Predicting the Severity of Car Accident

Shen Wang

October 6, 2020

1. Introduction

1.1 Background

The economic and societal impact of traffic accidents cost U.K. citizens hundreds of billions of dollars every year. Thousands of car drivers are killed or seriously injured in car crashes every year in the United Kingdom. And for many people who are not seriously injured, the injuries they experience can cause symptoms that last for months and even year, sometimes for the rest of their lives.

1.2 Problem

Several factors that can affect the possibility of people getting into a car accident and how severe it would be, such as the weather and the road conditions. Knowing these factors can provide drivers with a little insight into the severity of the accident, the drivers would drive more carefully or even change their travel if they can, and the government might be able to implement well-informed actions and better allocate financial and human resources to reduce the impact of car accidents.

1.3 Objective

The first objective of this project is to recognize key factors affecting the accident severity. The second one is to develop a model to predict accident severity based on the sophisticated traffic accident dataset. To be specific, for a given accident that just happened or a potential one, this model is supposed to be able to predict the likelihood of this accident being a severe one.

2. Data acquisition and processing

2.1 Data Sources

The dataset “Road Safety Data – Accidents 2018” is downloaded from Open Data Platform UK. The dataset is published by the Department of Transportation and is being shared under Open Government Licence. The dataset captured road accidents in the UK in 2018 and has 32 features/columns and 122635 rows. All the data variables are coded rather than containing textual strings. The dataset provider also attached lookup tables that can guide people who are going to use the dataset to understand the data.

2.2 Data cleaning and feature selection

Some features may be not useful to train the model., such as “Latitude” and “Local Authority (District)”. First of all, I simply selected some features that may be related to the analysis based on my preference, including “Accident Severity”, “Number of Vehicles”, “Number of Casualties”, “Day of Week”, “Speed limit”, “Junction

Detail", "Pedestrian Crossing-Human Control", "Pedestrian Crossing-Physical Facilities", "Light Conditions", "Weather Conditions", "Road Surface Conditions", "Carriageway Hazards", "Urban or Rural Area", "Did Police Officer Attend Scene of Accident". Since all the data variables are coded rather than containing textual strings, the dataset is relatively clean and formatted.

Data cleaning was performed to detect and handle corrupt or missing records. Because the number of observations is over 100k, which is large enough for generating a model, the observation with missing values will be directly deleted from the dataset. Based on the variable lookup table the data provider provided, all the missing values are coded as -1. Then all the observations whose values are -1 will be deleted. After deleting the missing data, 107037 observations and 14 features are left in the data.

2.3 Data Processing

After data cleaning and feature selection, the next step is to check the target variables, "Accident Severity". Based on the variable lookup table provided by the data provider, the target variable has 3 possible values 1, 2, and 3, which represents "fatal", "serious", and "slight" respectively. There are 84567, 20970, and 1550 observations marked as "slight", "serious", and "fatal", which is unbalanced. The unbalanced target variable may cause some problems in training models. Because the dataset has over 100k observations, I simply deleted 40000 observations whose target variable is "slight". And then I combined "serious" and "fatal" observations. Finally, there are 44567 "slight" and 22470 "serious" observations.

After cleaning the target variables, all the categorical variables are converted to binary variables, including, "Day of Week", "Junction Detail", "Pedestrian Crossing-Human Control", "Pedestrian Crossing-Physical Facilities", "Light Conditions", "Weather Conditions", "Road Surface Conditions", "Carriageway Hazards", "Urban or Rural Area", "Did Police Officer Attend Scene of Accident". The next step was to normalize the variables that were not categorical: "Number of Vehicles", "Number of Casualties", "Speed limit". I simply used the "StandardScaler" to normalize the variables.

3. Exploratory data analysis

3.1 Calculation of target variable

The target variable in the dataset is accident severity. 0 and 1 represent "slight" and "serious" accidents respectively. There are 44567 "slight" and 22470 "serious" observations after cleaning the dataset.

3.2 Relationship between accident severity and weather conditions

After converted the weather conditions to binary variables, "1" represents the weather condition is fine, and "0" represents the weather condition is not fine. I calculated the percentage of different accident severity when the weather conditions are different. Based on Figure 1, the table shows that there is no big difference in the severity of accident when the weather conditions are different. When the weather is fine, the number of different accident severity is all about 85% of the total number of accident severity. When the weather is not fine, the number of different accident severity is all about 15% of the total number of accident severity.

Accident_Severity	Weather_Conditions	
serious	fine	0.859769
	not_fine	0.140231
slight	fine	0.845962
	not_fine	0.154038

Figure 1. Relationship between accident severity and weather conditions

3.3 Relationship between accident severity and light conditions

After converted the light conditions to binary variables, "1" represents the light condition is "daylight", and "0" represents the light condition is "not daylight". Based on Figure 2, the "serious" accidents are more likely to happen when the light condition is not good.

Accident_Severity	Light_Conditions	
serious	daylight	0.699421
	not_daylight	0.300579
slight	daylight	0.763166
	not_daylight	0.236834

Figure 2. Relationship between accident severity and light conditions

3.4 Relationship between accident severity and junction detail

After converted the junction details to binary variables, "1" represents that the accident happens at the junction, and "0" represents that the accident doesn't happen at a junction or within 20 meters to a junction. Based on Figure 3, there is no big difference in the severity of accident when the accident happens at a junction or not.

Accident_Severity	Junction_Detail	
serious	at_junctino	0.505073
	not_at_junction	0.494927
slight	at_junctino	0.516705
	not_at_junction	0.483295

Figure 3. Relationship between accident severity and junction detail

4. Predictive Modeling

4.1 Model Development

5 classification models are used to predict the severity of accident, including logistic regression, K-nearest neighbors, decision tree, random forest, and support vector machine. First of all, I split the data into training and testing dataset. The training dataset owns 70% of the total data, the testing dataset contains the rest 30% data. I used n-neighbors = 7 as the parameter in KNN model, and n-estimators=100 as the parameter for random forest model. And then I calculate their accuracy, F1 – Score, and Jaccard – Score. Finally, I print out

the result in a table.

4.2 Model Performance

Based on Figure 4, the SVM model performed the best (67.80% accuracy, 62.13 F1- Score). The accuracy of the rest models is all above 64.00. And the F1 – Score of all models are over 57.50. Noticed that, the model accuracy is not high. A possible reason is the unbalance of the dataset.

	Model	Accuracy	F1 - Score	Jaccard - Score
0	Support Vector Machines	67.80	62.13	16.27
2	Logistic Regression	67.01	57.50	7.23
4	Decision Tree	66.48	61.71	17.06
3	Random Forest	66.46	62.11	18.12
1	KNN	64.45	61.28	19.13

Figure 4. Model Performance

5. Discussion

Noticed that the dataset is unbalanced, the accuracy of the models is not good enough and has room to improve. Another reason why the accuracy is not high is that the variables may not be good to predict the accident. More variables and data related to car accidents would be helpful to improve model performance, such as the drivers' background information and vehicle conditions.

6. Conclusions

In this project, I analyzed the relationship between accident severity and car accident data. I built 5 classification models to predict whether a car accident would be slight or serious. These models can be very helpful for the government to implement well-informed actions and better allocate financial and human resources to reduce the economic and societal impact of car accidents.