Coursera Data Science Capstone Report

Seattle Accident Severity

By Trevon Tewari

14th September 2020.

Introduction/Business Understanding

In a world where commuting is inevitable, for work, leisure or otherwise, accidents seem to be unavoidable due to the sheer density of cars on the road and a need to get from point A to B as fast as possible. Accidents also tend to affect others as a source of traffic congestion, especially in areas of higher population density, rendering transport as a time-consuming exercise and reducing productivity.

In a motor vehicle, the risk of an accident is always present, however, its likelihood as well as severity can be affected by many external factors such as visibility, road conditions, weather conditions etc. These factors are mostly out of our control, however, it would be extremely helpful if we were to know just how much these factors can contribute to the severity of an accident happening, to educate the driver to exercise more caution in certain conditions and circumstances as well as to provide alternative routes to avoid such an unfortunate event. The following study is intended for the education of drivers with respect to how they should exercise caution in various conditions.

Data

To gain insight into the contributions of external factors to collisions, we will review the Collisions - All Years dataset which comprises traffic collision data from the Seattle Police Department from the year 2004 to present.

The dataset was obtained via the Applied Data Science Capstone course on Coursera as the sample dataset for the project.

First and foremost, we identify the SEVERITYCODE attribute which is our target variable for this study and is a numeric reflection of the severity of the accident caused in terms of the degree of injury. We then look at some samples from the data and realize that most of the data are categorized as objects. Additionally, many are also worded descriptions of other attributes in the dataset. Many of these initial samples seem to be missing data and some seem to be keys used for identifying the report and its respective details.

Methodology

Description based values in the table were filtered out as they do not contribute numerically to the exercise. Some of the attributes are also observed to be taken as a response to an accident happening such as PERSONCOUNT, PEDCOUNT and PEDCYLCOUNT. These may skew the models from predicting severity in terms of external conditions as the SEVERITYCODE does not account for the number of people involved, just the degree of injury or fatality. This study will not consider the location of the accidents but can do so in future, to narrow the data in terms of areas of high density.

Attributes such as SPEEDING and INATTENTIONIND may have proven very useful to the study but contained too many null values to be used in the model and must be removed.

This leaves us with WEATHER, LIGHTCOND and ROADCOND, where we now explore their distributions:

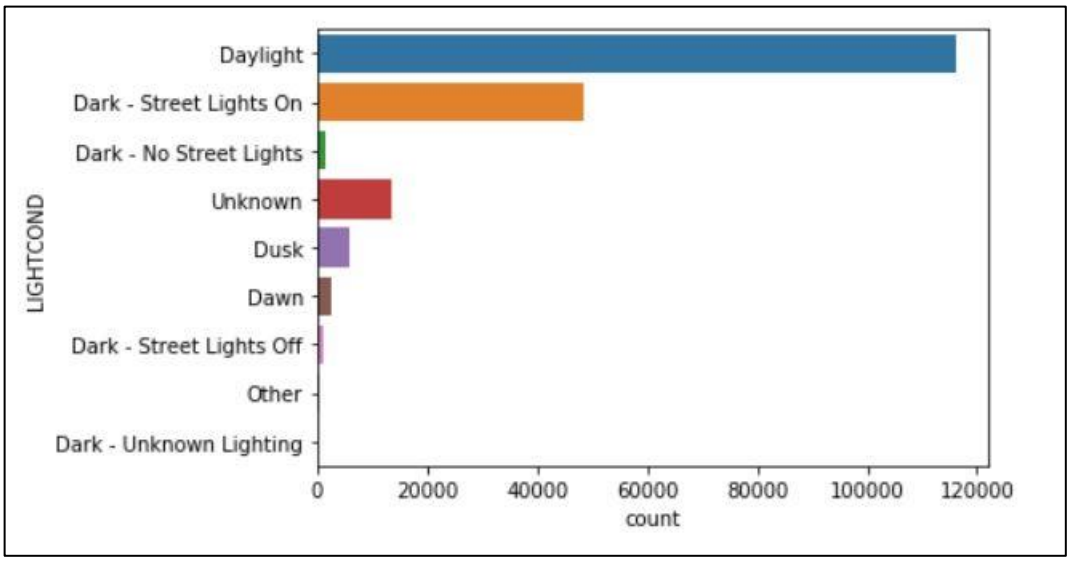

Figure 1: Count plot of LIGHTCOND.

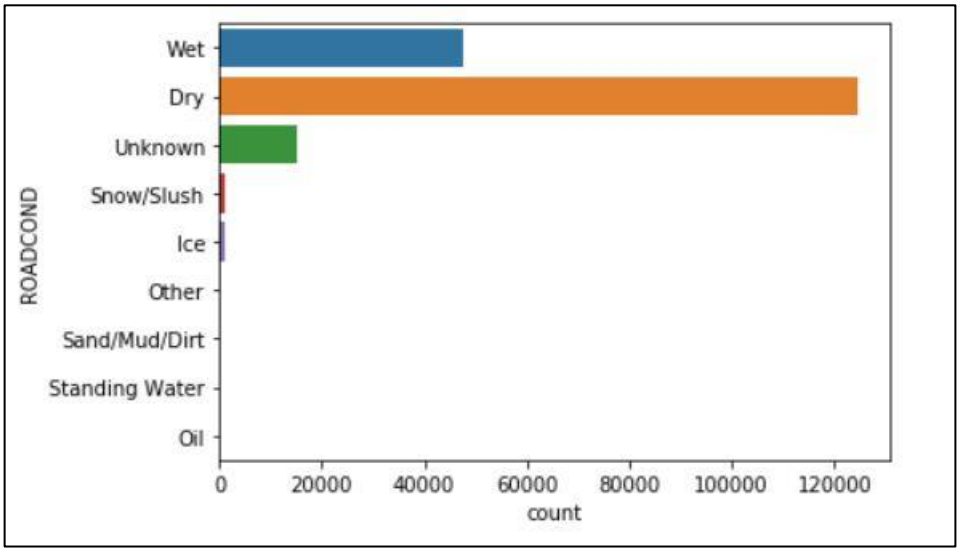

Figure 2: Count plot of ROADCOND.

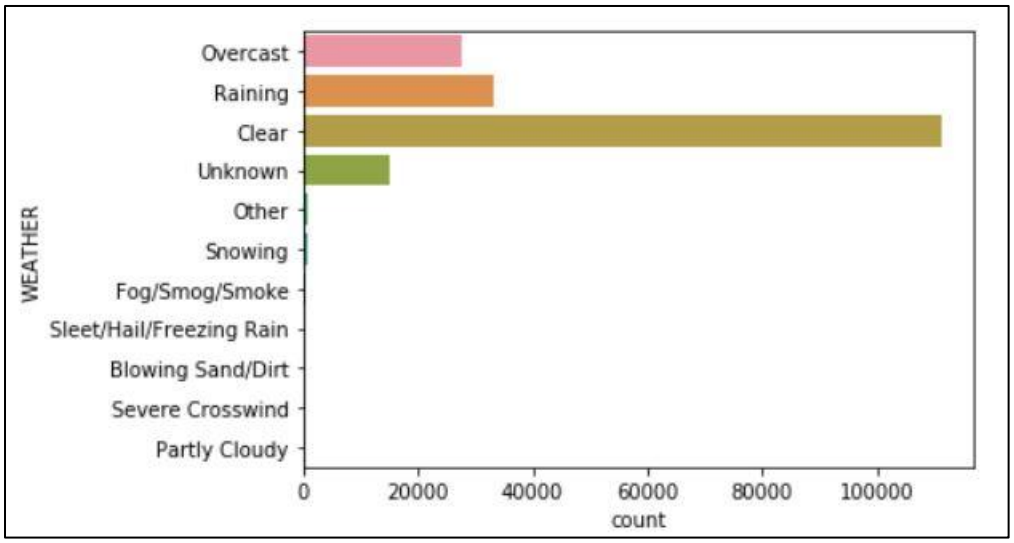

Figure 3: Count plot of WEATHER.

They are therefore chosen for the study.

	SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND
0	2	Overcast	Wet	Daylight
1	1	Raining	Wet	Dark - Street Lights On
2	1	Overcast	Dry	Daylight
3	1	Clear	Dry	Daylight
4	2	Raining	Wet	Daylight

Figure 4: Table of the chosen attributes.

The WEATHER, LIGHTCOND and ROADCOND are also observed to contain a lot of null values. Therefore, the corresponding rows are removed. All three attributes are also shown to have an “Unknown” and “Other” category, which are small relative to the size of the dataset and does not contribute to the effectiveness of the study, as a result, these values are removed.

```
df2['LIGHTCOND'].value_counts()
```

```
Daylight                    116077
Dark - Street Lights On      48440
Unknown                      13456
Dusk                          5889
Dawn                          2502
Dark - No Street Lights       1535
Dark - Street Lights Off      1192
Other                          235
Dark - Unknown Lighting         11
Name: LIGHTCOND, dtype: int64
```

Figure 5: Figure showing categories in LIGHTCOND.

We now look at the contents of the target variable, SEVERITYCODE, and we can see that it is heavily imbalanced. This can cause bias in our models and should be fixed. To do so, the dataset is shuffled, and the category of lesser value put aside and counted. The number of samples required are then taken from the category in excess and the datasets are then concatenated to form a randomized, balanced dataset as shown:

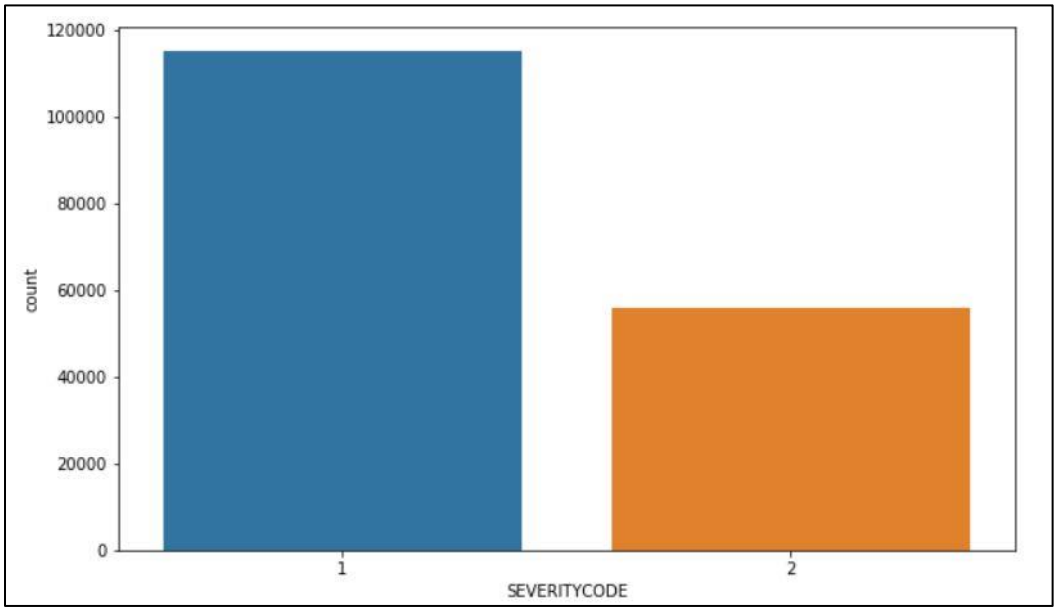

Figure 6: Count plot of SEVERITYCODE showing imbalance.

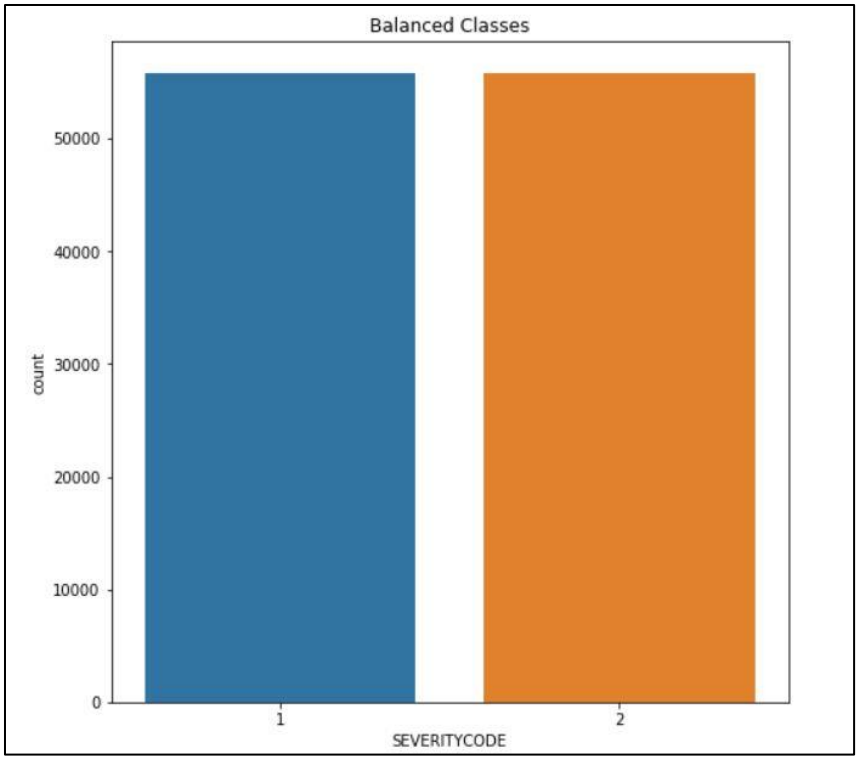

Figure 7: Count plot of SEVERITYCODE after balancing.

Now, the attributes chosen are also observed to be objects, meaning that they are text-based categories and cannot be used for modelling in this state. As a result, one hot encoding is used to enumerate the attributes as shown. To regulate the values, the scaler is used.

Results

The dataset is then split into the training and testing sets using a 70/30 split. The following classification models are then designed and applied in each case:

K Nearest Neighbours

The K Nearest Neighbours model is applied on the dataset, using the range of K from 1 to 30 and verifying the best choice for K using the mean accuracy. Therefore, the K that provides the highest mean accuracy is the best choice for the model. In our case, K was found to be 9 with a mean accuracy of 0.507.

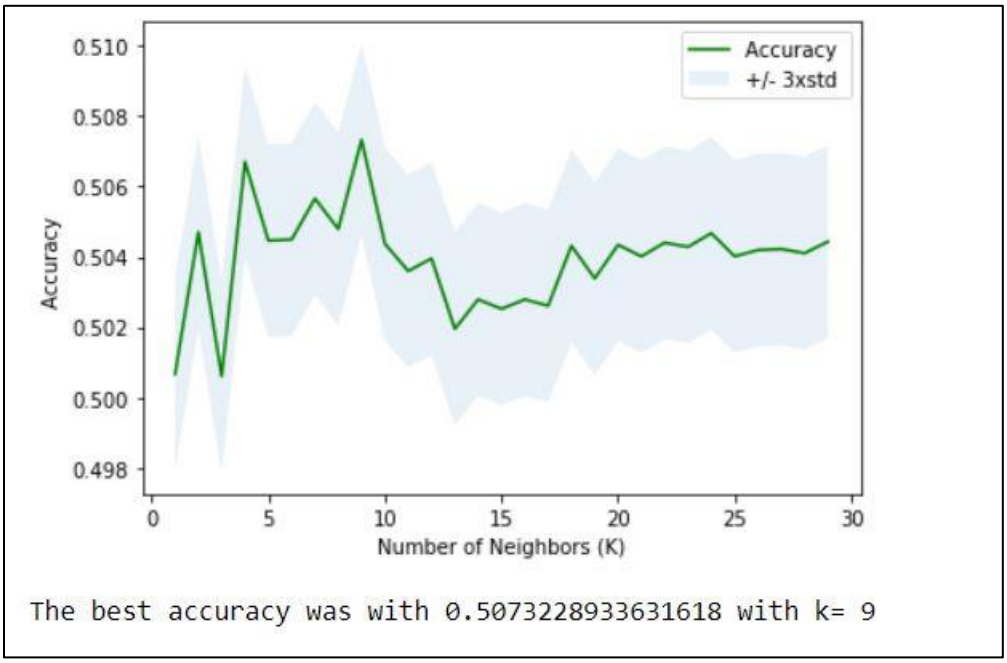

Figure 8: Finding the best K value.

The most effective K is then applied to the model for use on the test data later.

Support Vector Matrices

The support vector matrices model is then applied on the dataset using the rbf kernel. The accuracy score was found to be 0.516.

Logistic Regression

The logistic regression model was then applied on the dataset with C as 0.01 and the liblinear solver. The accuracy score was found to be 0.515.

Decision Trees

The decision trees model was then applied to the dataset with a max level of 6. The accuracy score was found to be 0.517.

The Jaccard similarity score and f1 scores were then found for all four models as well as the log loss for the logistic regression model as shown in the table below.

Algorithm	Jaccard	F1	LogLoss
KNN	0.507323	0.606822	NA
SVM	0.516450	0.334004	NA
LogReg	0.515377	0.390242	0.692251
Decision Tree	0.516957	0.333306	NA

Figure 9: Figure Showing Results obtained from models.

Discussion and Recommendations

As we can observe, the K Nearest Neighbours seemed to have the highest F1 score of 0.606 and the lowest jaccard index of 0.507. The other models had much lower f1 scores, around 0.3 and slightly higher jaccard index scores, around 0.51. Of the models used, the K Nearest Neighbours may have had the best performance. However, these models can generally be deemed as having poor performances as all the scores can be considered as relatively low.

Although the weather, road conditions and lighting conditions are initially expected to have a significant impact on accident severity, these low performances indicate that the WEATHER, ROADCOND and LIGHTCOND attributes are not sufficient. Useful attributes such as SPEEDING and INATTENTIONIND may have been very useful for these models as these are expected to be significant in more severe accidents, but too many null values rendered them unusable. Certain categories could have been grouped together such as the varieties in LIGHTCOND with variations of Dark. Additionally, location data could have been used to assist in narrowing the areas where accidents most occur to help clean some outlying data and allow drivers to be more aware of certain areas.

What could not be considered, due to the limitations of this dataset, was the contributions of these external conditions to the likelihood of an accident being caused rather than its severity. It can be stipulated that attributes such as the speed of the vehicle and the quantified inattention of the driver may have significantly contributed to the severity of accidents.

Conclusion

In conclusion, we can say that while it was expected that weather, lighting, and road conditions would have had an impact on the severity of accidents, this was not directly reflected in the study as shown. It is more likely that these external conditions may tend to affect the likelihood of an accident, rather than the severity.