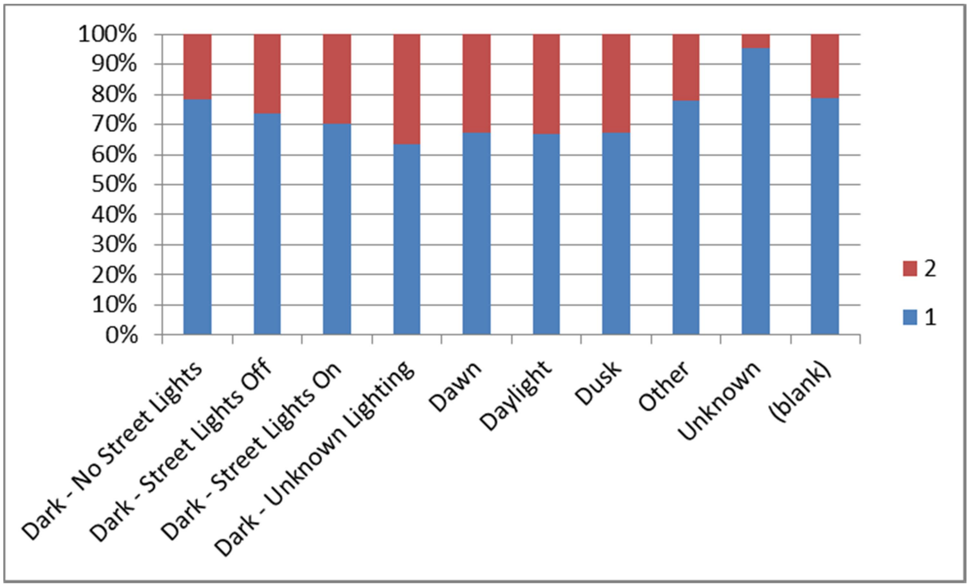
The relationship between “SPEEDING” and the severity of the accidents are illustrated in the chart. Intuitively, accidents due to speeding may lead to higher severity. The data seem to be in line with the intuition. The accidents with “Y” in the “SPEEDING” column have relatively more severe cases.



3.2.6 The light conditions during the collision

There is a column with header "LIGHTCOND". According to the attribute information document, it is the light conditions during the collision.

The distribution of "LIGHTCOND" is shown in the table below. There are some records that do not have meaningful light condition information for prediction of results, for instance "Other", "Unknown", "Dark – Unknown Lighting", blank. These records constitute less than 10% of the total records. It is expected that removing these records will not materially affect the results. Therefore, these records will be removed as part of the data cleaning process.



4 Machine learning approach

Broadly speaking there are two kinds of machine learning approaches, namely, regression and classification.

Regression is typically used to predict an outcome with continuous values. For example, if data is available, regression models may be used to the medical cost and the monetary value of property damage.

In contrast, classification is usually used to predict an outcome with discrete classes or categories. For example, depending on the availability of data, classification models may be used to predict whether there is fatality in the accident.

Based on the data provided, the focus is to predict whether there is bodily injury involved in the accident. So, it is more a classification problem and thus classification machine learning approach is used.

4.1 Data with categorical nature

It is noted that information for "Location of traffic accidents" (i.e. Alley, Block, Intersection) and "The light conditions during the collision" (e.g. dark, dawn, daylight) are in discrete categories.

To handle the categorical features, binary feature for each class is created. One-hot encoding is performed with the functionality of pandas. Additional columns like "ADDRTYPE_Alley", "ADDRTYPE_Block", "ADDRTYPE_Intersection" are created in the dataset.

4.2 Training data and testing data

There are almost 200,000 records in the raw data in total. To evaluate the performance of various models, the data is split into 75% for training and the remaining 25% for testing.

5 Results and discussion

Testing has been performed on a few classification models, including logistic regression, decision tree, random forest and support vector machines.

There are a number of metrics to evaluate the performance of models. The metrics include

1. Accuracy

   It is determined as $\frac{TP+TN}{TP+TN+FP+FN}$, where

   - TP is the number of True Positives, where prediction is positive and observation is positive indeed
   - TN is the number of True Negatives, where prediction is negative and observation is negative indeed
   - FP is the number of False Positives, where prediction is positive but the observation is actually negative

- FN is the number of False Negatives, where prediction is negative but the observation is actually positive

Accuracy is a metric that is very straight-forward. It measures the number of correct prediction, regardless of the class of prediction and observation.

2. Precision

   It is determined as $\frac{TP}{TP+FP}$

   Precision is a measure of proportion that when the prediction is positive, the actual observation is really positive.

3. Recall

   It is determined as $\frac{TP}{TP+FN}$

   Recall is a measure of ability to give a positive prediction when the actual observation is positive.

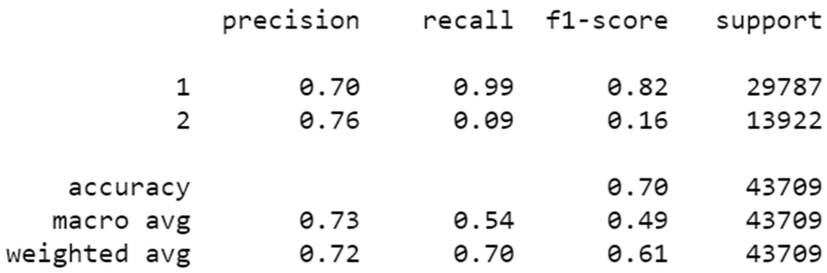
4. F1-score

   It is determined as $\frac{2 \times Precision \times Recall}{Precision+Recall}$

   F1-score is a balance between precision and recall.

5.1.1 Logistic regression

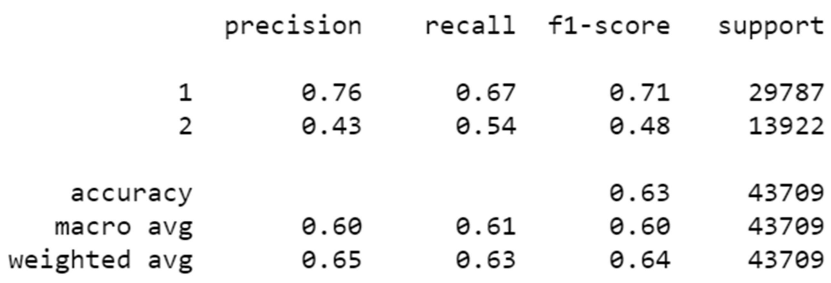
The various performance metrics under the logistic regression are shown as below.



```
              precision    recall  f1-score   support

           1       0.70      0.99      0.82     29787
           2       0.76      0.09      0.16     13922

    accuracy                           0.70     43709
   macro avg       0.73      0.54      0.49     43709
weighted avg       0.72      0.70      0.61     43709
```

It is noted that while the f1-score of class 1 (i.e. property damage) is quite high. The f1-score of class 2 (i.e. bodily injury) is quite low. The average of f1-score is 0.49, which signals room for improvement. The low f1-score of class 2 can be due to the fact that there are more actual observations in class 1 and the data is not as balanced as desired.

To handle the problem of imbalanced classes, the class weight in the logistic regression is set to “balanced” and the model is re-run.

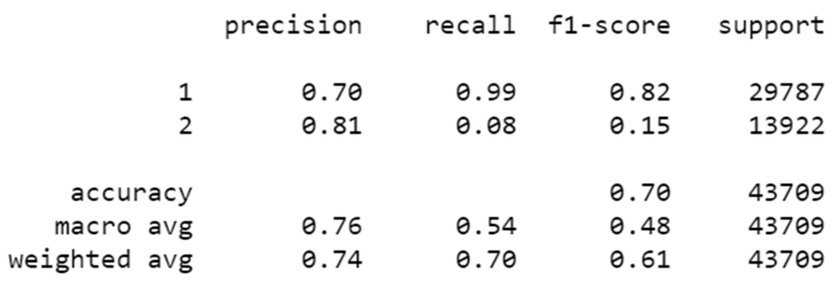


```
              precision    recall  f1-score   support

           1       0.76      0.67      0.71     29787
           2       0.43      0.54      0.48     13922

    accuracy                           0.63     43709
   macro avg       0.60      0.61      0.60     43709
weighted avg       0.65      0.63      0.64     43709
```

Comparing the performance metrics with the original model, f1-score of class 1 dropped a bit but the f1-score of class 2 improved quite significantly. The average f1-score improved to 0.60 as well. Therefore, it is considered that the model with class weight parameter of "balanced" gives better results.

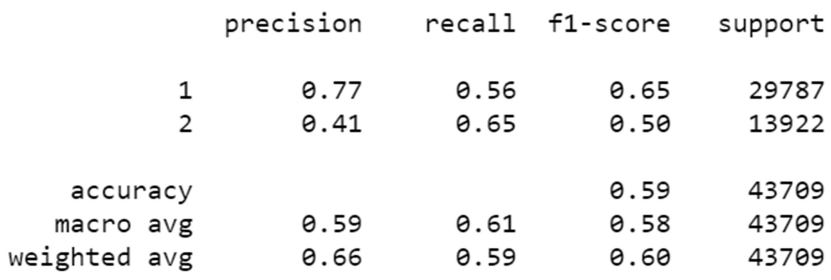
5.1.2 Decision tree

The various performance metrics under the decision tree classifier (without class weight parameter of "balanced") are shown as below.



```
              precision    recall  f1-score   support

           1       0.70      0.99      0.82     29787
           2       0.81      0.08      0.15     13922

    accuracy                           0.70     43709
   macro avg       0.76      0.54      0.48     43709
weighted avg       0.74      0.70      0.61     43709
```

The various performance metrics under the decision tree classifier (with class weight parameter of "balanced") are shown as below.

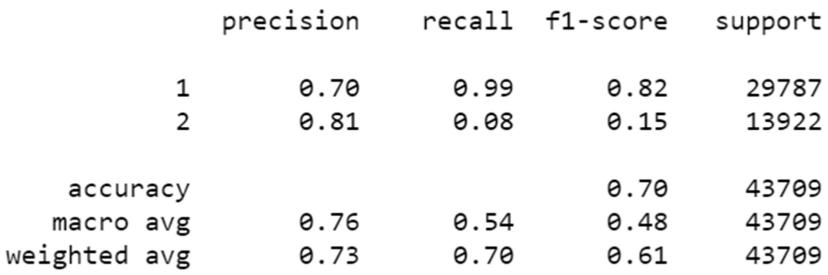


```
              precision    recall  f1-score   support

           1       0.77      0.56      0.65     29787
           2       0.41      0.65      0.50     13922

    accuracy                           0.59     43709
   macro avg       0.59      0.61      0.58     43709
weighted avg       0.66      0.59      0.60     43709
```

Similar pattern is observed when the performance metrics with and without the balanced class weight setting are compared. With a balanced class weight, the average f1-score is higher (0.58 vs 0.48).

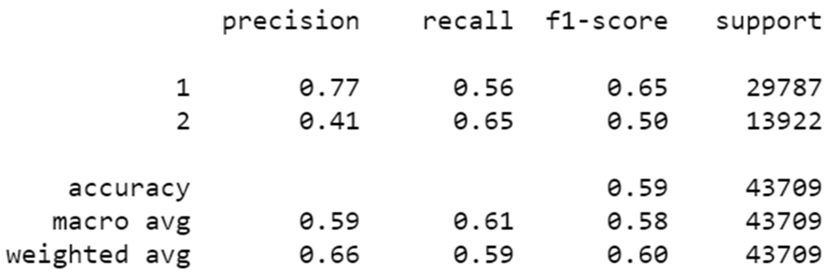
5.1.3 Random forest

The various performance metrics under the random forest classifier (without class weight parameter of “balanced”) are shown as below.



```
              precision    recall  f1-score   support

           1       0.70      0.99      0.82     29787
           2       0.81      0.08      0.15     13922

    accuracy                           0.70     43709
   macro avg       0.76      0.54      0.48     43709
weighted avg       0.73      0.70      0.61     43709
```

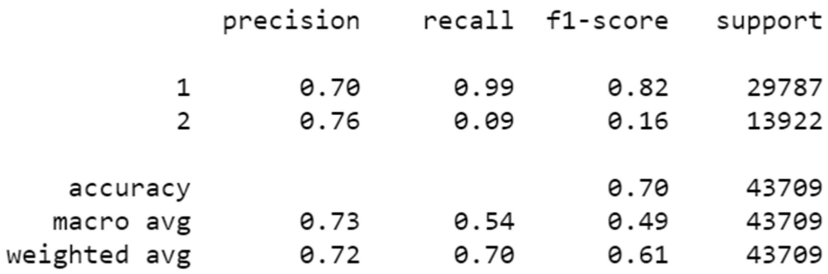
The various performance metrics under the random forest classifier (with class weight parameter of “balanced”) are shown as below.



```
              precision    recall  f1-score   support

           1       0.77      0.56      0.65     29787
           2       0.41      0.65      0.50     13922

    accuracy                           0.59     43709
   macro avg       0.59      0.61      0.58     43709
weighted avg       0.66      0.59      0.60     43709
```

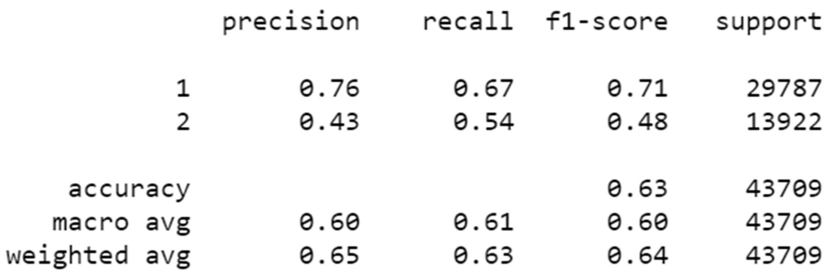
5.1.4 Support Vector Machines

The various performance metrics under the support vector machines (without class weight parameter of “balanced”) are shown as below.



```
              precision    recall  f1-score   support

           1       0.70      0.99      0.82     29787
           2       0.76      0.09      0.16     13922

    accuracy                           0.70     43709
   macro avg       0.73      0.54      0.49     43709
weighted avg       0.72      0.70      0.61     43709
```

The various performance metrics under the support vector machines (with class weight parameter of “balanced”) are shown as below.



```
              precision    recall  f1-score   support

           1       0.76      0.67      0.71     29787
           2       0.43      0.54      0.48     13922

    accuracy                           0.63     43709
   macro avg       0.60      0.61      0.60     43709
weighted avg       0.65      0.63      0.64     43709
```

5.1.5 Comparison of models

The performance metrics for different models are shown as below. Logistic regression model and support vector machines model give relatively better results in terms of average f1-score and accuracy.

Model	Class	F1-score	Accuracy	Precision	Recall
Logistic regression	1	0.60	0.63	0.76	0.67
	2			0.43	0.54
Decision tree	1	0.58	0.59	0.77	0.56
	2			0.41	0.65
Random forest	1	0.58	0.59	0.77	0.56
	2			0.41	0.65
Support vector machines	1	0.60	0.63	0.76	0.67
	2			0.43	0.54

Generally speaking, the accuracies of various models are about 0.6. There is room for improvement. Usually the performance of model can be enhanced with more data. Therefore, it is recommended that more data are collected to further sharpen the model accuracy.

6 Conclusion

This project is intended to evaluate the risk factors that may lead to more severe traffic accidents. Based on the data, it seems that several factors may factor the severity of the accidents, including location of accidents, light conditions, whether the accident was due to inattention of driver, whether the driver was under influence of drug or alcohol, whether the pedestrian right was granted and whether there was speeding.

Several machine learning models were used to predict the severity of traffic accidents based on the data provided. Logistic regression and support vector machines models generate relatively better results. To further improve the quality of model prediction, it is recommended to collect more data.