Applied Data Science Capstone – Predicting Car Accident Severity

Sreeja Matturu

Sept 28, 2020

Introduction

As per 2020, more than 100,000 people were facing accidental deaths. According to them, road accidents are one of the major categories where more than 38,000 people were dying due to many factors. It means that every day more than 100 people were dying because of road accidents. These accidents occur due to some of the major factors that can be influenced by weather, road condition, and also depends upon the driver like whether they are driving under any influence, etc. These accidents mainly occur in the highways especially when the weather is bad, the driver cannot be able to see other vehicles properly. If it rains heavy for suppose, the driver cannot be able to see the road and other vehicles such that, these may cause an accident that can cost their lifetime. Accidents can be categorized into two parts. One is the minor accidents and the other one is major accidents. Minor accidents include the small type of accidents where the vehicle gets a little damaged and the person has minor injuries and can recover within a short amount of time, whereas major accidents can cost a lifetime by completely damaging the vehicle and person gets heavy injuries that might take a long time to recover or sometimes death. These accidents majorly happen in big cities in the United States. For example, Seattle is a big city where there will be huge traffic and lots of vehicles travel every day on the highway. Due to that, it also carries many road accidents every day.

The main problem behind these road accidents in Seattle or any other city in the US is that the people do not know about the road conditions or sudden weather changes across some areas which are hard to predict. I propose to develop a machine learning model to predict the severity of the road accidents based on the weather, road condition and help the people to know whether it is safe or dangerous to drive on that road. If the driver travels along the danger route every day like, weather changes or if the road gets damaged, he/she can drive slowly in that road. If not for the everyday drivers, other people can decide whether they want to drive on that road or take a detour. This model does not only helps the drivers but also the road workers to know the bad road condition place so that they can go and repair the road as soon as possible to prevent accidents.

Data Collection

The data set is a CSV file format that consists of different road accidents that occurred in Seattle. The data set was acquired from Kaggle.com where it contains 37 different types of attributes and more than 190k rows of data describing each accident at each time. This data set contains outliers which are represented as NaN and thus represented as an unbalanced dataset. It includes different types of attributes like injuries, pedCounts, vehicle counts, and fatalities. Features that I want to use from this data set include categorical variables like type of collision, collision codes, whether or not the pedestrian was granted right of way, road, and weather conditions.

- 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

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.

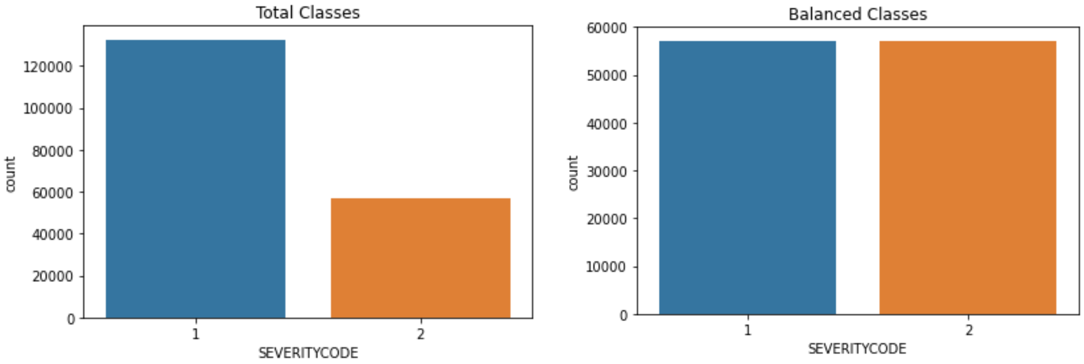


Figure 1 Dataset count before and after rebalancing

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.

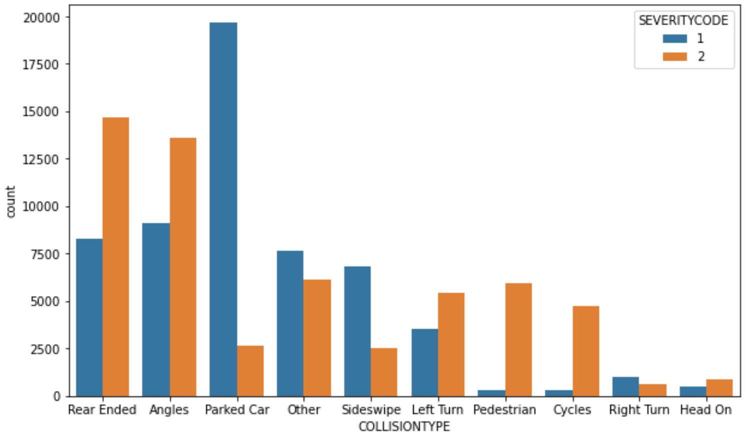
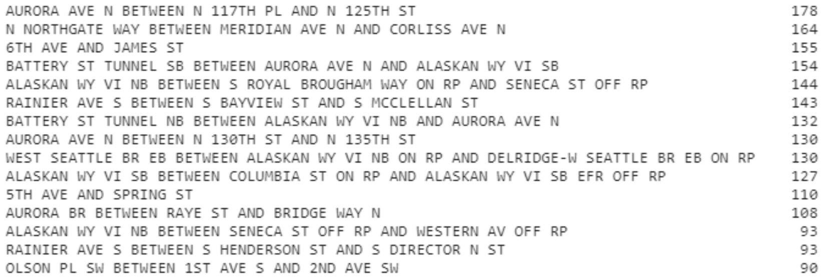


Figure 2 Count plot by collision type

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.

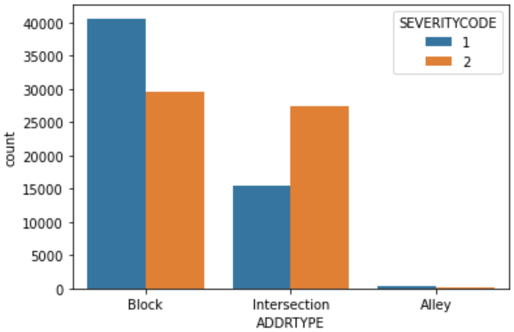


Figure 3 Count plot by address type

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 “unknown” 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.

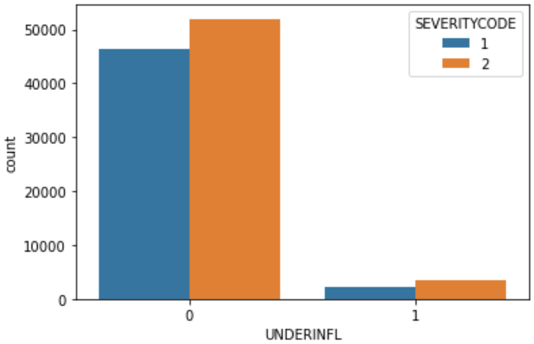


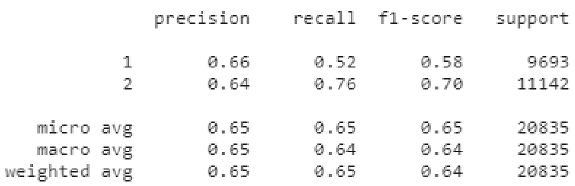
Figure 4 Count plot by whether-under-influence

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.

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 0.64. The recall for predicting severitycode-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.

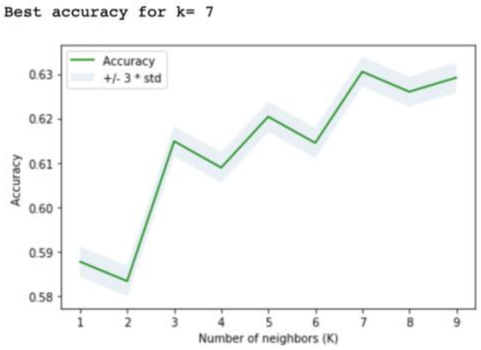


```
              precision    recall  f1-score   support

           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

Since above features are mainly about accident impact scale (people and vehicles involved) and collision type, they were not selected for 2nd round of modeling in order to better assess factors which could cause severe accidents more likely to happen. The model accuracy is 0.60 and the recall for predicting severitycode-2 (injury) is 0.66. From the tree hierarchy, address type, driver under-influence, driver inattention are the most important features, followed by speeding, light condition, weather and road condition. It shows the intersection is highly susceptible for severe accident occurrence and dangerous driving behaviors cause severe accidents more likely to happen especially under less favorable driving conditions. The recommendation from this analysis for traffic operations and polices includes better design of intersection with safer lanes, increased visibility, more guidance signs, as well as reinforced public education of safe driving, warning/reminder signs, strict punishment measures to prevent dangerous driving behaviors.

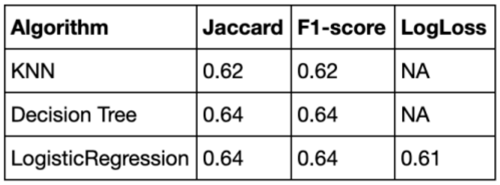
Along with that, K-nearest neighbor and Logistic Regression algorithms were also applied to check the accuracy in predicting the car accident severity. I calculated the evaluated metrics using Jaccard index, f1-score and log loss. When KNN algorithm was applied for 10 clusters, it was shown that the value of k = 7 was the best number of clusters to get the accurate prediction.



```
Best accuracy for k= 7
```

Figure 5 K-value Estimation for Accuracy

Below table shows the evaluation metrics for each machine learning algorithm that was applied.



Algorithm	Jaccard	F1-score	LogLoss
KNN	0.62	0.62	NA
Decision Tree	0.64	0.64	NA
LogisticRegression	0.64	0.64	0.61

Table 3 Evaluation Metrics Results

Conclusion:

In this case study, collision sample dataset was analyzed, and decision tree, KNN and Logistic Regression machine learning techniques were 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.