Project: Car Accidents Severity

Introduction

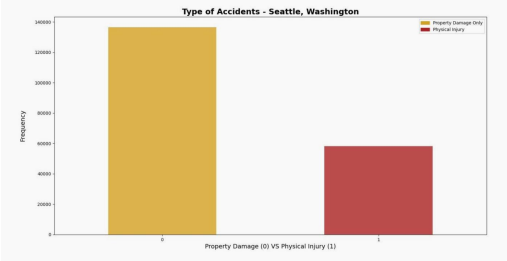
Urbanization is the main reason for the increase in the number of vehicles on the road. In this timely world, everyone is in their own thoughts and rush, which can lead to traffic accidents. There are many factors leading to the accident. It depends on weather, road conditions, vehicle conditions, driver conditions, etc. The main purpose of our project is to reduce car accidents. Here we deal with a data set of Seattle traffic records served as the management department in Seattle since 2004. These data include a collision occurred, and all necessary data.
This report is on what caused the accident and how to take measures to reduce these accidents

Dataset

This is an extensive data set provided by the Seattle Police Department, which has collected more than 190,000 observations in the past 15 years. In order to accurately build the model to prevent accidents and/or reduce their severity in the future, we will use the some attributes as indexes.

Methology

Considering that the feature set and target variables are categorical variables, for example, weather, road conditions, and light conditions are categorical variables higher than level 2, and their values are limited, usually based on a specific finite group, and their correlation may represent different images. Then what it actually is. Usually, it is important to consider the impact of these variables in a car accident, so these variables are selected. Some picture descriptions of the data set are to better understand the data.



I started by importing the csv file and to prepare the data, I dropped the columns we do not need from the dataset, i.e., columns that do not have values or where the values are unknown. Even though this is an important factor, I dropped Speeding entirely because it is missing over 180,000 values and this can hamper the results.

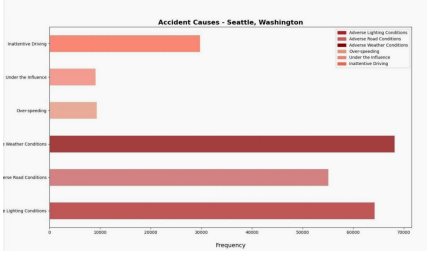
```
In [6]: df_vech['SEVERITYDESC'].value_counts()

Out[6]: Property Damage Only Collision    136485
        Injury Collision                   58188
        Name: SEVERITYDESC, dtype: int64
```

Of the total number of accidents in Seattle, most of them are property damage, and only collisions are more than injuries.

```
Car_Accidents = Car_Accidents[Car_Accidents['ROADCOND'] != 'Unknown']
Car_Accidents = Car_Accidents[Car_Accidents['WEATHER'] != 'Unknown']
```

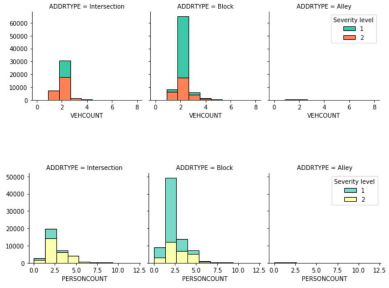
Upon further inspection, I found out that ROADCOND and WEATHER have unknown values. This will again hamper the analysis therefore I dropped the values where there is no information.



In order to balance the goal, the SMOTE in the imblearn library is used to obtain an unbiased classification model that trains instances of two elements under the same conditions of accident severity.

Results

I continue to understand the severity of the accident based on the selected variables. I noticed that accidents at intersections are more serious (level 2-injury), while most non-serious accidents (level 1-property damage) occur at intersections. I also found that the most serious accident occurred at an intersection and involved 2 to 3 people.



As we can see, most accidents occur during the day, on dry roads, not even at intersections and on the road during sunny weather. Most of them are not over speeding and the impact of collisions is very rare.

Discussion

At the beginning of the analysis, I tried to calculate the severity and frequency of road accidents based on weather conditions, road conditions and other factors. Even if our data is of the right

size, there are still many missing elements. We need to clean the data to get good results. We must delete speed because there are too many elements missing, but I think this is an important factor that should be considered. It is clear from the analysis that most accidents involve drivers, on wet roads, in bad weather, at intersections, and of very small nature. This may help police departments understand where to install more stop signs, or may add cameras at intersections to force people to slow down. We also live in a technology-friendly world, so maybe we can develop some built-in technology in the car to warn us when the road and weather conditions are bad or the car is approaching a stop sign.

Conclusion

When comparing all models by score, we can get a clearer understanding of the severity of the car accident in terms of the accuracy of the three models as a whole and their performance-Seattle, Washington 14 for each output of the target variable. When comparing these scores, we can see that the f1 score is the highest score of k nearest neighbors of 0.75. However, when we compare the accuracy later and recall each model, we can see that the accuracy of the k nearest neighbor model is 1 in 0.08. The variance is too large to choose this model as a viable option. When looking at the other two models, we can see that the decision tree has a more balanced degree between 0 and 1. When looking back at 0 and 1, the logistic regression is more balanced. The average f1-scores of the two models are very close, but for Logistic regression, it is 0.04 higher. It can be concluded that the two models can be used side by side for best performance.