Applied Data Science Capstone by IBM/Coursera

Week 3 - Car Accident Severity Report

Mohammad H Saeedian / Sept 30, 2020

1. Introduction: Business Problem

Road traffic accidents are responsible for millions of deaths and injuries every year in the world. Road traffic injuries cause significant economic losses from medical treatment cost to productivity loss and it costs most countries 3% of their gross domestic product . The World Health Organization describes the road traffic system is the most complex and the most dangerous system with which people have to deal every day.

To reduce traffic accidents is an important public safety challenge and big data analytics has emerged with powerful techniques to provide insights on factors leading to the increased risk of accidents. Therefore, it can be used for public to be more aware of potential accident risks when planning trips like locations which are susceptible for accident occurrence and to avoid dangerous driving behavior like drunk driving or speeding. Moreover, it can be used to develop public traffic prevention operations and policies to reduce overall accidents.

2. Data Engineering

2.1 Feature Selection

Example dataset was used and the goal was to predict accident severity. It was noticed that only 2 codes were included in dataset but no other severity codes . After reviewing all 32 attribute columns in the dataset, below 13 attributes were initially chosen for data analysis. Both features that indicating accident impact scale (people and vehicles involved) and factors that could potentially lead to accidents (location, weather, road, light conditions, driver behaviors) are included. Other columns with tracking information and columns with duplicated information were not selected. Column “status” was used to eliminate data rows with “unmatched” status.

- Accident Location
- Address Type
- Collision Type
- Person Count involved in accident
- Pedestrians involved in accident
- Bicycles involved in accident
- Vehicle Count involved in accident
- Weather
- Road Condition
- Light Condition
- Speeding
- Whether inattention
- Whether driver(s) under influence
- Whether hit a parked car

2.2. Dealing with Data Imbalance

From the dataset, there is data imbalance issue as there are more severe code 1 cases than 2. In reality, a severe road accident is a rarer event than minor accident (no injury). Since machine learning algorithms usually have difficulty learning from imbalanced datasets, this dataset was rebalanced using under-sampling method with smaller dataset for analysis.

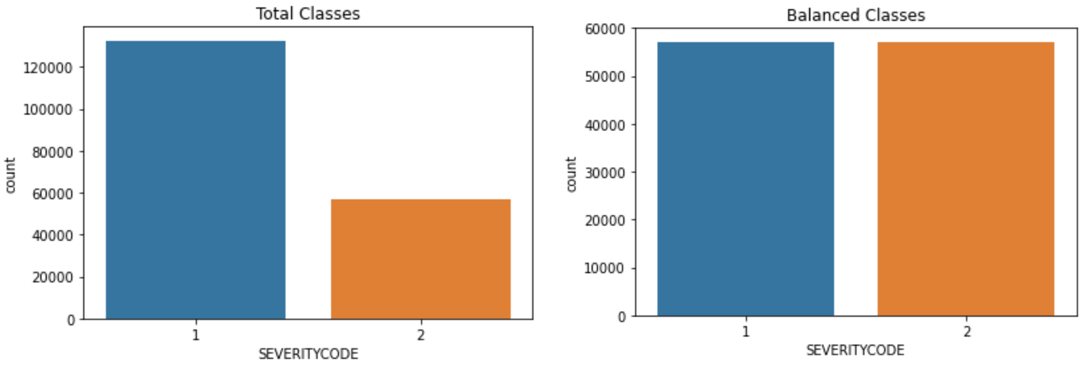


Figure 1 Dataset count before and after rebalancing

2.3. Data Understanding and Cleaning

Looking at count plot by collision types, it gives good information of what type collision is more likely to happen (e.g. more accidents from left turn than right turn and more likely to have injury). Other observations include if pedestrian(s) or cycle(s) involved, the chance of injury is significantly higher, and if hitting a parked car, the chance of injury is significantly lower. Based on these observations, whether hit a parked car, pedestrian and cycle involvement are considered useful features and these 3 attributes were selected for further model development.

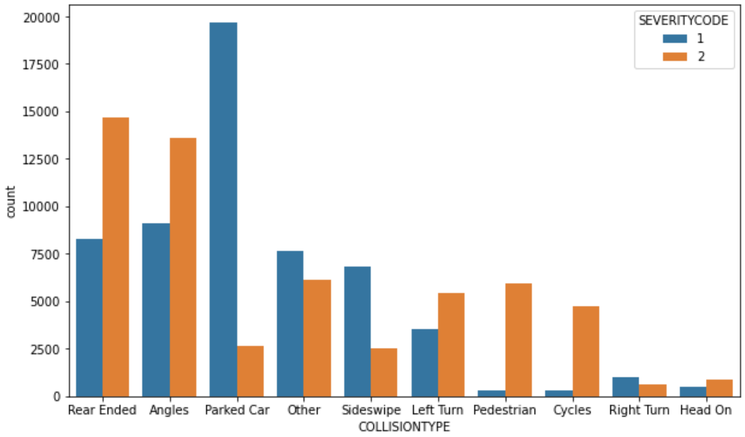
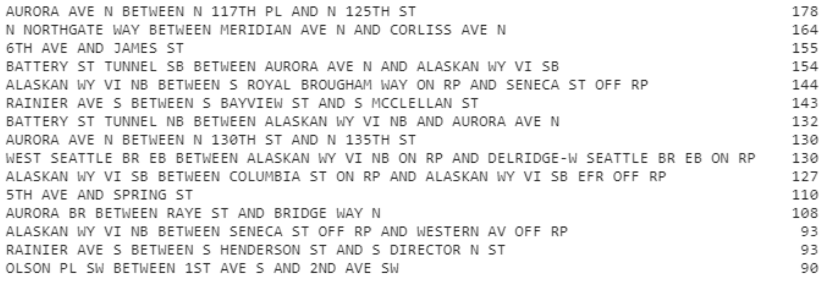


Figure 2 Count plot by collision type

From address count, table 1 lists top locations that are more prone to have accidents. Further data processing could be done to group addresses based on their accident injury occurrence ratio and one hot encoding technique can be used to convert data to binary variables and append to data frame, however the main interest of this case study was to understand effect of various factors, this address attribute was dropped and will be included in future study.



```
AURORA AVE N BETWEEN N 117TH PL AND N 125TH ST                               178
N NORTHGATE WAY BETWEEN MERIDIAN AVE N AND CORLISS AVE N                     164
6TH AVE AND JAMES ST                                                         155
BATTERY ST TUNNEL SB BETWEEN AURORA AVE N AND ALASKAN WY VI SB               154
ALASKAN WY VI NB BETWEEN S ROYAL BROUGHAM WAY ON RP AND SENECA ST OFF RP     144
RAINIER AVE S BETWEEN S BAYVIEW ST AND S MCCLELLAN ST                        143
BATTERY ST TUNNEL NB BETWEEN ALASKAN WY VI NB AND AURORA AVE N               132
AURORA AVE N BETWEEN N 130TH ST AND N 135TH ST                               130
WEST SEATTLE BR EB BETWEEN ALASKAN WY VI NB ON RP AND DELRIDGE-W SEATTLE BR EB ON RP    130
ALASKAN WY VI SB BETWEEN COLUMBIA ST ON RP AND ALASKAN WY VI SB EFR OFF RP   127
5TH AVE AND SPRING ST                                                        110
AURORA BR BETWEEN RAYE ST AND BRIDGE WAY N                                   108
ALASKAN WY VI NB BETWEEN SENECA ST OFF RP AND WESTERN AV OFF RP               93
RAINIER AVE S BETWEEN S HENDERSON ST AND S DIRECTOR N ST                      93
OLSON PL SW BETWEEN 1ST AVE S AND 2ND AVE SW                                  90
```

Table 1 Value count by address in descending order

From address type count plot, it shows accidents happen more frequently at block while accidents happening at intersection are more likely to cause injury. This categorical feature of “alley”, “block”, “intersection” was converted to numerical values 0, 1, 2 and was selected for model.

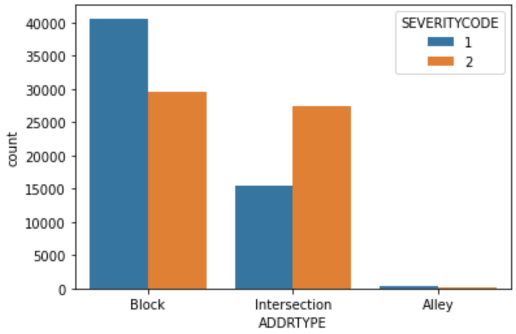


Figure 3 Count plot by address type

Weather, road condition, light condition attributes were processed with simplified categories and converted to ordinal numbers based on potential impact on increasing accident risk. For example, light condition was converted to 1-bright daylight, 2-dust/dawn, 3-dark with lights on, 4-dark with no light. “Other” or “unknow” data was converted to 0 and removed from dataset. Linear regression analysis was also run between weather and road condition attributes and the R-squared score is 0.61 indicating decent correlation between the two. Therefore, only one attribute “weather” was kept for next step.

For speeding, inattention, under-influence and hit-parked-car attributes, missing data entry was interpreted as “N” and then binary values were unified and converted to 0 or 1. For cases with driver under influence, it shows there is higher chance of 60% causing injure compared to 50% otherwise. All these attributes were kept for building model.

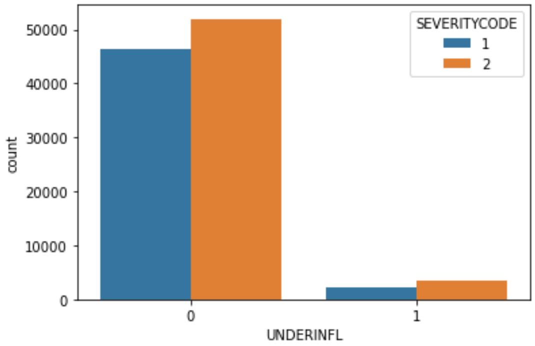


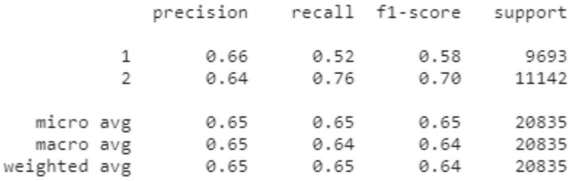
Figure 4 Count plot by whether-under-influence

At last for the whole dataset, data rows with missing data entry were dropped. After data engineering, 11 features were selected with collision type and address location dropped.

3. Model Development and Result Discussion

This accident severity prediction was defined as a classification problem (accident severity 1 or 2). Decision tree modeling was selected in this case considering its advantages in reflecting importance ranking of attributes in tree hierarchy and ease of interpretation. Model's accuracy was evaluated, and classification report was generated including precision, recall and f1 scores. Recall score is particularly of interest because for rare event prediction like road accident prediction or medical diagnosis, it is preferred to have a higher recall with lower precision because false positives could correspond to high-risk situations that we probably want to detect too.

For 1st round, all features were used, and the model accuracy is 0.65. The recall for predicting severity code-2 (injury) is 0.76. From the tree hierarchy below, pedestrians, bicycles and person count involved in accidents are the most important features, followed by vehicle count involved and hit-a-parked-car situation. It can be easily interpreted from the tree structure that if there are pedestrians, bicycles or more people involved in the accidents, more likely to have injuries and if it's hit-a-parked car situation, less chance of injury. From these observations, the recommendation is to have public traffic safety polices that are more protective of pedestrians and cycles on roads like having more marked crosswalks, bicycle lanes, pedestrian having right of way, more road warning signs for pedestrian and bicycles passing and so on.



```
              precision    recall  f1-score   support

           1       0.66      0.52      0.58      9693
           2       0.64      0.76      0.70     11142

   micro avg       0.65      0.65      0.65     20835
   macro avg       0.65      0.64      0.64     20835
weighted avg       0.65      0.65      0.64     20835
```

Table 2 1st classification report

4. Conclusion

In this case study, collision sample dataset was analyzed, and decision tree machine learning was used to predict the accident severity level. Pedestrian, bicycle and people count involved are the most important features predicting whether an accident has injury. Meantime, accident address type (intersection) and dangerous driving behaviors (driver under-influence, inattention, speeding) are the most important factors that affects accident severity. These observations from the models can be very helpful to guide traffic polices to focus on most important factors to prevent accident injuries.