Car Accident Severity Report

Yushan Sun

Introduction

Nowadays, people have witnessed and recorded many car collisions. Facing with so much collisions data, data scientists come up with new perspective to these data. Based on these past records, people are likely to gain the insights into the data, and to explore these possible factors which influence the happen of collisions. Therefore people want to know if some of the factors such as weather, road condition, light condition and speeding can influence and involve in the collisions. And do these factors influence the severity of the collisions. More important, scientists want to use these evidence to build a series of reliable models to predict how much these conditions can influence the collisions. So that people are capable of decreasing the possibility of car accident based on the weather or road condition.

Data

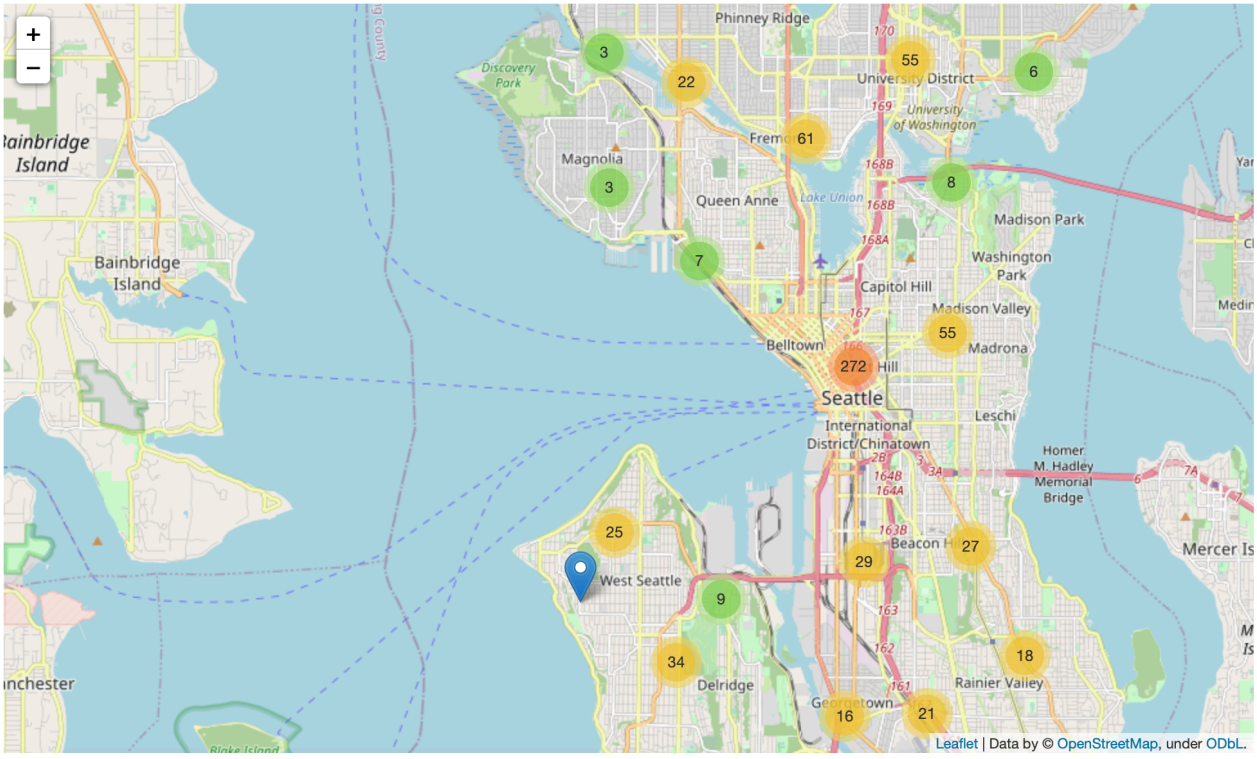
For this research, I used the data set from 'Collisions—All Years' table. This table includes all collisions provided by SPD and recorded by Traffic Records in Seattle region from 2004 to present. The table includes over one hundred ninety-four thousand rows and thirty-eight columns. It includes the explicit location and location type of collisions, the severity of the collision, the date and time of the collision, the total number of people involved in the collision, and also the weather, road condition, light condition, etc. The table also labeled whether the collision was due to inattention and whether the driver was under the influence of drugs or alcohol, which means we can exclude the human factors to search for the environment factors.

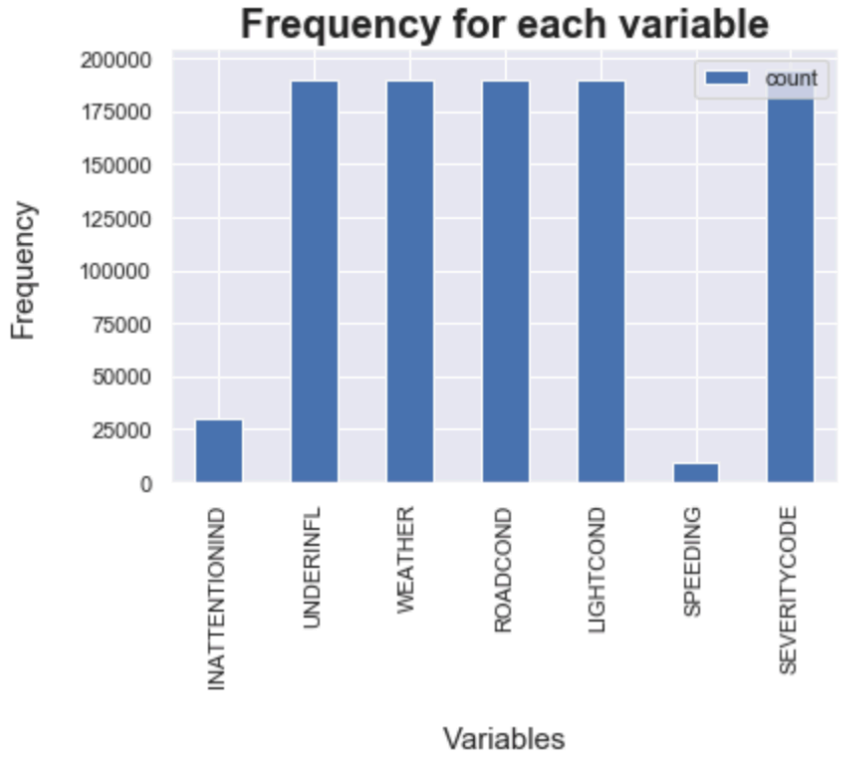
	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet	Daylight

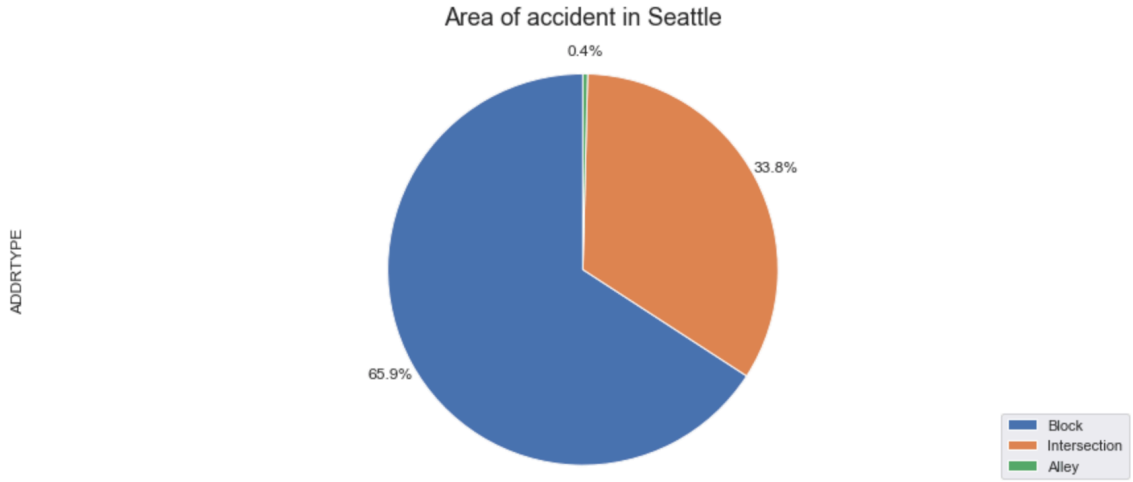
Firstly, we extract first 1000 car accidents from the dataset and plot the locations in the map of Seattle to have a basic idea about the distribution of car accidents. From the following map, we can observe that most of accidents happen near Belltown and Chinatown.

To simplify the model and select the main features, we choose the following variables and plot the frequency of car accidents related to each variable into the bar chart. From the chart, we can observe that the frequency of inattention and speeding are relative low and weather, road condition, light condition, are more important features affecting the car accidents.

I also draw a pie chart to present the percentage of avenue types when the car accident happened. We can see most of accidents happened in block and the second common place is the intersections.




Frequency for each variable
count
200000
175000
150000
125000
100000
75000
50000
25000
0
Frequency
INATTENTIONIND
UNDERINFL
WEATHER
ROADCOND
LIGHTCOND
SPEEDING
SEVERITYCODE
Variables


Area of accident in Seattle
0.4%
33.8%
65.9%
ADDRTYPE
Block
Intersection
Alley

Methodology

We use three models to predict the car accident and eventually we will compare the accuracy of these three models to choose the best one for the question.

Decision Tree makes decision with tree-like model. It splits data into two or more leaves based on the filter conditions. As the tree goes deeper and wider, the outcome is being more specific. To avoid overfitting and large amount of calculating time, we set the maximum depth of the tree is 6.

Random Forest Classifier is also a tree-based algorithm. It is a set of decision trees which are randomly selected subset of training set. So RFT has more extend degree of randomness and we do not need to set the specific conditions to add the leaves.

Logistic Regression is a classifier that estimates discrete values based on a given set of an independent variables. It predicts the probability of occurrence of an event by fitting data to a sigmond function. Largely, it is used for classification questions.

Result

In the result part, we will compare the confusion matrix and the accuracy of three models to choose the best one. After comparing the following result, we find three has similar accuracy in f1-score. Therefore, we decided to add more data for each model and compare again in the future.

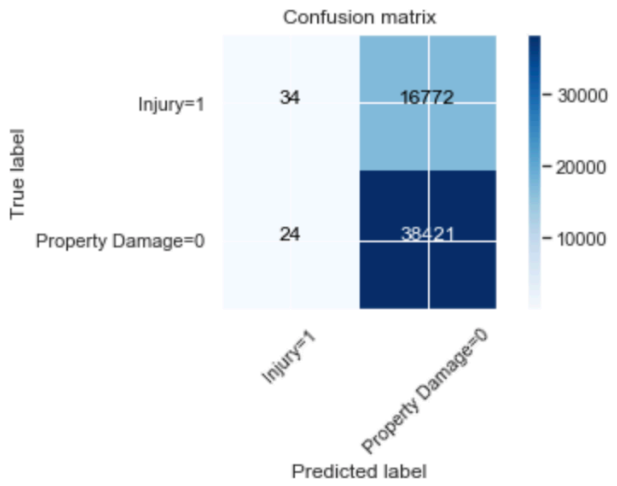
Result of Decision Tree Model

```
Confusion Matrix - Decision Tree
Predicted      0   1    All
True
0          38421  24  38445
1          16772  34  16806
All        55193  58  55251
              precision    recall  f1-score   support

           0       1.00      0.70      0.82     55193
           1       0.00      0.59      0.00        58

    accuracy                           0.70     55251
   macro avg       0.50      0.64      0.41     55251
weighted avg       1.00      0.70      0.82     55251
```

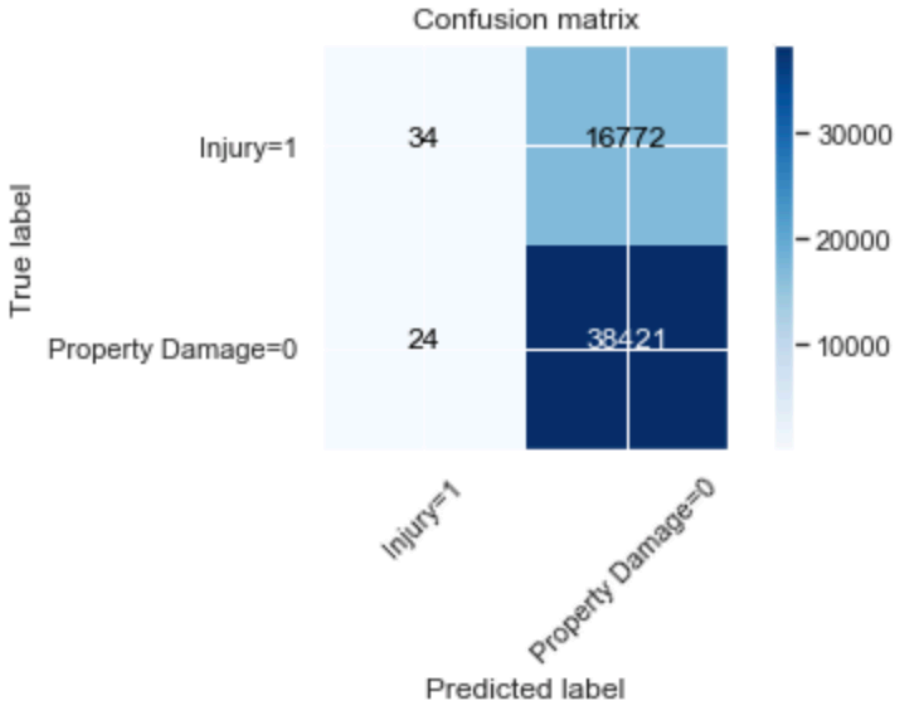
```
Confusion matrix, without normalization
[[   34 16772]
 [   24 38421]]
```



Result of Random Forest Model

	precision	recall	f1-score	support
0	0.70	1.00	0.82	38445
1	0.54	0.00	0.01	16806
accuracy			0.70	55251
macro avg	0.62	0.50	0.41	55251
weighted avg	0.65	0.70	0.57	55251

```
Confusion matrix, without normalization
[[   34 16772]
 [   24 38421]]
```



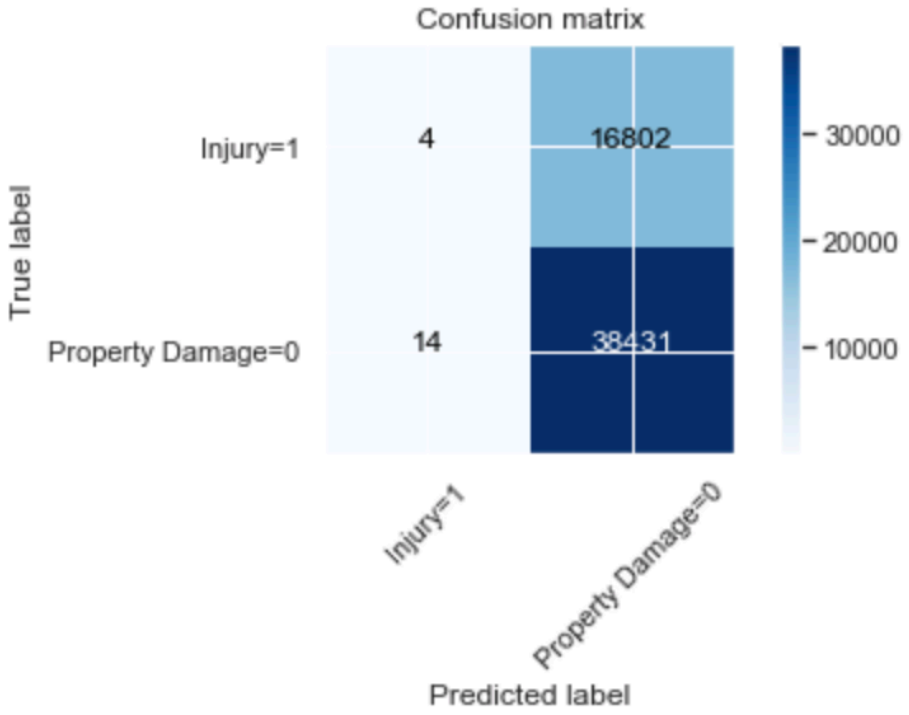
Result of Logistic Regression Model

```
0.6108557143303655
Accuracy 0.6956435177643844
              precision    recall  f1-score   support

           0       0.70      1.00      0.82     38445
           1       0.22      0.00      0.00     16806

    accuracy                           0.70     55251
   macro avg       0.46      0.50      0.41     55251
weighted avg       0.55      0.70      0.57     55251
```

```
Confusion matrix, without normalization
[[    4 16802]
 [   14 38431]]
```



Discussion and Conclusion

The accuracies of all models was around 60%-69% which means we can roughly predict the severity of an accident.

Initially, the classifiers had a prediction accuracy of 60%-69%, however, we find there are imbalance of data.

We can conclude that this model can roughly predict the severity of car accidents in Seattle. And the performance of models may be improved after adjusting the imbalance.

The trained model can be deployed onto governance and monitoring web and mobile applications to predict the accident severity for a given set of parameters.