Predicting Car Accident Severity

Peiming Wu

August 21,2020

1.Introduction/Business Problem

1.1 Background

There are numerous car accidents happen across the world every day. All car accidents will create damage to cars involved or even worse, take lives. People want to drive safe not only in terms of reducing the chance of damaging their cars but also less life risks. In fact, there are many factors that contribute to the severity of a car accident. Therefore, it is advantageous for related departments to accurately predict the severity of car accidents under those conditions. For example, does bad road conditions involves in large number of car accidents? If it does, the prediction provides solid reason for better road constructions. From that, warning information like road signs to the drivers will lead to positive impacts.

1.2 Business Problem

Data that might contribute to a car accident including locations, weathers, road conditions, light conditions, vehicles or pedestrians involved, speeding, whether the driver is involved was under the influence of drugs or alcohol, etc. This project aims to predict the car accident severity based on these data.

2.Data acquisition and cleaning

2.1 Data Sources

The dataset that will be employed in this project is the example dataset provided by the teaching staff. It includes all collisions provided by SPD and recorded by Traffic Records. The dataset could be downloaded from https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv and the Metadata could be downloaded from https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf . This dataset is a supervised with labeled severity of car collisions with numerous attributes like locations, weathers, road conditions, light conditions, etc. This dataset, however, needs to clean since there are empty inputs in some attributes like road conditions.

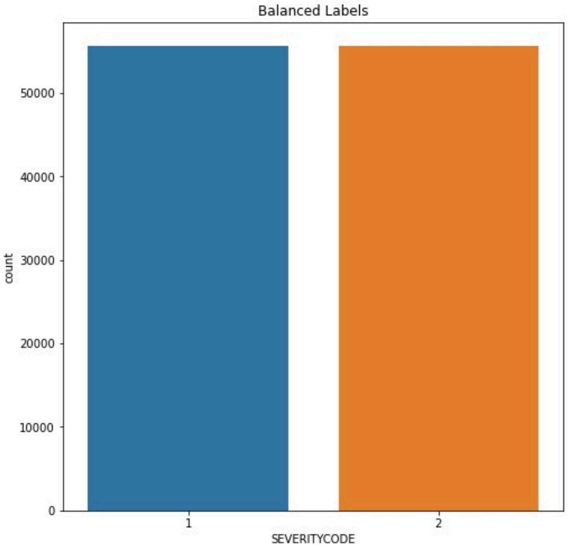
2.2 Data Cleansing

The feature sets I decided to use include 'PERSONCOUNT', which stands for the number of people involved in the collision, 'PEDCOUNT', which stands for the number of pedestrians involved in the collision, 'PEDCYLCOUNT', which stands for the number of bicycles involved in the collision, 'VEHCOUNT’, which stands for the number of vehicles involved in the collision, 'ROADCOND', which stands for the road condition during the collision, 'WEATHER', which stands for the weather during the time of the collision, 'LIGHTCOND', which stands for the light condition during the collision.

There are two major problems with the dataset. First, the labels are unbalanced. After getting the details of the dataset, I found this dataset only provides two different labels for ‘SEVERITYCODE’. They are ‘1’ for ‘prop damage’ and ‘2’ for ‘injury’. However, the number of these accidents for each has huge difference. The originally dataset has double size of label 1 accidents than that of label 2 accidents. Therefore, it’s necessary to shuffle and resample to create a balanced dataset before training. Otherwise, a biased predication will be generated.

The second problem I need to deal with is that for ‘ROADCOND’, ’WEATHER’, ’LIGHTCOND’ feature sets use categorical variables. Hence the conversion from categorical to numeric is necessary before employing machine learning methods. Beyond that, some inputs in these categorical columns are empty, which should be dropped from the dataset in order to provide better results.

Therefore, I first dropped all the rows with empty inputs in those categorical columns. After that, I converted the categorical variables into numeric variables. Finally, I shuffle the dataset and resample randomly to create a dataset with equal size rows for both label 1 and label 2 to ensure an unbiased predication. The resampled and balanced labels could be examined below:



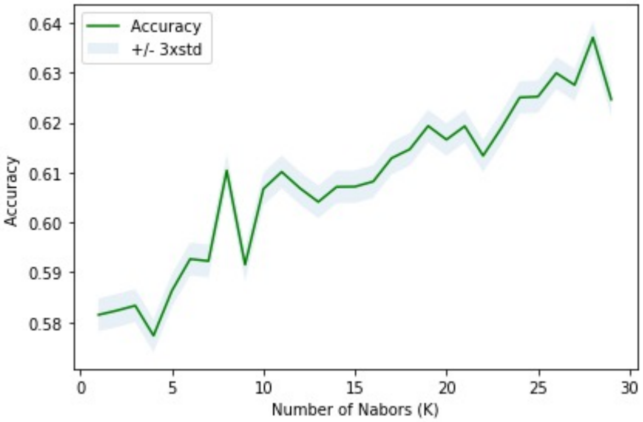
3. Methodology

I decided to use classification method since the problem is find a way to predict the severity of a car accident. The reason is that classification is about predicting labels with supervised data. In this project, the severity of all car accidents collected in the data is provided with 'SEVERITYCODE' labels, which suggests that classification is the ideal machine learning method to be employed in this project.

Among available classification methods, I choose KNN since most of the feature sets after cleansing is highly trainable using KNN model.

4. Results

By exploring different k values from 0-30 and feature sets I mentioned in previous sections, I got maximum test set



accuracy of ~64% with k=28 and f1-score ~0.63. This is an acceptable result. However, there's a large room to improve this since I only use the provided dataset and build a naïve model. Considering that the purpose of this project is to predict potential car accidents, feature sets like people involved could be hard to measured when we try to provide warnings. Instead, road conditions, weathers and light conditions are more “stable” factor when we try to predict. If one wish to optimize the model, one might drop the number of people, vehicles involved and try to use feature sets similar to geographical conditions since these are more measurable features before accidents happen.

5. Conclusion

This project allows predication of car accidents by training feature data sets that related to road conditions, weathers or light conditions during the time of collected car collisions. With decent accuracy of predication and potential future optimization, I believe that people could use this classification model to predict car accidents and hence provide warnings under certain combination of different conditions. (i.e. different features)