Car Accident Severity Predictions

IBM Data Science Capstone Project, October 2020

1. Introduction

1.1 Background

Automobile incidents can have a varying severity. Depending on the severity, the outcome can a small or large impact on other people. A lower severity may have a smaller impact on emergency resources and traffic patterns. While a large severity would require a greater response and drastically impact traffic patterns. Being able to know the difference would provide a better way to respond to an incident, decrease the severity of on an incident, or, better yet, being able to prevent an incident. The data collected from previous accidents can be utilized to better improve awareness and predictability of future accidents. The data can be manipulated and used to train models to better inform drivers and key decision makers (e.g. Emergency dispatchers) in response to traffic incidents.

1.2 Problem

The problem at hand is to use data available to be able to provide a useful prediction to stakeholders. Stakeholders such as traffic enforcement, emergency responders, and commuters are some examples of those who could be helped by exploring this data.

2. Data acquisition and cleaning

2.1 Data Source

The project allowed for any open source for the data as long as it had a feature similar to the example data set. In the end, the data set chosen for the Capstone project was the example data provided by Coursera. With the dataset, there is provided information on the Metadata that helps add clarification or details to data attributes.

2.2 Data Cleaning

Upon review of the Metadata, it can be seen that there are several usable attributes from the dataset. Furthermore, there are also several attributes that are administrative codes and references that will need to be removed in part of the data cleaning. Below in Table 1 - Attribute of Example Dataset that were removed is a list of the information in the example dataset that was not useful and was removed to reduce the information included in a dataframe. There is no point in wasting computing time by carrying this information forward.Table 1 - Attribute of Example Dataset that were removed

Table 1 - Attribute of Example Dataset that were removed

Attribute	Reason to remove
'OBJECTID', 'INCKEY', 'COLDETKEY','REPORTNO','STATUS', 'LOCATION', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC', 'SEVERITYDESC', 'SDOT_COLDESC', 'ST_COLDESC',	Reference values to data tables, database keys, descriptions, etc.
'INTKEY', 'EXCEPTRSNCODE', 'SDOTCOLNUM', 'PEDROWNOTGRNT', 'SEGLANEKEY', 'CROSSWALKKEY'	Mostly Null values or missing data
'SEVERITYCODE.1', 'INCDATE'	Duplicate data/information

Next, there is a few of the key attribute that contained null values. It is suspected that over the years either systems might have been upgraded and information might not have carried forth. Another possible reason is that the information was just never recorded to begin with. Of the two cases, it is most likely human error and the latter. Due to some of the fields being categorical (i.e. Weather), it would be difficult to determine and fill in that missing data. Since the data set is relatively large compared to the null values, it seemed reasonable to just drop the rows with null values.

Secondly, it was noticed that there were two fields with a significant amount of data missing: “SPEEDING” and “INATTENTIONIND”. The description from the Metadata said “SPEEDING” was “Whether or not speeding was a factor in the collision. (Y/N).” Similarly, the description for “INATTENTION” was “Whether or not collision was due to inattention (Y/N).” It was early on hypothesized that that if these two fields were indicative of a “Yes”, that it would have a greater influence over the severity. Therefore, the decision was made to covert the data to binary where 0 would mean No or Unknown and 1 would be a affirmed case. This was expected to be explored further into the data analysis.

Thirdly, many of the driving conditions were different categories of string inputs. However, what was missed on the initial null-value removal is that most of these categories contained a small number of “Unknown” fields - which were removed. Using “Unknown” as a predictor would only overfit the model.

Lastly, the datetime attribute “INCDTTM” had mixed data. It appeared that older data before 2010 did not include a time portion. If this was converted without being managed, it would screw the data with adding erroneous incidents at time “00:00:00” or at 0 hour of the day during the analysis. Consequently, a new column in the dataframe was created with the time portion only. Once that was created, the dataframe had any entries removed that contained precisely “00:00:00” resulting from the datetime conversion.

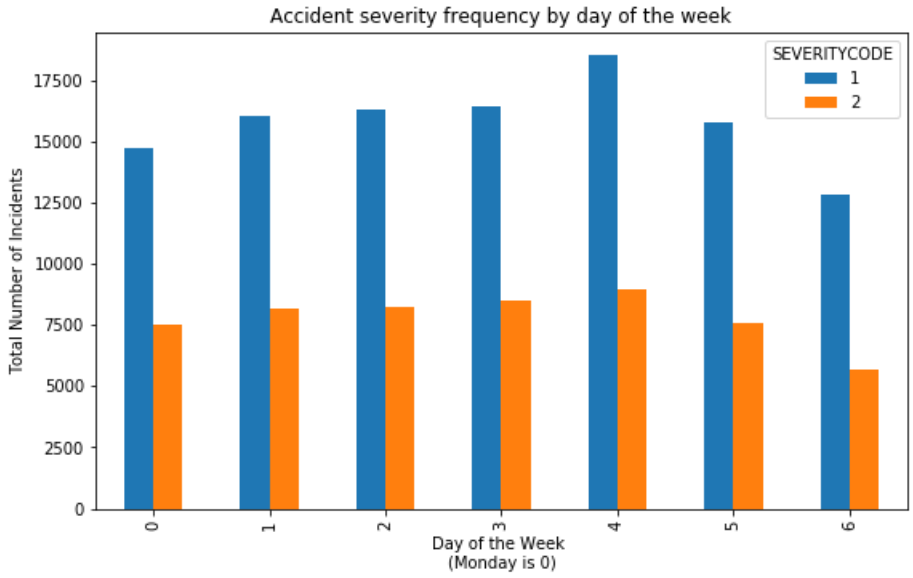
3. Exploratory Data Analysis

3.1 Date and Time factor of incidents

Traffic is busier and more congested during some parts of the day compared to others. Furthermore, some months of the year have more treacherous driving conditions. The hypothesis is that time can be a possible factor to help aid in determining when accidents and different severities occur.

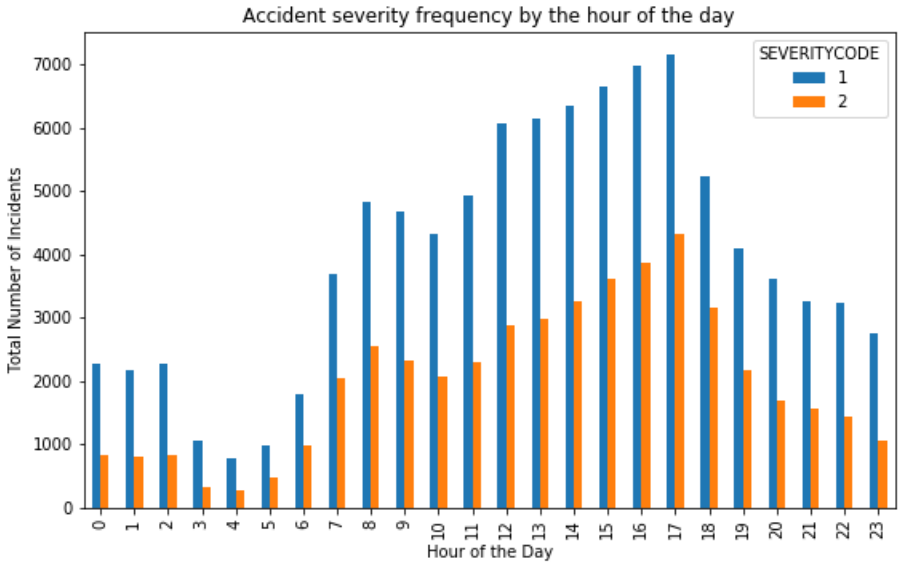
A few observations can be made from viewing the data by days of the week show in Figure 1 - Incident Severities for days of the week Looking at the distribution, it appears to show that overall incident rate totals are on Fridays. Secondly, Saturdays show that there is as lower amount of severity code 2 incidents to code 1. Lastly, Sunday shows an overall incident total lower on Sundays. This could be useful for the model in predicting the severity depending on which day of the week the incident occurs.

Figure 1 - Incident Severities for days of the week



The time of day also plays a big factor on when incidents are expected to occur. In Figure 2 - Incident Occurrences by hour of the day, the hours of 3 through 6 am tend to have the lower amount of incidents and steadily increase afterwards. The trend in incident rates starts declining from 6 pm, which makes sense since this is after the normal business hours and when people have returned home.

Figure 2 - Incident Occurrences by hour of the day



The last time periods do not provide very good data for modelling. While Figure 3 - Incident Occurrence totals for 2004 through 2020 shows there are trends in the traffic accidents by years, this will not be useful for predicting current year nor future year data. Using the yearly data could overfit the data to the training data and make it less useful. Lastly, the monthly data, while statistically significant, does not show a significant trend in line with seasonal expectations (eg winter driving being worse). However, this is possible limited by the location (Seattle) and could be useful for a different data set in other locations. See Figure 4 - Incident Occurrences by month of the year below for details.

Figure 3 - Incident Occurrence totals for 2004 through 2020

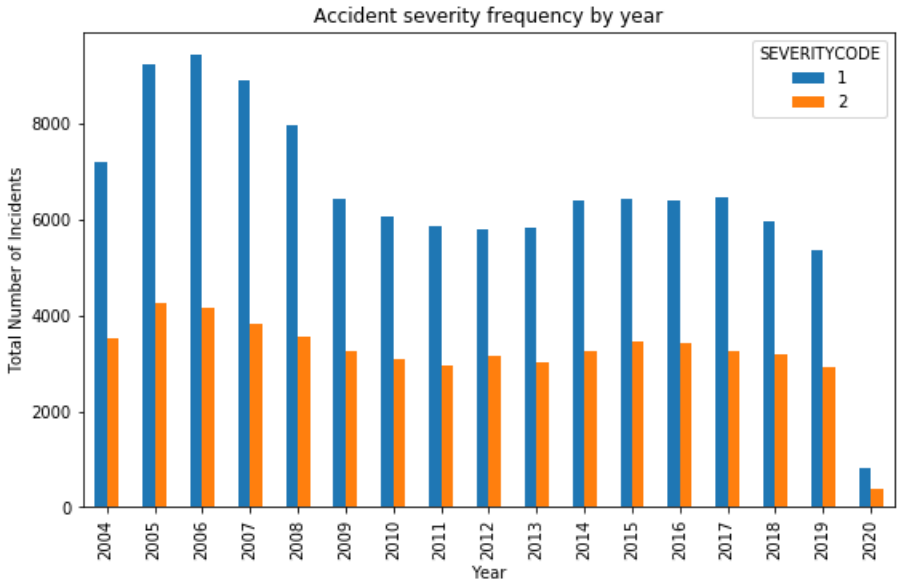
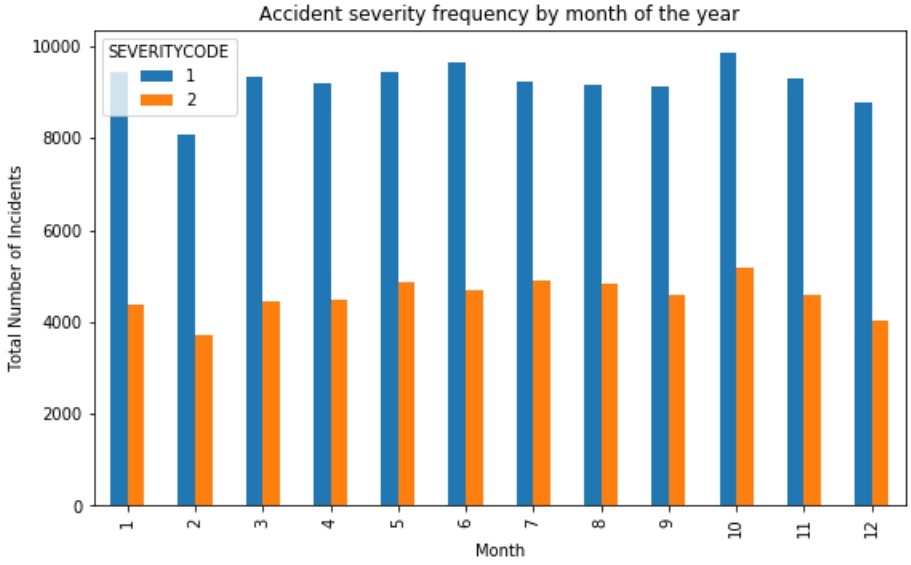


Figure 4 - Incident Occurrences by month of the year



3.2 Driving/Road Conditions

There were three different attributes that were grouped together for driving/road conditions: weather, road, and lighting conditions. They were analyzed individually to see if there was any specific factor that lead to higher ratio Severity 1 codes versus Severity 2 codes. It was quickly realized that there was an imbalance of data between the number of Severity 1 and Severity 2 occurrences.

The first condition that was investigated was weather. The majority of the weather conditions recorded were “Clear”, “Raining, and “Overcast.” These three conditions accounted for 98.9% of the incidents. Additionally, Severity 1 accounted for two-thirds for almost of the incidents shown in Table 2 - Severity totals for weather condition. The only exceptions were “Sleet/Hail/Freezing Rain”, “Partly Cloudy”, and “Snowing.” For “Partly Cloudy”, there were only 5 reported incidents and therefore was not of significance. Similarly, “Sleet/Hail/Freezing Rain” had less than 100 incidents. This could suggest that there might be larger error due to smaller numbers. Unexpectedly, accidents were deemed Severity 1 for 80% of cases involving “Snowing”.

Table 2 - Severity totals for weather condition

SEVERITYCODE	Blowing Sand/Dirt	Clear	Fog/Smog/Smoke	Other	Overcast	Partly Cloudy	Raining	Severe Crosswind	Sleet/Hail/Freezing Rain	Snowing
1	25.0	61171.0	322.0	131.0	15202.0	2.0	17817.0	16.0	65.0	555.0
2	13.0	30681.0	165.0	58.0	7309.0	3.0	9311.0	7.0	23.0	138.0

The next driving condition isolated was road conditions and the distribution of severity can been seen in Table 3 - Severity totals for road conditions. The majority of the roads conditions were either “Dry” or “Wet” and amounted to 98.6% of the total road conditions. Similar to the data in Table 2 - Severity totals for weather condition, the main data showed about a two-third ratio of Severity 1 to 2. Correspondingly, the “Ice” and “Snow/Slush” showed a much lower occurrences of Severity 2. This makes sense considering those road condition occur during winter weather discussed for the previous data.

Table 3 - Severity totals for road conditions

SEVERITYCODE	Dry	Ice	Oil	Other	Sand/Mud/Dirt	Snow/Slush	Standing Water	Wet
1	68294.0	726.0	22.0	52.0	29.0	578.0	52.0	25553.0
2	34138.0	225.0	15.0	34.0	16.0	135.0	23.0	13122.0

The last driving condition analyzed was for lighting conditions. As can be seen in Table 4 - Severity total for lighting conditions, approximately 93% of the lighting during accidents was either Daylight or Dark with Streetlights on. Low light situations (Dawn and Dusk) only accounted for about 4.6% of the accidents. Similar to the other majority of weather and road conditions, accidents occurred at a rate of approximately 2 Severity 1 codes to 1 Severity 2 codes for lightning conditions. It is noted that when its dark and there are streetlights available, there is a slightly higher (68%) occurrence of Severity 1 codes.

Table 4 - Severity total for lighting conditions

SEVERITYCODE	Dark - No Street Lights	Dark - Street Lights Off	Dark - Street Lights On	Dark - Unknown Lighting	Dawn	Daylight	Dusk	Other
1	881.0	634.0	26970.0	4.0	1295.0	62367.0	3058.0	97.0
2	269.0	269.0	12184.0	4.0	701.0	32652.0	1597.0	32.0

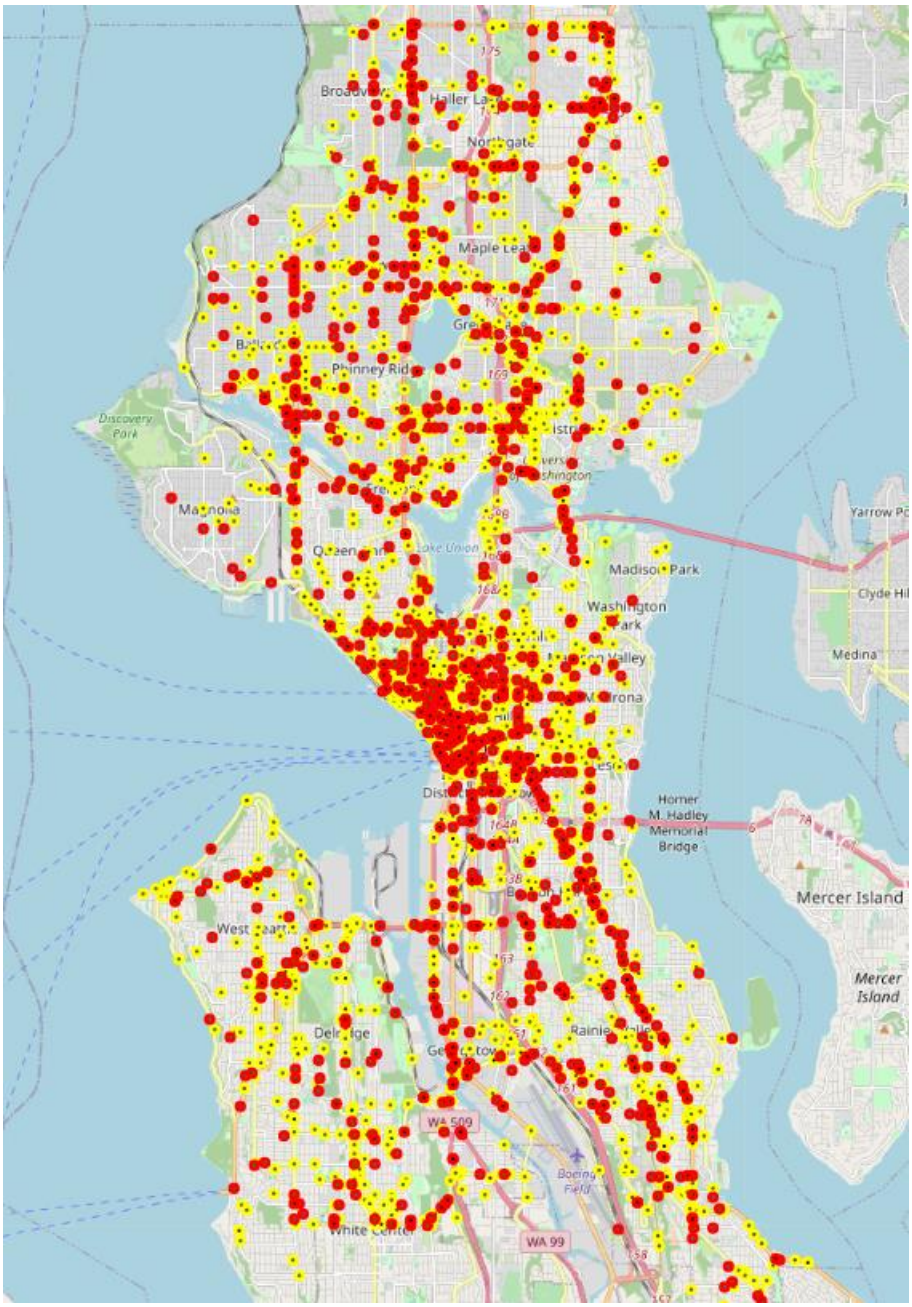
Overall, there weren't many correlations noted in the road or driving conditions. The majority of the data that did not appear to occur at a 2:1 ratio were for the categories that did not occur regularly.

3.3 Mapping of Incidents

One way that the data can be understood is by plotting the incidents where the occur to visualize the issue. This was accomplished by installing the Folium package and leveraging the longitude and latitude of each reported incident. Since the number of incidents were well over 100,000, I choose to randomly select 3000 points to show. As seen in Figure 5 - Folium map with incidents overlayed. Yellow is Severity 1 and red is Severity 2, with only 3000 markers, the might starts to get congested at that zoom level. Additionally, using 5000, the processing time takes noticeably longer and no extra information was gained.

At first glance, it's easy to a majority of all accident occur downtown. Further to previously analyzed data, you visually see there are roughly two yellow dots (Severity 1) to one red dot (Severity 2). However, upon further inspection, there is a bit of a trend to see that red dots trace along the larger roads and highways. In addition, red dots do seem occur more at intersections than mid-section. This suggests keeping and using the 'JUNCTIONTYPE' attribute in the model would be helpfully. Furthermore, using location data could help improve the determination of the severity of an accident.

Figure 5 - Folium map with incidents overlayed. Yellow is Severity 1 and red is Severity 2 (3000 random points).



4. Predictive Modeling

4.1 Model Selection

4.1.1 Decision Tree

Taking a look at the data given, it is essentially all categorical and not continuous. Thinking about this suggests a Decision Tree might be the best model to break down the randomness and be able to predict a severity. Therefore, a Decision Tree model will be trained and tested for its effective to predict an incident severity.

The categorical data for the driving/road conditions had to be transformed by converting with the LabelEncodeer method. This converted the categories to numerical values for the model. Afterwards, the data was split into testing and training data to help evaluate the model. The decision tree was initial tested using a max_depth of 4 but was later determined that 10 was more suitable. This can be seen below in the Model Evaluation section.

4.1.2 K-Nearest Neighbor

It is prudent to check more than one model to see which is best suited to predict the data. Due to the categorical and discrete nature of the data, K-Nearest Neighbor (KNN) would be another useful model to utilize. The data preparation and encoding use from the previous model could be used for the KNN model.

4.2 Model Evaluation

4.2.1 Decision Tree

Since the data is imbalanced, one of the best ways to evaluate the model was using a confusion matrix. Before the data was upsampled, the model had an F1 score for Severity 1 was 0.82. However, the predictability was very poor for Severity 2. As is included in Figure 6 - Confusion Matrix of Decision tree model (before upsampling), the recall was very low for Severity 2. In other words, true values a Severity 2 incident were only predicted correctly 35% of the time. Using the upsampling technique and increasing the amount of Severity 2 occurences, the model was able to improve the recall to 72% (Figure 7).

Figure 6 - Confusion Matrix of Decision tree model (before upsampling)

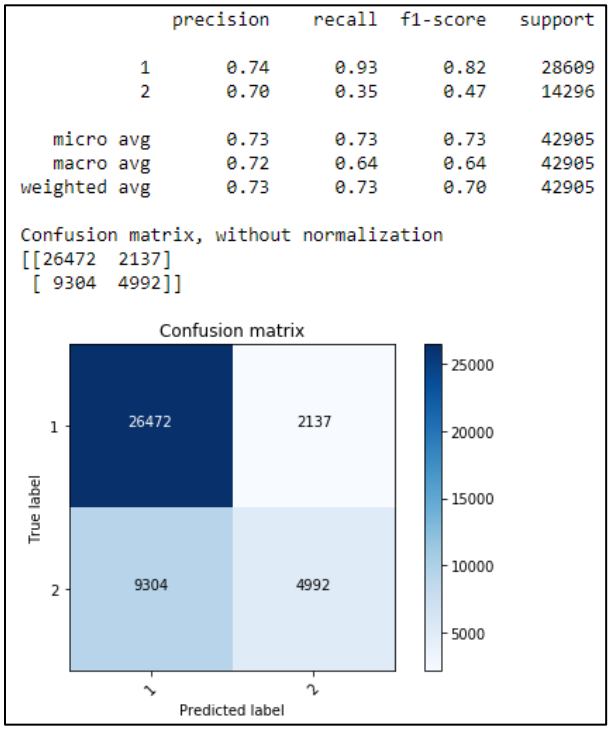
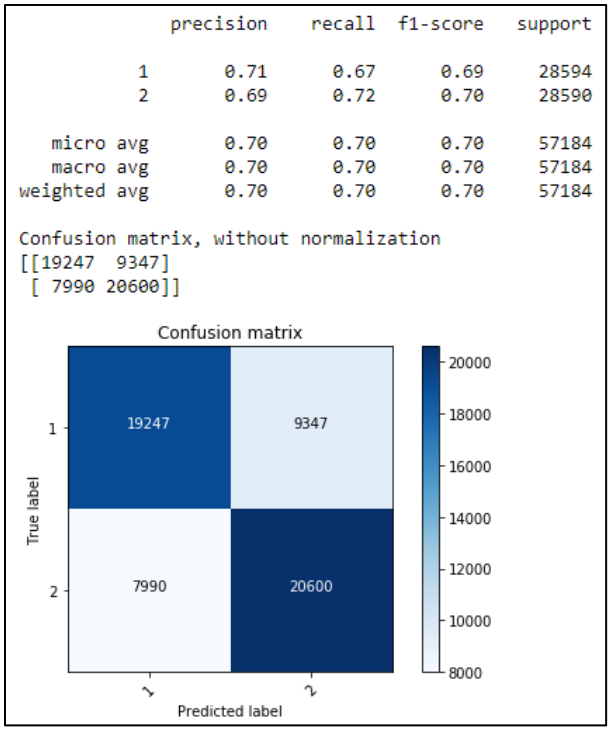


Figure 7 - Confusion Matrix of Decision tree model (after upsampling)



4.2.2 K-Nearest Neighbor

Different values for k gave different results for the KNN model. It was prudent to check for several values of k and compare F1_score to see which would be the best predictor for the model. From the graph in Figure 8 - F1-Score for each k value selected, it can be seen that the value of k=1 gave the best F1_score. The k=1 value was used to train the model and evaluated using a confusion matrix (Figure 9 - Confusion Matrix of KNN model, k=1 (after upsampling)). The precision of predicting Severity 1 and Severity 2 was 86% and 77%, respectively. Furthermore, it can see that the model using k=1 had a high recall and that both severities had the true values predicted accurately. In other words, almost three-quarters of Severity 1 were predicted correctly while Severity 2 was at 88%.

Figure 8 - F1-Score for each k value selected

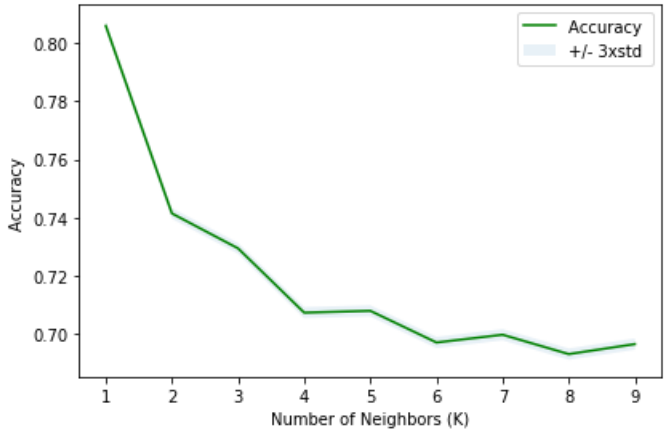
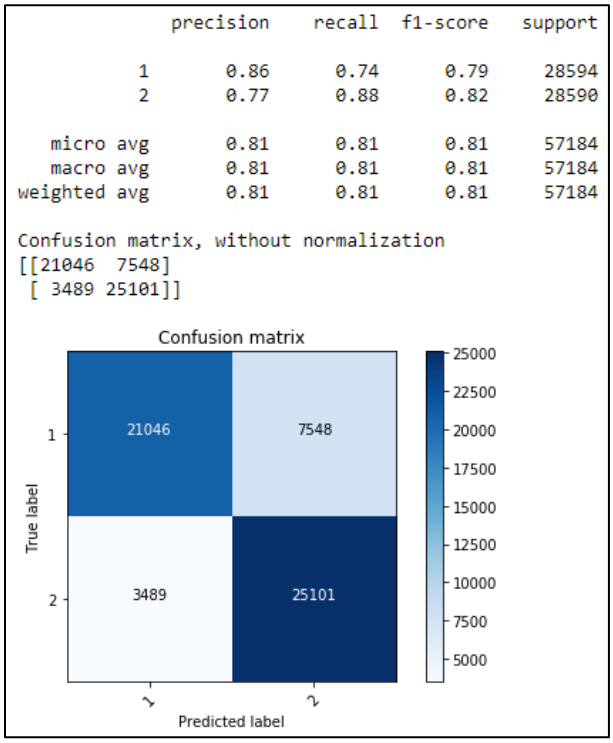


Figure 9 - Confusion Matrix of KNN model, k=1 (after upsampling)



	precision	recall	f1-score	support
1	0.86	0.74	0.79	28594
2	0.77	0.88	0.82	28590
micro avg	0.81	0.81	0.81	57184
macro avg	0.81	0.81	0.81	57184
weighted avg	0.81	0.81	0.81	57184

```
Confusion matrix, without normalization
[[21046  7548]
 [ 3489 25101]]
```

5. Conclusion

In the Capstone project, the data was analyzed for different driving conditions such as road conditions, light conditions, and weather conditions. Unexpectedly, the severity of an accident did not increase in severity with adverse driving conditions. Severity 2 accidents actually occurred less frequently. The data was analyzed for accidents at different times, days of the week, or months. It was observed that there were trends to the frequency of accidents and severity depending on the time frame looked at. Lastly, it was determined that the Speeding and Inattention data was either insufficient in size or an ineffective indicator in predicting the severity of an accident.

The data was used to train both a Decision Tree and KNN model. The Decision tree model was able predict the severity of an accident with better reproducibility than average occurrence of Severity 1 to 2. The improvement was approximately 4-6%. After evaluation of the KNN model, the predictability of the model provided an improvement upon the Decision tree model. Furthermore, a higher percentage of true values were correctly predicted. This is especially important when providing warning to drivers on driving risk or correctly sending out emergency services to accidents.