IBM Applied Data Science Capstone Report

1. Introduction

1.1 Background

Car accidents are a very significant cause of serious injury and death in the United States. According to the National Safety Council, 38,800 people lost their lives in car crashes in 2019, and about 4.4 million people were in accidents severe enough to require that they seek medical attention for their injuries. Some of these individuals will suffer from permanent disability, and possibly chronic pain. This results in significant distress, decreased quality of life, and financial burden for many individuals every year. Car crashes also have greater societal impacts in the form of lost productivity and subsequent impacts on the economy, as well as being burdens on the medical system.

There are a great number of factors that may contribute to whether an accident will be severe or minor. These might include, but are not limited to, road conditions, light conditions, the weather, driver inattention, driver intoxication, speeding, and the area in which people are driving. This project seeks to determine if there are any reliable indicators that particular conditions significantly elevate the risk of getting into a more serious car accident – one that involves bodily injury rather than just property damage. This project will focus on one geographical area, Seattle, and attempt to determine if there are significant risk factors for car accident related injuries that may be addressed so as to improve outcomes.

1.2 Interest

There are numerous parties who may have an interest in this type of assessment. From a public health standpoint, government bodies could use the information gathered from this study to determine the best changes that can be made to infrastructure in order to reduce fatalities and serious injuries (such as increasing lighting in certain areas, dealing more aggressively with poor road conditions, such as heavy snow, or modifying how streets or intersections are constructed), or to determine the best driver behaviours to target with policing or public education efforts (such as distracted driving, intoxicated driving, or speeding).

Private companies may have an interest in the assessment as well, as the information gathered here may allow for the development of consumer products aimed at improving driver safety. The manufacturers of GPS units in vehicles, or tech companies that provide driving and GPS related products (such as Google Maps), could use this type of information to provide added value to their customers. This could look like expanding the functionality of GPS units to include current weather, road conditions, lighting conditions, or other information that would output an accident risk assessment to the user, and provide the user with guidance on how they should proceed according to the current risk level.

2. Data Acquisition and Cleaning

2.1 Data Sources

I opted to use the provided data set for the course, as I am very new to data science and coding, and wished to keep things relatively simple so that I could focus on the rather large number of new skills I need to practice. As such, the data was acquired from the IBM Applied Data Science Capstone course content. Additional information on the data set was acquired from the Seattle

GeoData website (https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0).

The dataset originates from Seattle Department of Transportation, and represents data from 2004 to the present. It provides 194,673 observations, with most observations having 38 attributes. The attributes are diverse, and include things such as the weather, road conditions, lighting conditions, whether or not a driver was speeding, the number of people involved, the seriousness of injuries, details of how the accident occurred, where the accident occurred, and much more.

2.2 Data Cleaning and Feature Selection

My goal for this project was to focus on attributes that are most easily observable and/or modifiable prior to an accident occurring. As such, I decided to use weather, road conditions, light conditions, speeding, inattention, being under an influence, and address type (intersection, block, or alley). Other attributes were omitted because:

1) They were too similar to the outcome being predicted (accident severity), such as injuries, serious injuries, fatalities, and collision type.
2) They contained information that could only be determined after the accident occurred, and would be unlikely to be helpful in the goal of reducing accidents through public health or technological intervention.
3) They were simply not useful for my intended approach (e.g. GPS coordinates, ESRI identifiers, etc.)

For attributes with a binary classification, i.e. driver inattention and speeding, null values were assumed to be “no” (values were either null or “Y”), and the data was updated accordingly. Null values and values indicated as “unknown” for all other attributes resulted in the observation being removed, as there was no reasonable method of filling in the data with a mean, or another similar measure. This is primarily due to the fact that all the attributes being assessed contain categorical values. At this time, the number of observations being utilized has been reduced to 169,578, with 8 attributes (including the target attribute) associated with each observation.

The target variable, accident severity, is limited to either a “1” or “2” in the data set, despite the fact that the accompanying documentation indicates 5 potential severity codes. This is therefore to be considered a binary outcome. As such, solving the desired problem requires a predictive classification model, which necessitates a machine learning approach. Many of the machine learning approaches covered within the certification courses may be applied, including K-Nearest Neighbours, Decision Trees, Logistic Regression, and Support Vector Machine. I intend to assess the accuracy of each prior to determining which would be best suited to the task of predicting accident severity. Measures of accuracy such as Jaccard Index, F1-score, and Log Loss will be utilized for this purpose.

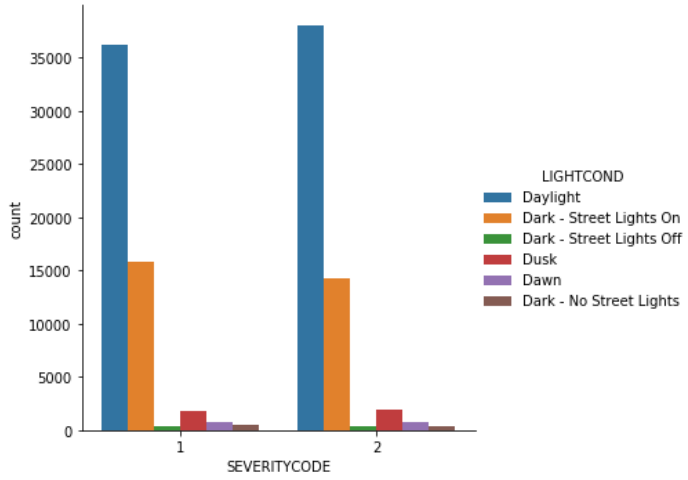
Because the number of severity 1 accidents far outweighed the severity 2 accidents, a random selection of 55654 observations were selected from the severity 1 subset in order the balance the data. This reduced the number of observations being used to 111,308.

3. Exploratory Data Analysis

3.1 Relationship Between Accident Severity and Light Conditions

Accidents were more likely to be severe at dawn, dusk, and surprisingly, during daylight hours. The poorest driving conditions, i.e. darkness with no street lights, were the conditions least likely to result in injury accidents relative to non-injury accidents.

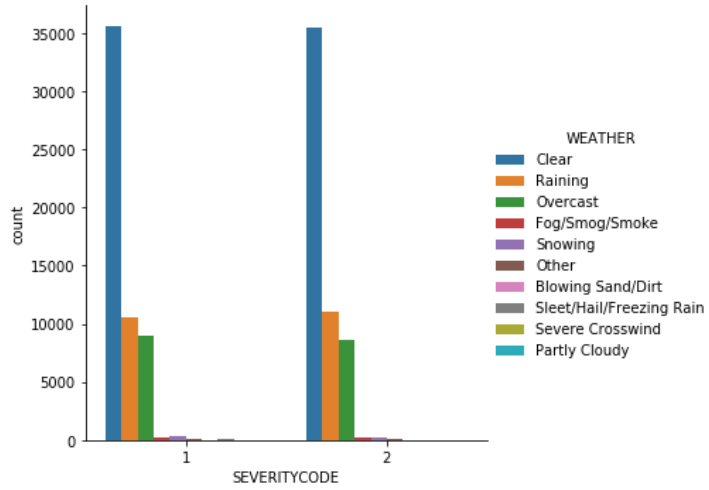
LIGHTCOND	SEVERITYCODE	
Dark - No Street Lights	1	0.616114
	2	0.383886
Dark - Street Lights Off	1	0.569038
	2	0.430962
Dark - Street Lights On	1	0.526361
	2	0.473639
Dawn	2	0.516603
	1	0.483397
Daylight	2	0.511860
	1	0.488140
Dusk	2	0.509106
	1	0.490894



3.2 Relationship Between Accident Severity and Weather

The weather conditions most likely to result in injury accidents included Fog/Smog/Smoke, Partly Cloudy, and Raining. Surprisingly, snow and sleet/hail/freezing rain were the least likely to result in injury accidents relative to non-injury accidents.

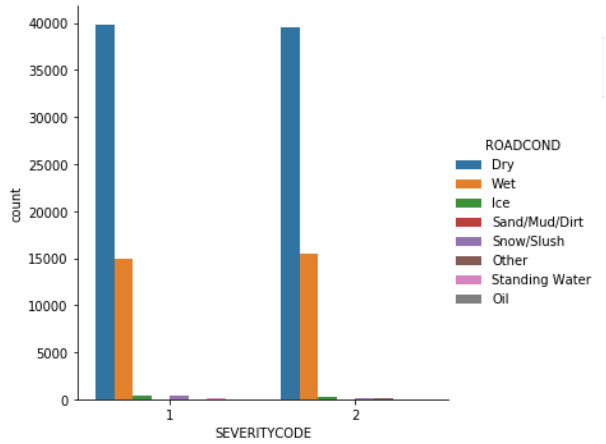
WEATHER	SEVERITYCODE	
Blowing Sand/Dirt	1	0.535714
	2	0.464286
Clear	1	0.500401
	2	0.499599
Fog/Smog/Smoke	2	0.505556
	1	0.494444
Other	1	0.519231
	2	0.480769
Overcast	1	0.509291
	2	0.490709
Partly Cloudy	2	1.000000
Raining	2	0.512900
	1	0.487100
Severe Crosswind	1	0.500000
	2	0.500000
Sleet/Hail/Freezing Rain	1	0.602941
	2	0.397059
Snowing	1	0.663223
	2	0.336777



3.3 Relationship Between Accident Severity and Road Conditions

The most likely road conditions to result in injury accidents were oil, other, sand/mud/dirt, and wet. Snow/Slush and Ice were not more likely to result in injury accidents relative to non-injury accidents. When viewed in conjunction with the previous sections, it appears that especially poor driving conditions are more likely to result in property damage type collisions. This could be due to drivers being more highly alert and cautious when conditions are particularly bad, resulting in slower impact speeds.

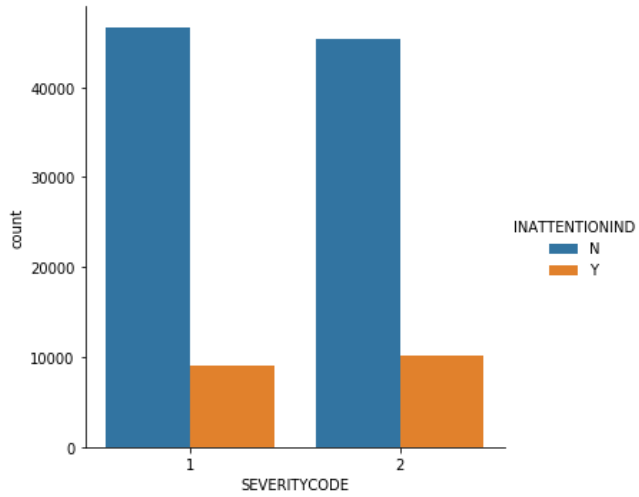
ROADCOND	SEVERITYCODE	
Dry	1	0.501303
	2	0.498697
Ice	1	0.607988
	2	0.392012
Oil	2	0.615385
	1	0.384615
Other	2	0.583333
	1	0.416667
Sand/Mud/Dirt	2	0.512195
	1	0.487805
Snow/Slush	1	0.674419
	2	0.325581
Standing Water	1	0.640000
	2	0.360000
Wet	2	0.508474
	1	0.491526



3.4 Relationship Between Accident Severity and Driver Inattention

Not surprisingly, accidents were more likely to involve bodily injury when a driver was deemed to be inattentive.

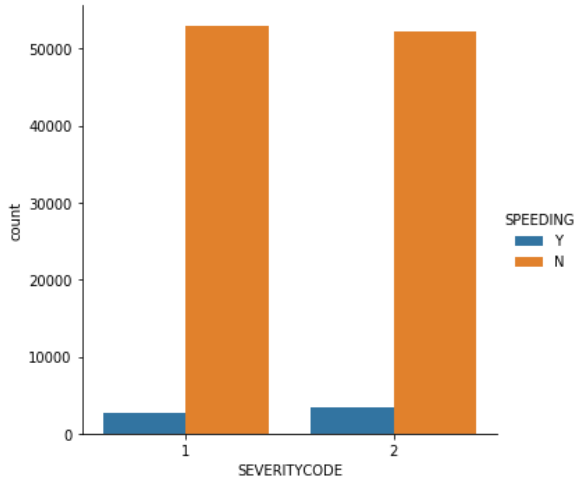
INATTENTIONIND	SEVERITYCODE	
N	1	0.506576
	2	0.493424
Y	2	0.531479
	1	0.468521



3.5 Relationship Between Accident Severity and Speeding

Also unsurprisingly, accidents were more likely to involve bodily injury when a driver was deemed to be speeding.

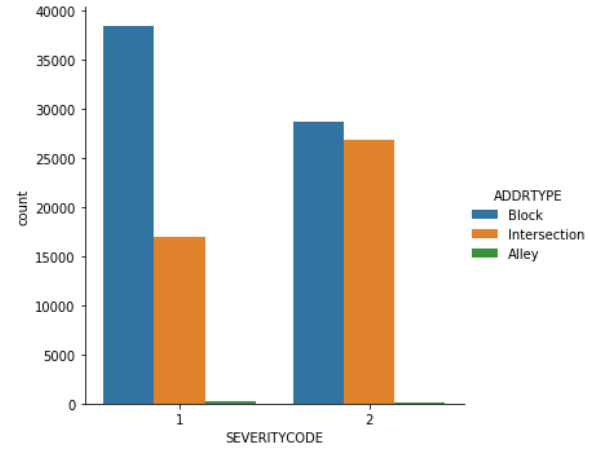
SPEEDING	SEVERITYCODE	
N	1	0.503604
	2	0.496396
Y	2	0.561586
	1	0.438414



3.6 Relationship Between Accident Severity and Address Type

This turned out to be the most prominent determinant of accident severity from the data visualization section. Accidents were more likely to involve injury if the accident occurred in an intersection, and were significantly less likely to involve injury if the accident occurred in an alley. The latter is logical, as drivers are unlikely to be traveling at high speeds in an alley.

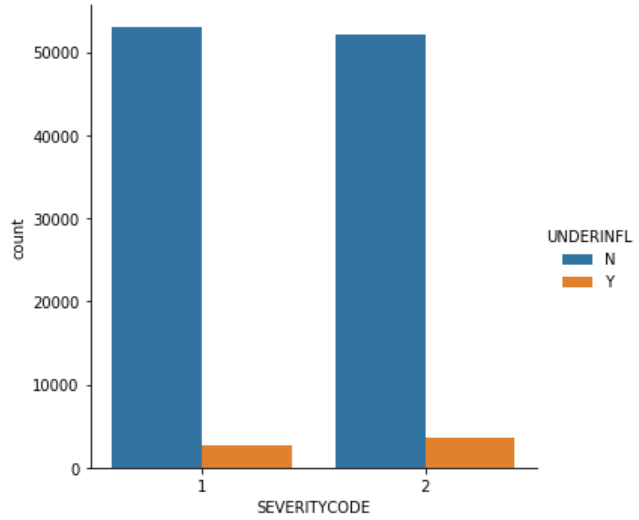
ADDRTYPE	SEVERITYCODE	
Alley	1	0.765244
	2	0.234756
Block	1	0.572375
	2	0.427625
Intersection	2	0.613045
	1	0.386955



3.7 Relationship Between Accident Severity and Being Under an Influence

A driver being under the influence of alcohol or drugs was more likely to result in a more serious accident. Similar to inattention and speeding, this was not a surprising finding.

UNDERINFL	SEVERITYCODE	
N	1	0.503966
	2	0.496034
Y	2	0.567563
	1	0.432437



3.8 Additional Comments on Exploratory Data Analysis

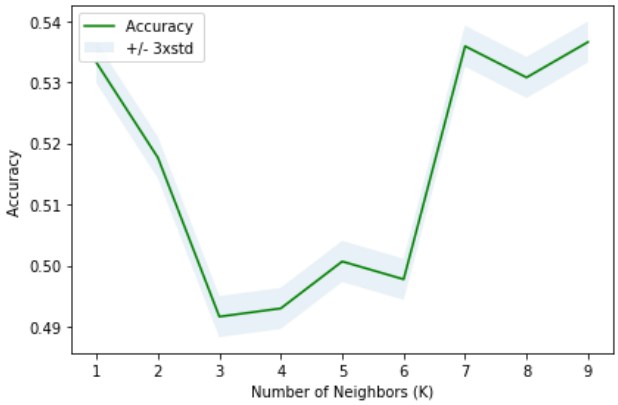
One of the challenges associated with visualizing and assessing this particular data set is that all of the data attributes are categorical, which significantly limits options for graph selection. This characteristic may also provide challenges with the machine learning models, which will be discussed in the following section.

4. Predictive Modeling

Because the dataset's attributes are all categorical, including the target variable, classification models were used. I decided to test 4 different models in order to assess which would perform the best for the question at hand. A train/test split was created, with 80% of the data being used for training.

4.1 K Nearest Neighbours

A KNN model was created with a starting k value of 4, using the Scikit-learn library. The model performed poorly, with a train set accuracy of 0.503 and a test set accuracy of 0.596. As assessment to determine the optimum value of k was performed, and the best performing k value was found to be 9.



The model was retrained with k = 9, which produce a train set accuracy of 0.542 and a test set accuracy of 0.536. Although the k value adjustment improved the consistency between train and test set accuracy, the predictive ability of the model performs barely better than a coin toss, and is therefore rather poor.

4.2 Decision Tree

Using the Scikit-learn library, a decision tree model was created with a max depth of 20, to ensure that the numerous classification type attributes could be accommodated. This model produced an accuracy score of 0.596.

4.3 Support Vector Machine

An SVM model was created using the Scikit-learn library. It produced an accuracy of 0.597.

4.4 Logistic Regression

A logistic regression model was created using the Scikit-learn library. The probability predictions produced by the model were poor, with the highest confidence observed in the first 5 outputs being 0.626. The general accuracy of the model was 0.596.

```
#predict probabilities
yhat_prob = LR.predict_proba(X_test)
yhat_prob[0:5]
```

```
array([[0.58662205, 0.41337795],
       [0.62561619, 0.37438381],
       [0.38966538, 0.61033462],
       [0.58662205, 0.41337795],
       [0.58662205, 0.41337795]])
```

4.5 Model Performance Metrics

	Algorithm	Jaccard	F1-score	Log Loss
0	KNN	0.541309	0.528938	NA
1	Decision Tree	0.592859	0.592369	NA
2	SVM	0.597046	0.596881	NA
3	Logistic Regression	0.593354	0.591432	0.671243

Jaccard and F1-scores for all models were in the range of 0.5 – 0.6. K Nearest Neighbours produced the poorest results, while the Decision Tree, SVM, and Logistic Regression models produced very similar results. The SVM model produced marginally superior scores compared to the other models. Log Loss for the Logistic Regression model was higher than I would have liked.

5. Conclusions

On the whole, the performance of the models was disappointing, as the Jaccard Index results indicate that the model predictions perform only slightly better than a coin toss. As they currently exist, I would not recommend using these models for any practical application. Further analysis on the models was not completed, as the low predictive accuracy would need to be improved before any meaningful insights could be drawn from the model.

I suspect that many of the attributes may contain too many possible classification options (for example, the weather attribute had 10 possible values, and the road conditions had 8. In future iterations of this project, I would likely attempt to cluster some of these values, and possibly eliminate some of the values that had very low counts, in order to simplify and hopefully improve the model outcomes.

It is also possible that either too many or too few attributes were used for model training – experimenting with this may also lead to some improvements.

I am new to data science (and coding in general), so this project was a very humbling, but quite fun, exercise. It has helped me realize where I need to seek further understanding, particularly when it comes to fully understanding the code so that I can trouble shoot errors. For example, I must admit that I could not get a visualization of my decision tree to work. I spent a significant amount of time Googling the error, pulling up the Scikit-learn documentation, and trying many things to fix it, but did not have any success. I suspect this is because I lack the in-depth knowledge of Python that would allow me to fully understand all the components of the code. As this certification program is introductory, I expect that I will need to seek quite a bit more knowledge in order to fully understand not only the code, but also the statistics and mathematical concepts involved in producing high quality analyses and machine learning models. This capstone has provided an excellent opportunity to test my current knowledge level. It has been very encouraging to see that I can pull together some functional code that will at the very least run without errors, and that I am quite comfortable with certain components of the project. The exercise has also helped me understand the areas in which I need to improve, which will assist me with focusing the next steps on my data science journey.

6. Future Directions

There is likely much that could be done to improve the existing model, as discussed in the conclusions section. There are a large number of potential changes that may improve the existing model, including condensing some of the attribute value options, using greater or fewer attributes in the model, or choosing a different set of attributes. Implementation of this model in any capacity would require significantly higher predictive accuracy.

There are some interesting new questions raised that could potentially be explored with new models and data exploration. For example, I found it interesting that when going through the data visualization stage, the most prominent difference between injury and non-injury accidents appeared to be related to where the accident occurred (intersections were more dangerous than blocks or alleys). This observation could be further explored using the dataset, as the details of accidents in intersections could be assessed for angle of impact, involvement of pedestrians, cyclists, or other vehicles, and a number of other factors. This could potentially help provide some ideas on how to best reduce these accidents, whether it's through better infrastructure and intersection design, or improved driver education.

Future iterations of this project may be able to improve model accuracy, however it is also possible that redoing the project from scratch, once I have obtained further training, would be more beneficial. Since this is my very first data science project, I expect that there are many areas in which my code is not optimal, or in which I made less than perfect decisions regarding the data or model development. I do not expect this project to ever be deployed in any capacity, however it would be interesting to consider it again as a future project, perhaps using a data set for somewhere closer to my region, perhaps even the city in which I live. The climate here is certainly a bit different than Seattle, and it would be interesting to see if there are some regional differences in the data.