Capstone Project - Car Accident Severity

Weihan Tang

September 6, 2020

Contents

1 Introduction

When it comes to car accidents, there can be a variety of external factors that contribute and lead to the causes. As one can imagine, road intersections might be the hot spots where accidents happen. Also, accidents may happen more frequently during a raining day due to lack of vision. A combination of such external factors may lead to road accidents with different severity. It is of high interest to predict severity of road accidents based on the readily observable and attainable data of external conditions, so that first responders are aware of the type of accidents they are dealing with.

In this study, we will examine across a large amount of car accidents with the power of data science, and correlate the severity of the accidents with their attributes, such as locations, road conditions, weather, and etc. Such correlation can facilitate in training a regression model that predicts severity of accidents.

2 Data Understanding and Preparation

2.1 Data Sources

In this study, we will use the dataset provided by the course of Applied Data Science Capstone project. This dataset contains nearly 200000 cases of car accidents, which is sufficient for our study.

2.2 Data Selection and Cleaning

2.2.1 Duplicate Columns

First of all, the target variable that we try to predict in this study is the severity code. This variable has two values, 1 and 2, whereas a value of 1 suggests property damage only, and a value of 2 indicates injury is involved. A glance at this dataset shows that there are two duplicate columns that contain data of this variable. Thus, we start by dropping one of the columns.

2.2.2 Feature Selection

This vast dataset contains information that is clearly not related to the cause of a road accident, such as OBJECTID, INCKEY, and etc. Here we only keep meaningful data that potentially have an impact on the cause of a road accidents, which include the following:

- Location of the accident: 'X', 'Y'
- Is the location a intersection or a block: 'ADDRTYPE'
- Weather condition: 'WEATHER'
- Under influence of achohol: 'UNDERINFL'
- Road condition: 'ROADCOND'
- Light condition: 'LIGHTCOND'item Speeding or not: 'SPEEDING'

2.2.3 Missing Values

The dataset contains many values of NaN. These values need to be dealt with in different ways.

- The location data 'X'and 'Y'have a few rows of missing data. In this case, the entire rows are dropped. The same approach is applied to the other features where the number of NaN values contributes to only a small portion of the whole columns.
- An exception is the data contained in 'SPEEDING'. Due to the data format contained in the raw spreadsheet, only a speeding case has an entry 'Y', the rest are missing values. In this case we replace the NaN values with 0.

2.2.4 Data Formatting and Filtering

Many features in the dataset has string values. To prepare this dataset for the downstream analysis, we need to replace the categorical values with quantitative values. Here we use one-hot binning to convert categorical data into a number of indicator columns. For example, the 'ROADCOND'column that contains the road condition data is converted into 9 separate columns - 'Dry', 'Ice', and etc, each containing binary values indicating whether an accident happened under such road conditions.

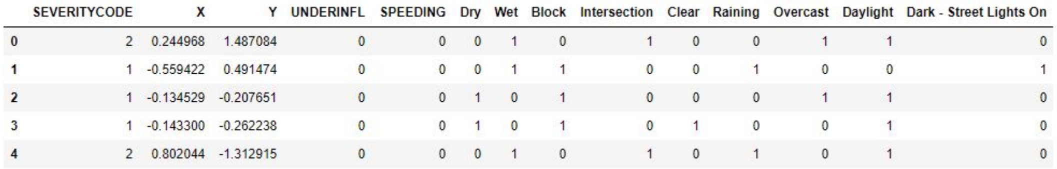
Further investigation shows that 'Dry'and 'Wet'conditions contribute to 91% of the total number of cases. To simplify the dataset, we only keep 'Dry'and 'Wet'columns, and drop the rest. The two columns are concatenated with the dataset, and the original column 'ROADCOND'is dropped. This approach is also applied to other columns 'ADDRTYPE', 'WEATHER', and 'LIGHTCOND'.

Another issue the data set has is that columns such as 'UNDERINFL'contains mixed values of Y, 0, and N. Both N and 0 indicate the driver is not influenced by achohol. Here we convert values of N to integer 0, and Y to integer 1. We apply the same approach to column 'SPEEDING'. Also, upon checking the data types of each column, we notice that column 'UNDERINFL'have a type of 'object'. We need to cast these two variables into integers for further analysis.

2.2.5 Data Normalization

The location parameters 'X'and 'Y'are clustered around -122 and 47, with small deviations. In order to facilitate further analysis, data normalization is necessary. Here we use z-score normalization.

After the above process, the data is now cleaned, formated, and filtered for modeling and downstream analysis, as shown below.

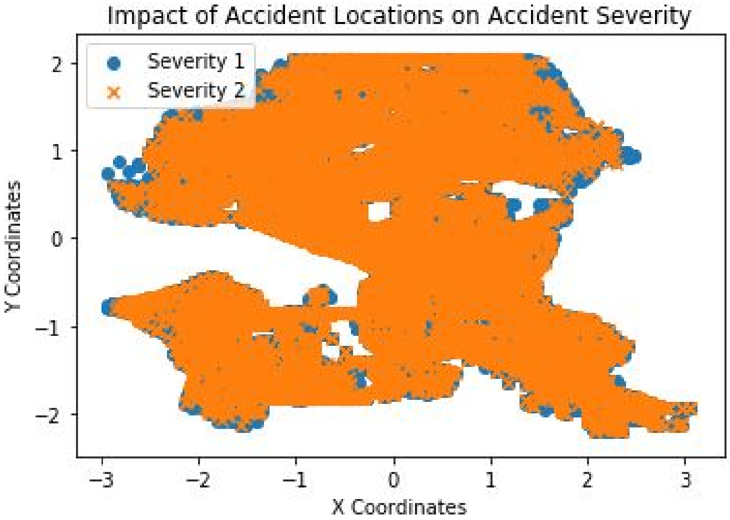


	SEVERITYCODE	X	Y	UNDERINFL	SPEEDING	Dry	Wet	Block	Intersection	Clear	Raining	Overcast	Daylight	Dark - Street Lights On
0	2	0.244968	1.487084	0	0	0	1	0	1	0	0	1	1	0
1	1	-0.559422	0.491474	0	0	0	1	1	0	0	1	0	0	1
2	1	-0.134529	-0.207651	0	0	1	0	1	0	0	0	1	1	0
3	1	-0.143300	-0.262238	0	0	1	0	1	0	1	0	0	1	0
4	2	0.802044	-1.312915	0	0	0	1	0	1	0	1	0	1	0

3 Exploratory Data Analysis

3.1 Relationship between Accident Locations and Accident Severity

Certain locations can have higher rate of accidents, such as two-lane roads, parking lots and etc. In this session, we examine the relation between coordinates of accident location and accident severity. The scatter plot below shows that accidents of severity 1 and 2 nearly covers the same area. In this case, accident locations hardly play a role.



3.2 Relationship between Influence of Alcohol and Accident Severity

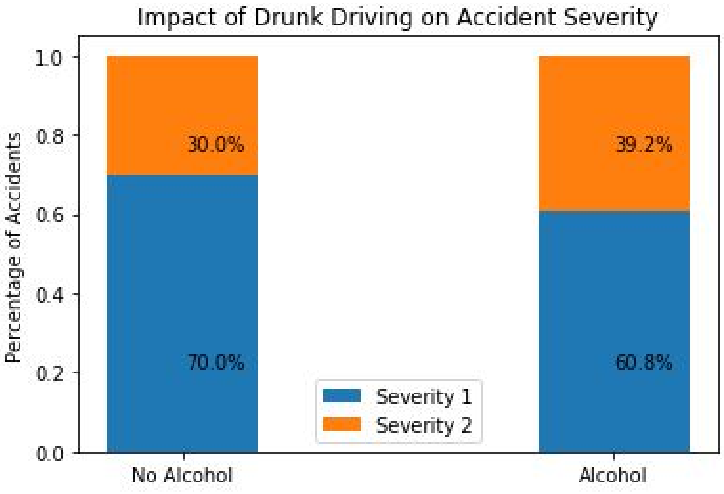
It is commonly know that druck driving contribute significantly to road accidents. In this dataset, the majority of accidents does not involve drunk driving. Comparing number of accidents directly can be misleading. In this comparison, numbers of accidents of severity 1 and 2 are converted into percentage values for each of the two categories: drunk driving and not drunk driving, as shown in the plot below. It is observed that under the influence of alcohol, percentage of accidents of severity 2 increases by 9.2%, indicating our hypothesis is correct.

3.3 Relationship between Speeding and Accident Severity

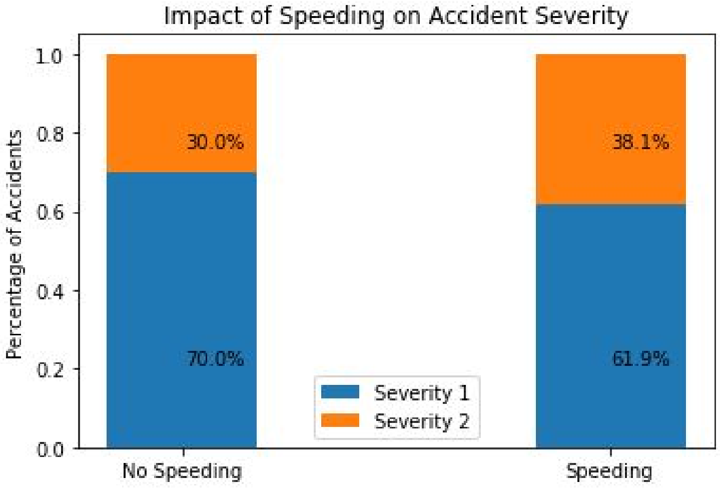
Similar to the influence of alcohol, the hypothesis in this case is that speeding contribute to the increase of accident severity. We apply the same approach used in the previous session, and the comparison is shown below. It is observed that speeding results in an increase percentage of accidents of severity 2 by 8.1%, indicating our hypotheses is correct.

3.4 Relationship between Road Conditions and Accident Severity

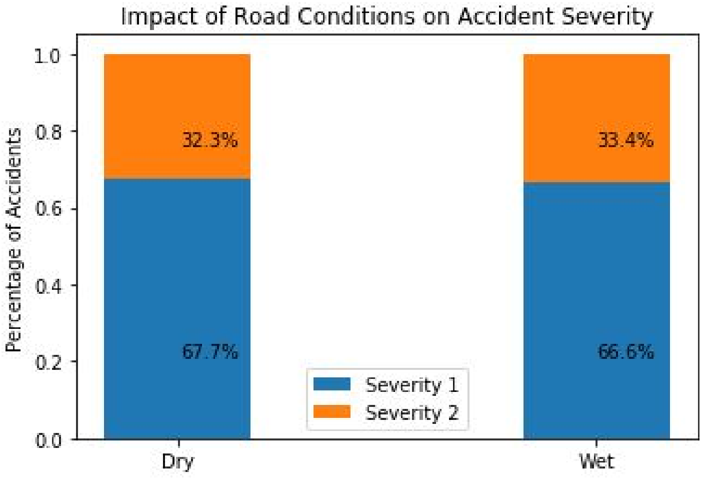
The hypothesis we use in this case is that wet road condition causes slippery road, which leads to higher occurrence of accidents. We apply the same visualization approach used in the previous session, and the comparison is shown below. The plot indicates that the percentage of accident severity does not differ much under dry or wet road conditions. Thus the hypothesis is incorrect.



Impact of Drunk Driving on Accident Severity



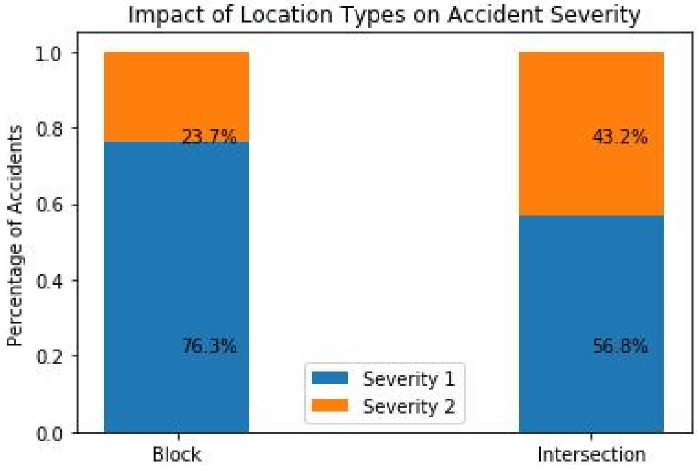
Impact of Speeding on Accident Severity



Impact of Road Conditions on Accident Severity

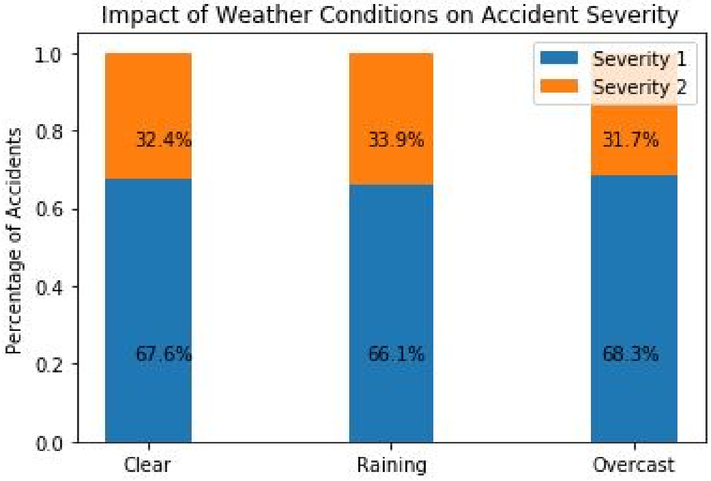
3.5 Relationship between Types of Accident Locations and Accident Severity

Preliminary investigation reveals that most accidents occur at a block or at an intersection. The hypothesis is that an intersection has higher accident severity as cars are coming from different directions, and passengers can be hit and injured from the side. The plot below shows our hypothesis is correct. Accidents of severity 2 occurred at an intersection is higher than those happened at a block, by 19.5%.



3.6 Relationship between Weather Conditions and Accident Severity

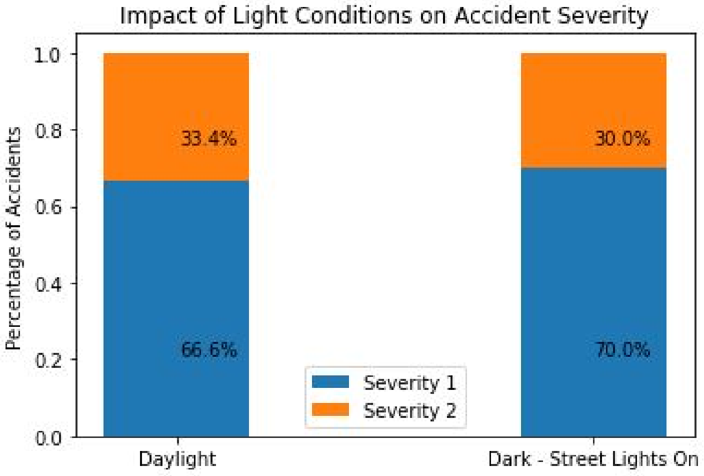
The hypothesis is that a raining day or a clear day can cause a driver to have lower visions due to rain and sunshine, respectively. However, the plot below shows our hypothesis is incorrect. Similar percentage of accident severity is observed for all three weather conditions.



3.7 Relationship between Light Conditions and Accident Severity

The common knowledge and hypothesis is that driving during a night lead to higher occurence and severity of accidents. On the contrary, the plot below shows our hypothesis is incorrect.

Similar percentage of accident severity is observed for both light conditions.



4 Modeling

In this section, we will explore three different machine learning techniques that are suitable for predict categorical target values (i.e. classification problem), namely K-Nearest Neighbours (KNN), decision tree, and logistic regression. A cluster of 50 neighbors is used for the KNN method. A decision tree is built with the maximum entropy gain criterion, and a maximum depth of 4. As for the logistic regression model, a regularization parameter of 6 and the 'liblinear' solver is used.

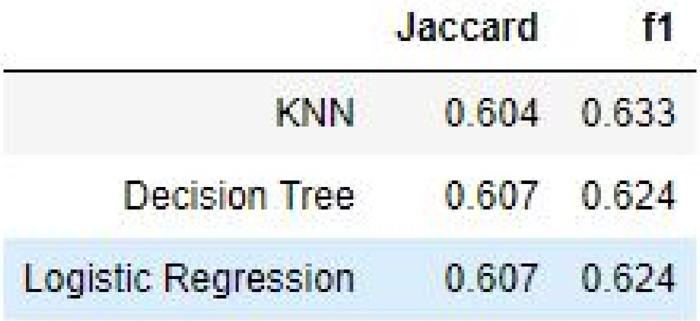
As observed in the previous section, location coordinates, light conditions, weather conditions, and road conditions have insiginificant effect on accident severity, and are thus excluded for the modeling process. Data retained for modeling are shown below.

Note that the dataset has been balanced prior to be fed into our models. Also, The dataset is split into training set and test set. Here we use 30% of our data for testing and the rest 70% for training. Predicted results are evaluated using metrics of Jaccard similarity score and f1 score.

5 Results and Discussion

A comparison of the prediction by KNN, decision tree, and logistic regression, evaluated by Jaccard similarity score and f1 score, is shown below.

It seems that the performance of all three models are on a similar level. Even with the model parameters tuned, the highest score is still around 0.6. A closer look at the confusion matrix shows a high number of false positive and true negative cases (7452 and 5754). These numbers are on the same level of true positive and false negative cases (10992 and 9410). This suggests that we need to include more impactful features to properly categorize these cases.



	Jaccard	f1
KNN	0.604	0.633
Decision Tree	0.607	0.624
Logistic Regression	0.607	0.624

6 Conclusion

In this study, we examined the dataset that contains potential causes that contribute to severity of road accidents. The dataset contains duplicate columns, missing values, and irrelevant information, and has been cleaned, formatted, and filtered prior to the analysis stage.

Exploatory analysis reveals relationship between each feature and the accident severity. It is interesting to note that road, light, and weather conditions have insignificant effect on the target value. This observation defies our commom sense and shows some of our hypotheses are incorrect. These features are dropped from our analysis as they are less impactful than the rest.

The simplified dataset is then fed into three machine learning models - K-Nearest Neighbor, decision tree and logistic regression. The models are tuned and evaluated by the metrics of jaccard similarity score and f1 score. Results show that additional features might be necessary to improve the accuracy of predictions.