Predicting and reduce the car accident in Seattle

Xuhuantao Bei

09/08/2020

1. Introduction

1.1 Background and Problem

The Seattle government just got enough fund to make investment to improve the traffic system in order to decrease the occurance of car accident and reduce the damage rate caused by those accidents. However, there are many factors can influence the car accidents and some of them are changeable, some of them are not. In order to make the maximum benefit from the investment, the Seattle government provides the data of car accident record in order to make analysis based on the previous data in real life.

1.2 Interest

Obviously, the local government would be interested in this project because if the accident rate is reduced, government can save lots of money to make improve the city from other aspects like healthcare,etc. On the other hand, the insurance company would be interested as well, if the accident rate is reduced, insurance company will take more profit from its product. The third party which interested in this project is the local residents because they are the taxpayers who provides the money to government and this project is the security of their life safety.

2. Data acquisition and cleaning

2.1 Data sources

The data set being used is provided by IBM data science on Coursera. It provides 194673 record of car accidents happened in Seattle in past twenty years. Information include accident location, severity level, number of cars and person involved in the accident, weather, road condition, light condition,etc. are provided in the data frame.

2.2 Data cleaning and Feature selection

Because the data set includes a lot of columns that we do not need, the first step I do is created a new data frame that only contains information that we need. After that, I realize that the target variable, severity code, is not balanced in this data set because only level 1 and level 2 of severity code are shown in this data frame; However, the number of accidents classified as level 1 are much larger than the number of accidents which classified as level 2. Therefore, I re-balanced the data frame by randomly reducing the number of level 1 accidents that contains in the data set so that I can guarantee the accuracy of the analysis. Finally, I transformed the values of

WEATHER, ROADCOND, LIGHTCOND from categorical value into numerical value so that we can apply the machine learning methods onto those columns.

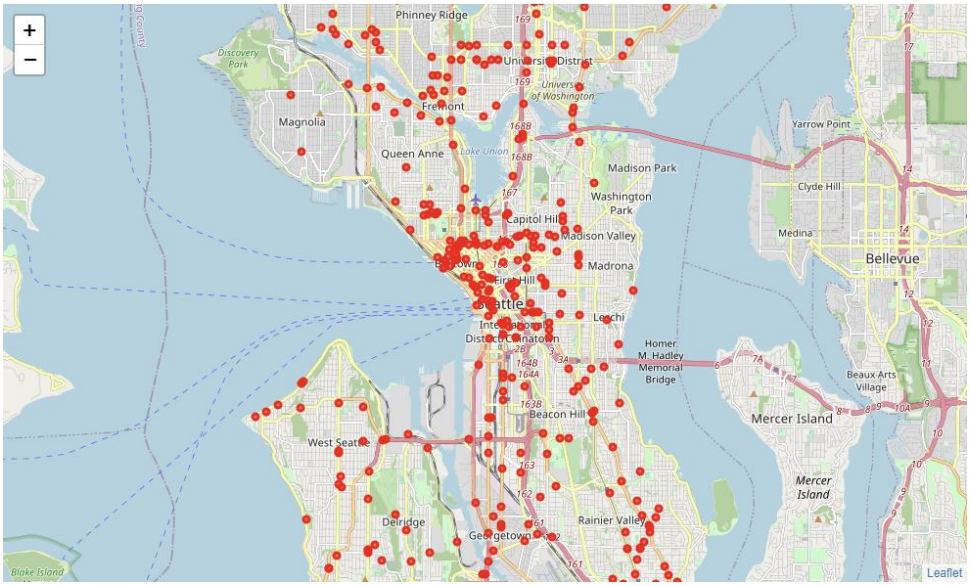
Table 1. Data Transformation

Column name	Previous value	Value after transformation
WEATHER	Clear, Raining, Overcast, Unknown,Snowing, Other, Fog/Smog/Smoke, Sleet/Hail/Freezing Rain, Blowing Sand/Dirt, Severe Crosswind, Partly Cloudy	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
ROADCOND	Dry, Wet, Unknown, Ice, Snow/Slush, Other, Standing Water, Sand/Mud/Dirt, Oil	1, 2, 3, 4, 5, 6, 7, 8, 9
LIGHTCOND	Daylight, Dark-Street Lights On, Unknown, Dusk, Dawn, Dark-No Street Lights, Dark-Street Lights Off, Other, Dark- Unknown Lighting	1, 2, 3, 4, 5, 6, 7, 8, 9

3. Data analysis

3.1 Mapping

In order to find the areas that require the improvement and traffic system, I dropped the the rows that do not have longitude and latitude and apply the remaining rows to make a graph. The the center of the map is automatically set on the center of Seattle and each red dots represents one accident that recorded in the data set.



As a result, we find that most of accidents located in areas of International District and Belltown.

3.2 Modelling

The weather condition, road condition, light condition are chosen as independent variable and the severity code is chosen as target variable. I applied three machine learning methods, KNN, Decision Tree, Logistic Regression, on the data set after dropping rows that contain Nan value. The most accurate model is logistic regression, which has 0.7 accuracy. The remain models have similar values of accuracy which is 0.698 for both KNN and Decision Tree. The results is great and it implies that our independent variables are important factor to predict the car accident in Seattle.

3.3 Evaluation

In order to make sure that our models are meaningful and accurate. I use Jaccard Score, F1 Score and Log loss methods to evaluate models. The results shown that Jaccard Score on those three models are very close which is 0.697 for KNN and 0.698 for both Decision Tree and Logistic Regression. The F1 Score for KNN is 0.575 and the F1 Score for Decision Tree and Logistic Regression is 0.574. The Log Loss for Logistic Regression is 0.604.

Table 2 Evaluation

	Model Name	Jaccard Score	F1 Score	Log Loss
0	KNN	0.697	0.575	NaN
1	Decision Tree	0.698	0.574	NaN
2	Logistic Regression	0.698	0.574	0.604

3.4 Exploration

Considering that the weather condition and road condition are hard to change, I want to test if the light condition is important factor to influence the damage rate of car accidents by assuming that the accidents involving more than five cars are high damage accidents. We see that most of car accidents happen during the daylight and most of the high damage accidents happen during the night although most of accidents happened when the street lights were turned on. The result implies that more lights are needed.

4. Conclusion

Based on the results, we can conclude that Weather Condition, Road Condition, Light Condition are three important factors to influence the everity level. Therefore, government can make investment to build a mobile application or plug-in component of navigation application to make warnings to drivers. However, Severity Code except 1 and 2 are not shown and listed in the data frame, the models still have large space to make improvement by inserting more various records of car accidents. On the other hand, government can make investment on the light system especially around the area of Belltown and International District since the weather condition and road condition are more difficult to make changes and it is possible that the light is not strong enough during the night based on the mapping and analysis.