Applied Data Science Capstone Project

Car Accident Severity Analysis: Seattle, Washington

Miller Palaniswamy |IBM-Coursera| 10-Sep-20

Contents

1.Introduction/Business Problem

1.1 Background

For an individual & the family dependent on him, his life & health is more important for survival. 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. In fact, there are many factors that contribute to the severity of a car accident & human safety.

At the same time growth of transportation plays an indispensable role in our society and their ever-increasing dependency to enables communication, trade and exchange of goods and services across the globe, is unstoppable. But it has very unfortunate impact on the society in terms of accidents.

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. Since transportation industry is ripe for advancement, data science can bring about an evolution in this sector.

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.

1.2 Case Study – Seattle traffic data

Washington State's largest city, Seattle, is a home for large tech giants such as Microsoft and Amazon headquartered in its metropolitan area. As of 2020, it has a total metro area population of 3.4 million (https://www.macrotrends.net)The total number of personal vehicles in Seattle in the year 2016 hit a new high of nearly 444,000 vehicles. In one South Lake Union census tract, the car population has more than doubled since 2010(http://www.seattletimes.com). The increase in car ownership rates can lead to higher numbers of accidents on the road because of a simple probability. Worldwide, approximately 1.35 million people die in road crashes each year, on average 3,700 people lose their lives every day on the roads and an additional 20-50 million suffer non-fatal injuries, often resulting in long-term disabilities. We will be trying to find possible insights from this data about the reasons as to why accidents occur, which area is more prone to accidents and what are the aftereffects of accidents on traffic flow.

The world suffers due to car accidents, including the USA. National Highway Traffic Safety Administration of the USA suggests that the economical and societal harm from car accidents can cost up to $871 billion in a single year. According to 2017 WSDOT data, a car accident occurs every 4 minutes and a person dies due to a car crash every 20 hours in the state of Washington while Fatal crashes went from 508 in 2016 to 525

in 2017, resulting in the death of 555 people. The project aims to predict how severity of accidents can be reduced based on a few factors.

1.3 Beneficiaries & Stakeholders

Data analytics can provide an in-depth knowledge of methods for analyzing and implementing intelligent transportation systems. For example: connected vehicles can be a way forward in the transportation industry. Sensors placed around cities that are then connected to apps can help drivers find parking spots faster, reducing traffic and emissions. The benefits of big data and analytics helps transportation firms to precisely enhance the model capacity, demand, revenue, pricing, customer sentiments, cost. Some of the assets include implementing real time monitoring system for enhancing operational efficiency, public transit system and traffic management system to improve bus transportation and reduce traffic congestion. By developing this project, we can help society with implementing some laws needed for transportation which can prevent the accidents and also helps them to know at where they have to be safer on roads which are more prone to accidents.

2.Data description, acquisition & Preprocessing

2.1 Data Sources

The dataset of all the collisions sahred by SPD – Traffic department can 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

There are 194,673 observations and 38 variables in this data set. Since we would like to identify the factors that cause the accident and the level of severity, we will use SEVERITYCODE as our dependent variable Y, and try different combinations of independent variables X to get the result. Since the observations are quite large, we may need to filter out the missing value and delete the unrelated columns first. Then we can select the factor which may have more impact on the accidents, such as address type, weather, road condition, and light condition.

The target Data to be predicted under (SEVERITYCODE 1-prop damage 2-injury) label.

Other important variables include:

• ADDRTYPE: Collision address type: Alley, Block, Intersection

• LOCATION: Description of the general location of the collision

• PERSONCOUNT: The total number of people involved in the collision helps identify severity level

• PEDCOUNT: The number of pedestrians involved in the collision helps identify severity level

• PEDCYLCOUNT: The number of bicycles involved in the collision helps identify severity level

• VEHCOUNT: The number of vehicles involved in the collision identify severity level

• JUNCTIONTYPE: Category of junction at which collision took place helps identify where most collisions occur

• WEATHER: A description of the weather conditions during the time of the collision

• ROADCOND: The condition of the road during the collision

• LIGHTCOND: The light conditions during the collision

• SPEEDING: Whether or not speeding was a factor in the collision (Y/N)

• SEGLANEKEY: A key for the lane segment in which the collision occurred

• CROSSWALKKEY: A key for the crosswalk at which the collision occurred

• HITPARKEDCAR: Whether or not the collision involved hitting a parked car

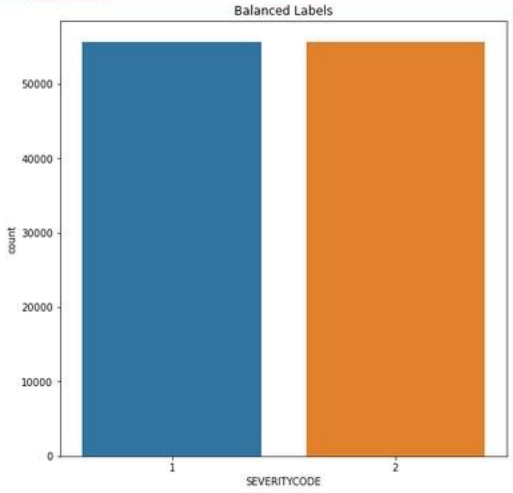
There are two major problems with the dataset. First, the labels are unbalanced. After getting the details of the dataset, I found this dataset only provides two different labels for 'SEVERITYCODE'. They are '1' for 'prop damage' and '2' for 'injury'. However, the number of these accidents for each has 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.

```
### Balance the dataset ###

# Shuffle the Dataset.
shuffled_df = df_clean.sample(frac=1,random_state=4)

# Put all the severity_code = 2 in a separate dataset.
severity_2_df = shuffled_df.loc[shuffled_df['SEVERITYCODE'] == 2]

#Randomly select len(severity_code == 2) observations from the 'severity_code == 1' (majority class)
resample_severity_1_df = shuffled_df.loc[shuffled_df['SEVERITYCODE'] == 1].sample(n=len(severity_2_df),random_state=42)

# Concatenate both dataframes again
norm_df = pd.concat([severity_2_df, resample_severity_1_df])

#plot the dataset after the undersampling
plt.figure(figsize=(8, 8))
sns.countplot('SEVERITYCODE', data=norm_df)
plt.title('Balanced Labels')
plt.show()
```

```
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages/seaborn/_decorators.py:43: FutureWarning: Pass the follow
an error or misinterpretation.
  FutureWarning
```




```
## Double check to make sure the data now has balanced labels
norm_df['SEVERITYCODE'].value_counts()
```

```
2    55683
1    55683
Name: SEVERITYCODE, dtype: int64
```

Even this sample values can still be pruned and reduced by another 5000+, if we consider removing columns with more than 20% values missing such as (INATTENTIONIND, PEDROWNOTGRNT, SPEEDING) & rows for columns with less than 20% values missing (COLLISIONTYPE,JUNCTIONTYPE,UNDERINFL,WEATHER,ROADCOND,LIGHTCOND etc.