Predicting Car Accident Severity

Colin Olson

December 9, 2020

1. Introduction

1.1 Background

In the United States, there are around 6 million car accidents every year, on average. In 2019 alone, there were an estimated 38,800 deaths due to car accidents. To put it into perspective, that is a little more than 106 deaths per day. These 6 million accidents result in about 3 million people being injured every year. And out of these 3 million injuries, about 2 million of them are permanent injuries. According to a study conducted by the National Highway Traffic Safety Administration in 2014, the "economic and societal harm from motor vehicle crashes" cost $871 billion in one year. These few statistics show just how big of an issue driver safety is in the United States.

1.2 Problem

The intention for this project is to analyze car accident data in Seattle in order to provide accurate predictions on the severity of a car accident, given various factors and conditions.

1.3 Interest

First and foremost, drivers would be very interested in accurate predictions of the severity of car accidents. Accurate predictions would allow drivers to be more cognizant of their surroundings in more dangerous areas or even avoid these areas entirely. Secondly, several government entities would be very interested in these findings. Accurate predictions would allow the government to improve conditions by enforcing more safety measures in higher risk areas. Emergency personnel and law enforcement could also be stationed closer to these higher risk areas in order to cut down on response time and potentially be able to save a significant amount of lives every year. Lastly, many different private companies would find this information beneficial. Car insurance companies could leverage this data in order to correctly adjust premiums given the severity of potential accidents. Other companies working on new technology to improve driver safety could also use this data to make key business decisions.

2. Data

2.1 Data Source

The collisions data used for this project was obtained through the CSV file shared in class. The labeled dataset was recorded by the city of Seattle, Washington from the year 2004 through 2020. The data shows the details of car collisions, including the severity, time, and conditions under which the collision occurred. You can download the CSV file here and the Metadata file here.

2.2 Data Cleaning

The data from the aforementioned source was downloaded and saved into a table. Many columns in the dataset were added in order to identify the exact occurrence of the accident. These irrelevant columns were dropped from the table since the final goal is to accurately predict the severity of a future accident given the conditions. The table below shows the columns which were dropped and the reasoning behind dropping them.

Dropped Column	Reasoning
OBJECTID, INCKEY, COLDETKEY, REPORTNO, STATUS, INTKEY, EXCEPTRSNCODE, EXCEPTRSNDESC, SDOTCOLNUM	Only added to dataset in order to identify the specific accident.
ST_COLCODE, ST_COLDESC, PEDROWNOTGRNT, COLLISIONTYPE, SDOT_COLCODE, SDOT_COLDESC, HITPARKEDCAR	Only can be known once an accident has occurred.
SEVERITYDESC, SEVERITYCODE.1	Repeat of information given in SEVERITYCODE.
LOCATION, SEGLANEKEY, CROSSWALKKEY, X, Y	Goal of model is to be independent of location. ADDRTYPE and JUNCTIONTYPE were kept as location descriptors.
INCDATE, INCDTTM	Goal of model is to be independent of time. LIGHTCOND was kept as a time of day descriptor.

Table 1 - Dropped Feature Columns

Out of the remaining columns, several contained many missing values and data inconsistencies. Instead of dropping these entries or columns and potentially lose valuable data, I decided to replace the null objects with ‘Other’ or ‘Unknown’. I was able to do this for the categorical variables ADDRTYPE, JUNCTIONTYPE, WEATHER, ROADCOND, and LIGHTCOND. The remaining features which contained null values were INATTENTIONIND, UNDERINFL, and SPEEDING.

The INATTENTIONIND feature is a categorical variable in which there is supposed to be a ‘Y’ for ‘Yes’ values and an ’N’ for ‘No’ values. The data source had an inconsistency in which the ‘Yes’ values were properly labeled with a ‘Y’, but the ‘No’ values were shown as null values. To correct this, the null values were replaced with 0’s and the ‘Y’ values were replaced with 1’s.

The UNDERINFL feature is a categorical variable similar to INATTENTIONIND. However, there was a slightly different data inconsistency with this feature. Some of the 'Yes' values were labeled as 'Y' and some as 1's. Likewise, some of the 'No' values were labeled as 'N' and some as 0's. I decided to convert all of the 1's to 'Y' values and the 0's to 'N' values. Additionally, there were 4884 missing values which I converted to 'UNKNWN' values.

The categorical feature SPEEDING had a data inconsistency similar to INATTENTIONIND in which the 'Yes' values were properly labeled with a 'Y' but the 'No' values were shown as null values. To correct this, the null values were replaced with 0's and the 'Y' values were replaced with 1's. The datatype for SPEEDING was originally listed as float64, but was converted to int64 for uniformity.

2.3 Feature Selection

The final cleaned dataset consisted of 194,657 rows. The goal of this project is to accurately predict the severity of a future car accident based on several conditions. To keep consistent with this goal, none of the features were dropped in order to further understand their correlation in the Exploratory Data Analysis phase.

3. Methodology

3.1 Exploratory Data Analysis

3.1.1 Severity of an Accident vs. Address Type

In order to better understand which address types contributed to which severity of accidents, the four address types were visualized as seen below. It was found that accidents occurring in an intersection resulted in the highest percentage of severity code 2 accidents (i.e, accidents resulting in an injury).

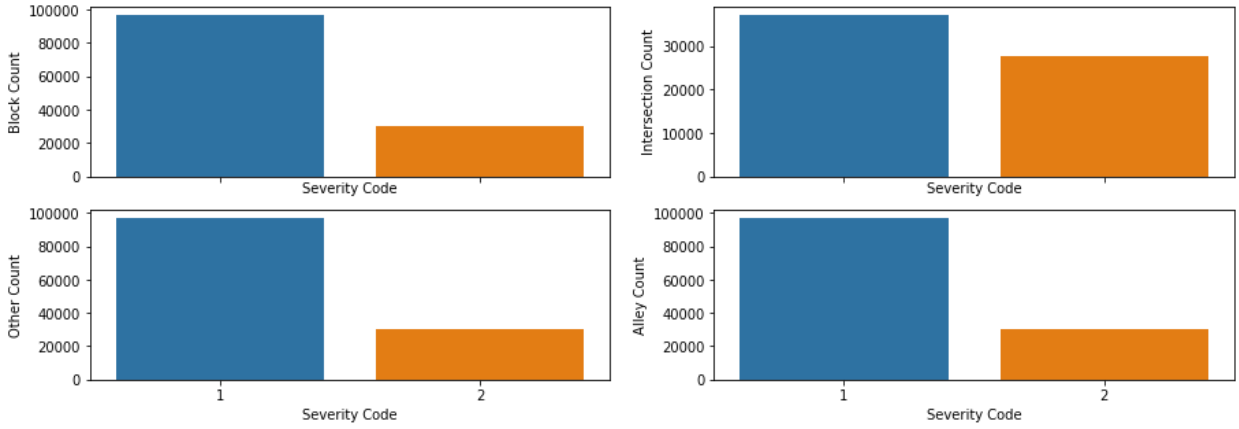

Figure 1 - Address Type vs. Severity

3.1.2 Severity of an Accident vs. Junction Type

In order to better understand which junction types contributed to which severity of accidents, the eight junction types were visualized as seen below. It was found that accidents occurring at an intersection in which the intersection was related to the incident resulted in the highest percentage of severity code 2 accidents.

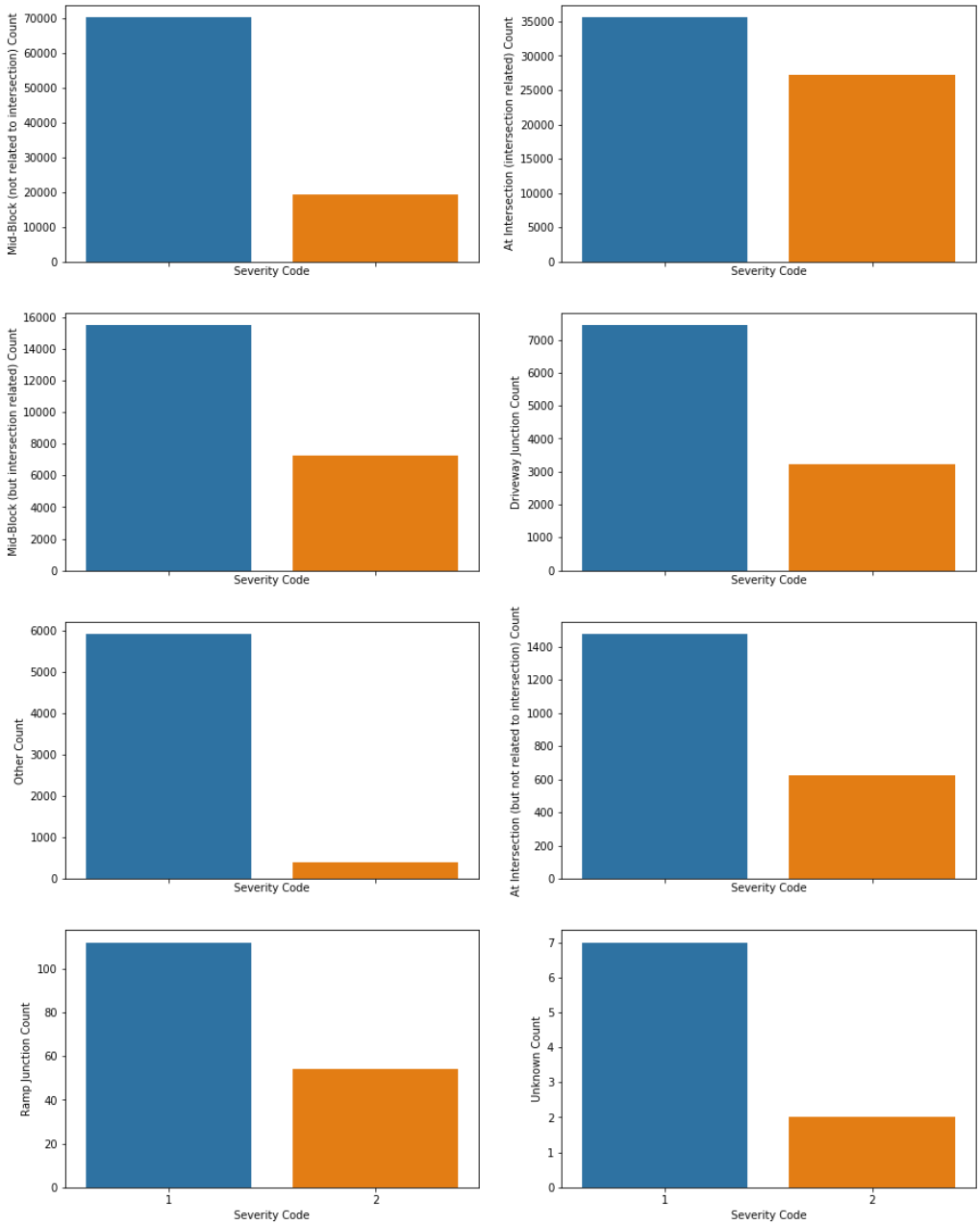

Figure 2 - Junction Type vs. Severity

3.1.3 Severity of an Accident vs. Person Count

In order to better understand how the number of persons involved in an accident affects the severity of the accident, the person count was plotted against the severity codes as shown below. As expected, it was found that the majority of accidents occurred included less than 15 persons. The visualization shows that the higher person count accidents tend to only result in a severity code of 1.

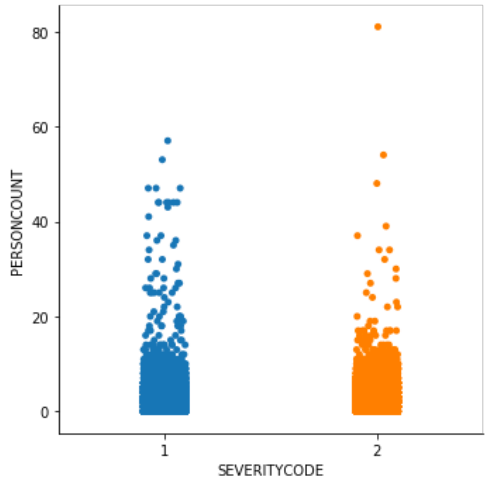

Figure 3 - Person Count vs. Severity

3.1.4 Severity of an Accident vs. Pedestrian Count

In order to better understand how the number of pedestrians involved in an accident affects the severity of the accident, the pedestrian count was plotted against the severity codes as shown below. The figure shows that the higher pedestrian counts involved, the more likely it is that the accident is of severity code 2.

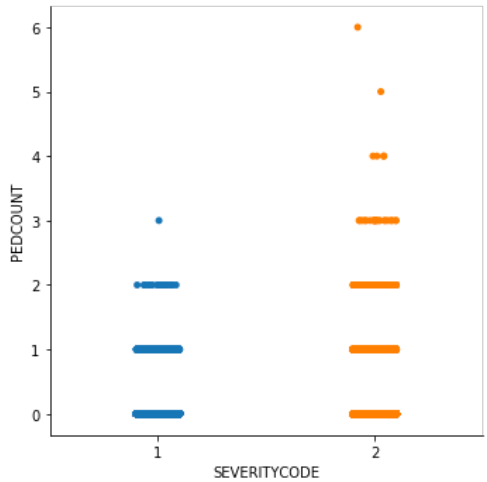

Figure 4 - Pedestrian Count vs. Severity

3.1.5 Severity of an Accident vs. Cyclist Count

In order to better understand how the number of cyclists involved in an accident affects the severity of the accident, the cyclist count was plotted against the severity codes as shown below. The figure shows that if more than one cyclist was involved in an accident, it was always an accident of severity code 2.

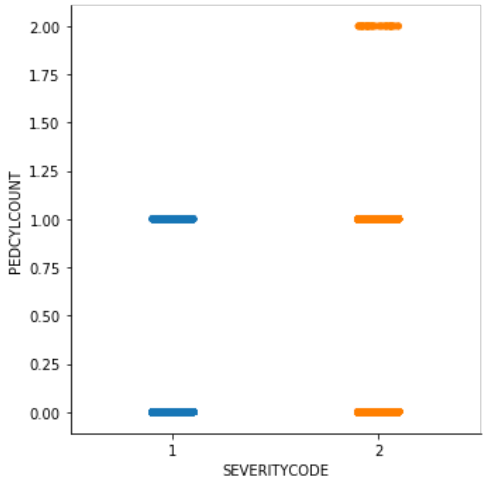

Figure 5 - Cyclist Count vs. Severity

3.1.6 Severity of an Accident vs. Vehicle Count

In order to better understand the relationship between the number of vehicles involved in an accident and the severity of that accident, the vehicle count was plotted against the severity code as shown in the figure below. As shown in the visualization, there is not a very strong correlation between the amount of vehicles involved in an accident and the severity of that accident.

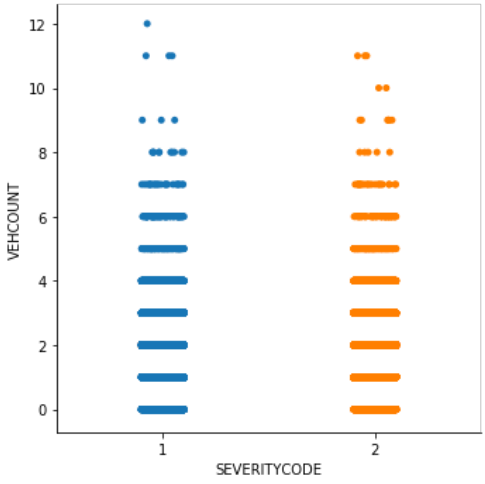

Figure 6 - Vehicle Count vs. Severity

3.1.7 Severity of an Accident where Inattention Occurred

In order to better understand whether or not inattention causes an accident to be more severe, the number of accidents where inattention occurred was plotted against their severity codes as shown below. The plot shows that a majority of these accidents resulted in a severity code of 1. The ratio of severity code 1 accidents to severity code 2 accidents is nearly 2:1.

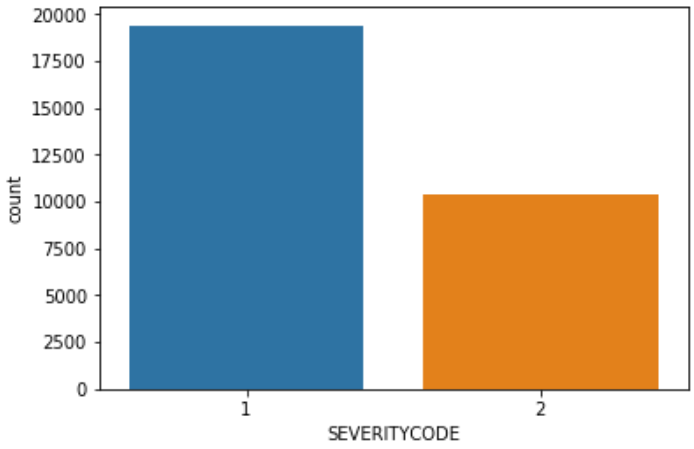

Figure 7 - Accident Count where Inattention Occurred vs. Severity

3.1.8 Severity of an Accident where the Driver was Under the Influence of Drugs/Alcohol

In order to better understand whether or not the driver being under the influence of drugs or alcohol is related to the severity of an accident, the number of accidents in which the driver was under the influence was plotted against the severity codes. As shown below, more accidents of severity code 1 were recorded, but the ratio of severity code 1 accidents to severity code 2 accidents went down, to around 3:2.

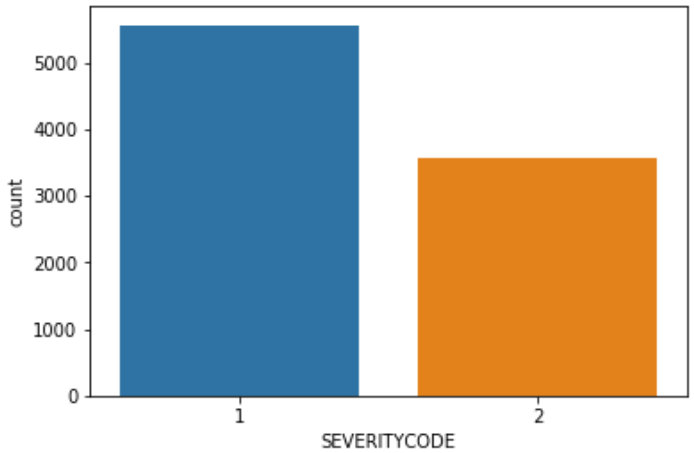

Figure 8 - Accident Count where Driver was Under the Influence vs. Severity

3.1.9 Severity of an Accident where the Driver was Speeding

In order to better understand whether or not a driver speeding is related to the severity of an accident, the number of accidents in which the driver was speeding was plotted against the severity codes. The plot below shows the correlation found. Out of a total of 9333 accidents reported in which the driver was speeding, 3531 of those had a severity code of 2. This equates to a severity code of 2 nearly 38% of the time.

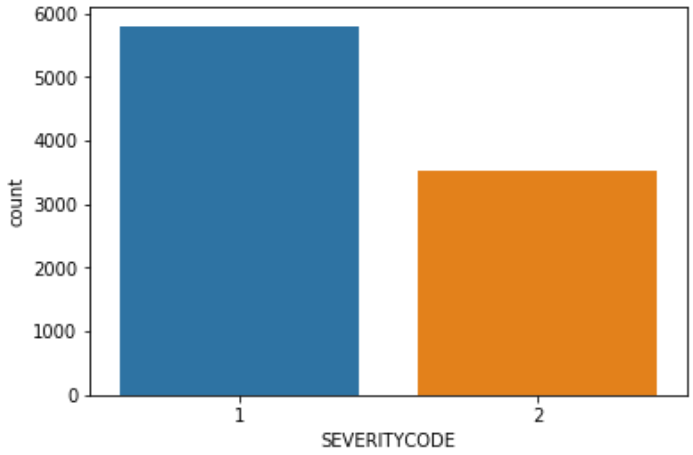

Figure 9 - Accident Count where Driver was Speeding vs. Severity

3.1.10 Severity of an Accident vs. Weather Conditions

In order to better understand the correlation between weather conditions and accident severity, ten plots were created. Each plot examined the number of accidents versus the severity code for each type of reported weather condition. Since the number of accidents for each weather condition varies by quite a bit, each weather condition was examined on its own. As one can see below, each weather condition resulted in more severity code 1 accidents, but certain weather conditions (i.e, raining and fog/smog/smoke) resulted in a higher percentage of severity code 2 accidents.

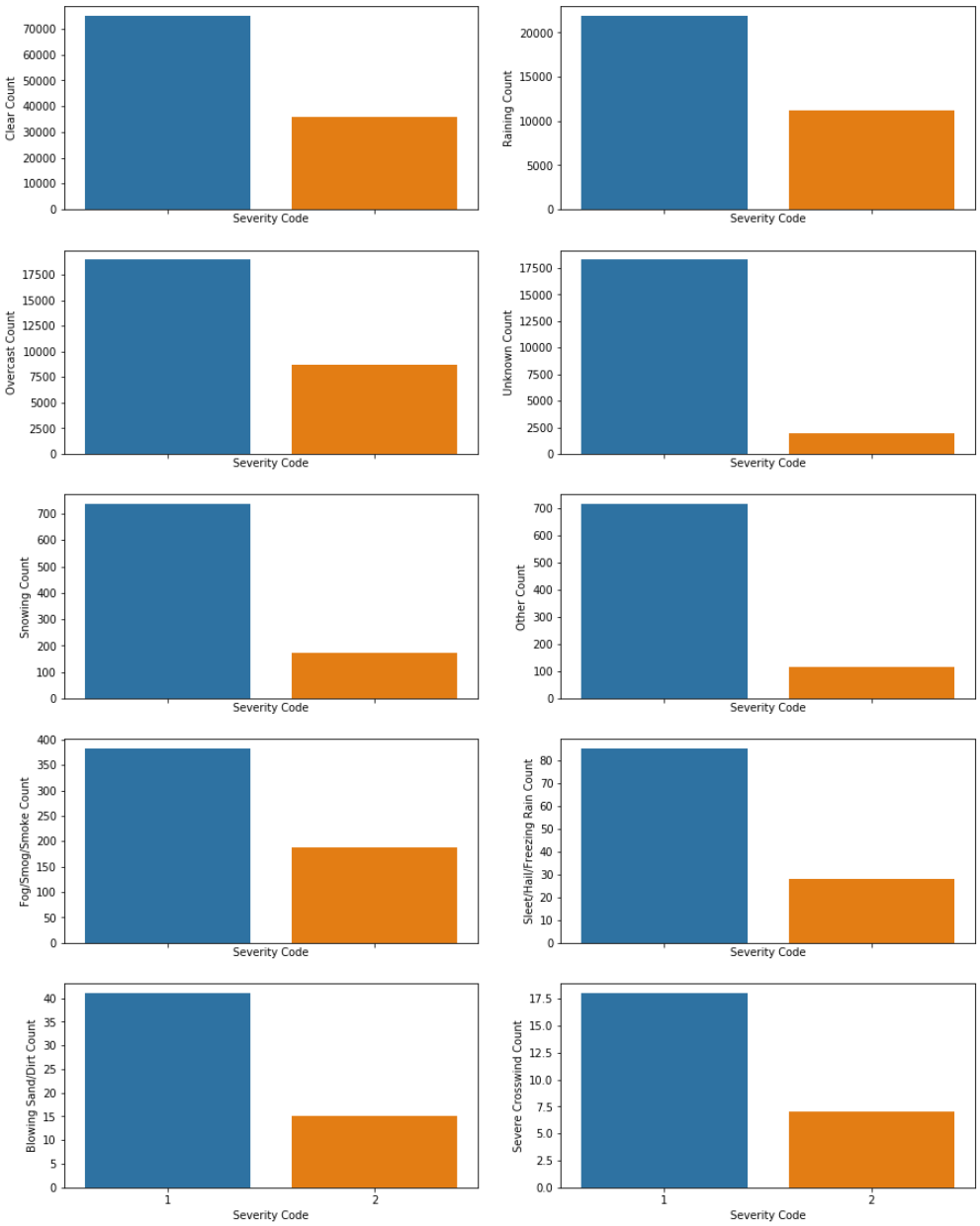

Figure 10 - Accident Count vs. Severity for Various Weather Conditions

3.1.11 Severity of an Accident vs. Road Conditions

In order to better understand the relationship between road conditions and accident severity, nine plots were created for each of the reported road conditions. Each plot examined the number of accidents versus the severity code for each road condition. Each road condition was plotted on its own due to the variance of total numbers of accidents for each road condition. As one can see below, similarly to the weather conditions, each condition resulted in more severity 1 accidents than severity 2 accidents. However, certain road conditions (i.e, oil, other, and wet) resulted in a higher percentage of severity code 2 accidents than others.

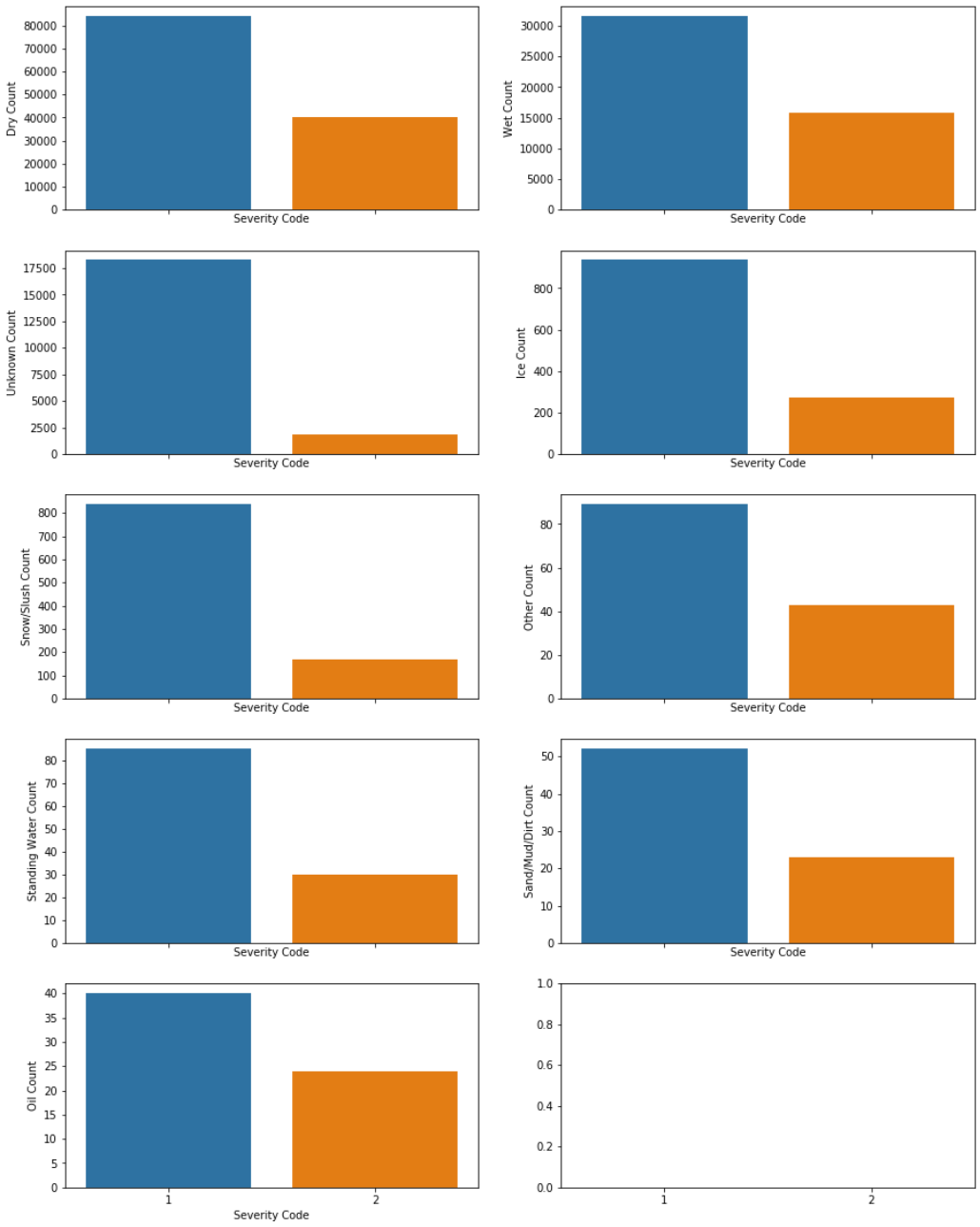

Figure 11 - Accident Count vs. Severity for Various Road Conditions

3.1.12 Severity of an Accident vs. Lighting Conditions

In order to better understand the correlation between lighting conditions and accident severity, eight plots were created for each of the reported lighting conditions. Each plot examined the number of accidents versus the severity code for each lighting condition. Each lighting condition was plotted on its own due to the variance of total numbers of accidents for each lighting condition. As one can see below, each condition resulted in more severity 1 accidents than severity 2 accidents. However, certain lighting conditions (i.e, dawn and dusk) resulted in a higher percentage of severity code 2 accidents than others. However, many factors can play into these lighting conditions. For example, it is possible that dawn and dusk result in more severity code 2 accidents due to more people rushing to get to and from work during heavy traffic hours.

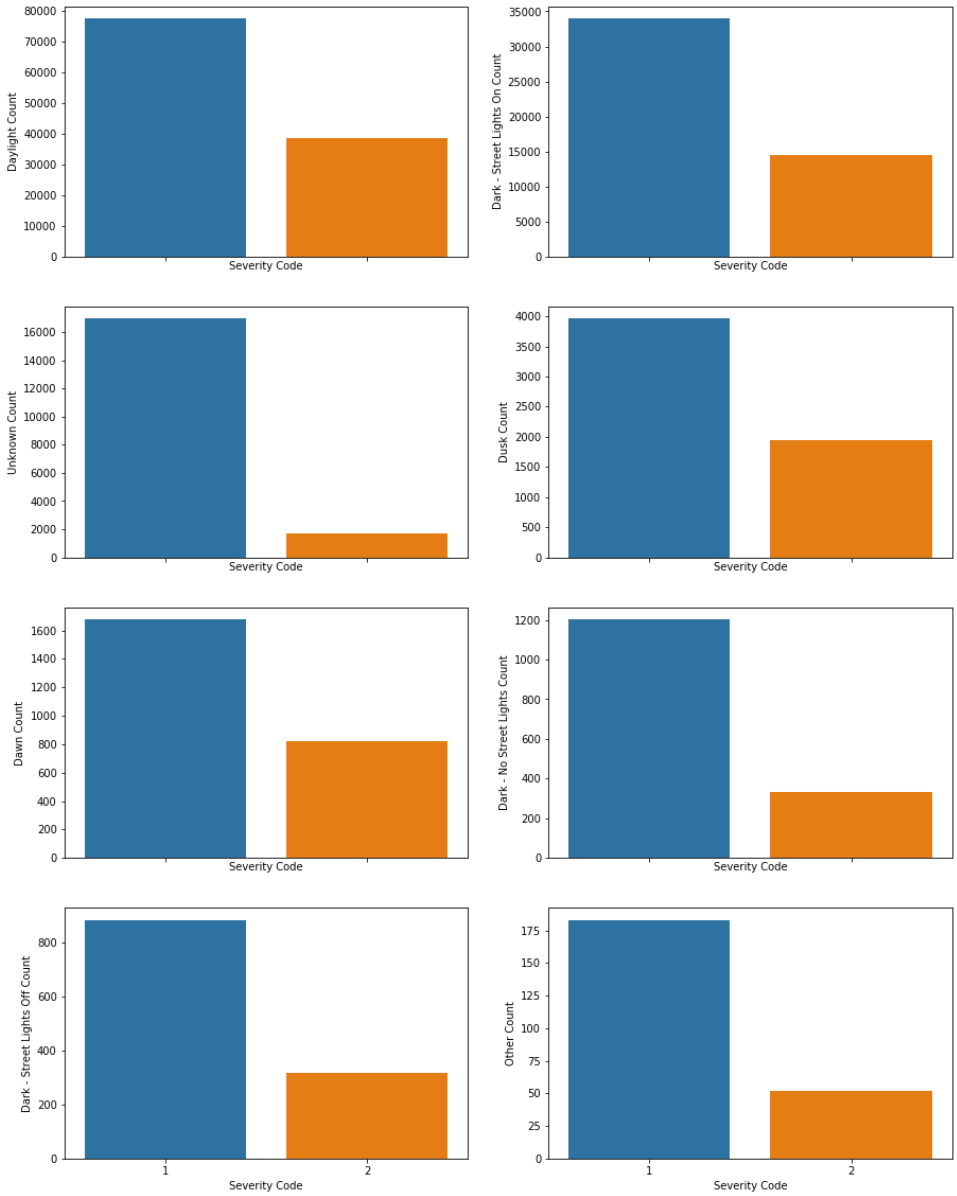

Figure 12 - Accident Count vs. Severity for Various Lighting Conditions

3.2 Predictive Modeling

The main goal of this project is to create a model which can accurately predict the severity of a future accident given a set of various features. There were two different variables to describe the severity of past accidents: 1 and 2. A severity code of 1 means just property damage resulted from the accident. A severity code of 2 means that an accident occurred from the accident. The data was split into a training data set and a testing data set to be used in the various algorithms and models. In order to find the most accurate model to complete this task, three separate models were built. Various accuracy scores were found for each of the three models, and the most accurate model was selected for implementation.

3.2.1 K-Nearest Neighbors Classification Model

The first model that was created was the 'K-Nearest Neighbors' classification model. In these type of models, it is important to first find the optimal value for the hyper parameter 'k'. In the interest of saving time, a small fraction of the total data was split into training and testing data to be used to find the optimal 'k' value. Once the optimal value for 'k' was found, the model was fitted using that value and the original training data.

3.2.2 Decision Tree Classification Model

The second model that was created was the 'Decision Tree' classification model. In these type of models, it is important to first find the optimal value for the hyper parameter 'max_depth'. To find this value, the training data was further split into testing and training data sets. The Decision Tree model was then repeatedly trained and tested with these new data sets and various 'max_depth' values. Each time the model was trained and tested with a different 'max_depth' value, the mean accuracy was found. The 'max_depth' value which resulted in the highest mean accuracy value was selected as the optimal 'max_depth' value and used in the final model. The final Decision Tree model was then fitted with the original training data and the optimal 'max_depth' value.

3.2.3 Logistic Regression Classification Model

The third model that was created was the 'Logistic Regression' classification model. In these type of models, it is important to first find the best 'solver' to use for the model. Similarly to the 'K-Nearest Neighbors' model, the training data was further split into testing and training data sets in the interest of saving time. The Linear Regression model was then fitted with each one of the five solvers and the new, smaller training data set. Each of these models was then ran with the new, smaller testing data set and the mean accuracy for each was found. Four of the five solvers produced a mean accuracy of 0.72 (newton-cg, lbfgs, sag, and saga). The 'liblinear' solver produced a mean accuracy of 0.715, so that was ruled out. After some additional research, it was found that the 'saga' solver is very time efficient for large datasets. Thus, the 'saga' solver was used along with the original training and testing data sets for the final Linear Regression model.

4. Results

The three fitted models were used to predict severity codes using the testing data set. The various accuracy scores of the models' predictions were found using four different evaluation metrics. Confusion matrices were then formed for each of the three models used.

4.1 Accuracy Scores

	Jaccard Score	F1 Score	Subset Accuracy Score	Log Loss
K-Nearest Neighbors	0.752235	0.844439	0.752235	N/A
Decision Tree	0.757526	0.850335	0.757526	N/A
Logistic Regression	0.755420	0.848211	0.755420	0.514902

Table 2 - Accuracy Scores for Models

4.2 Confusion Matrices

Confusion matrices are very useful for visualizing how many of the severity codes were correctly and incorrectly predicted. The confusion matrices for the three models are shown below.

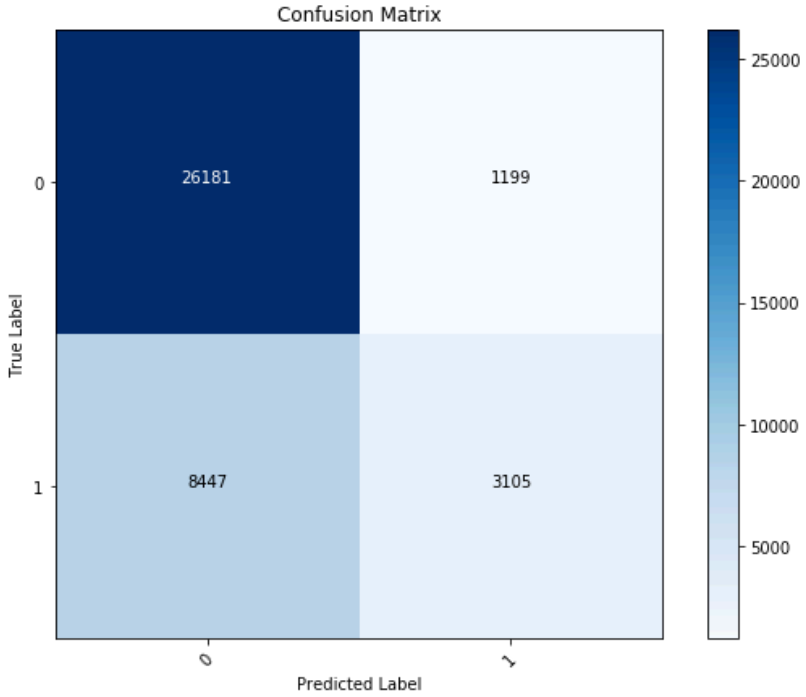

Figure 13 - K-Nearest Neighbors Confusion Matrix

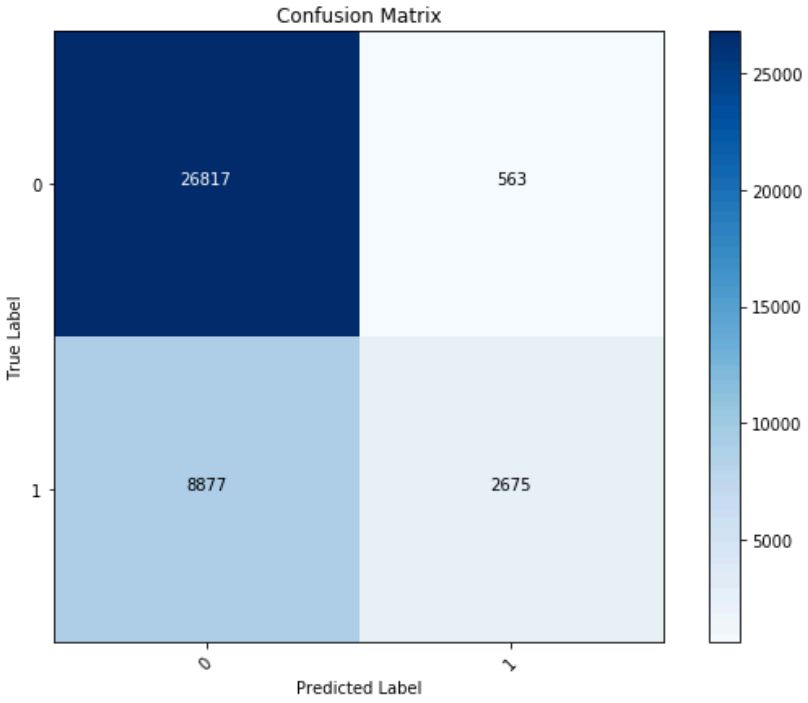

Figure 14 - Decision Tree Confusion Matrix

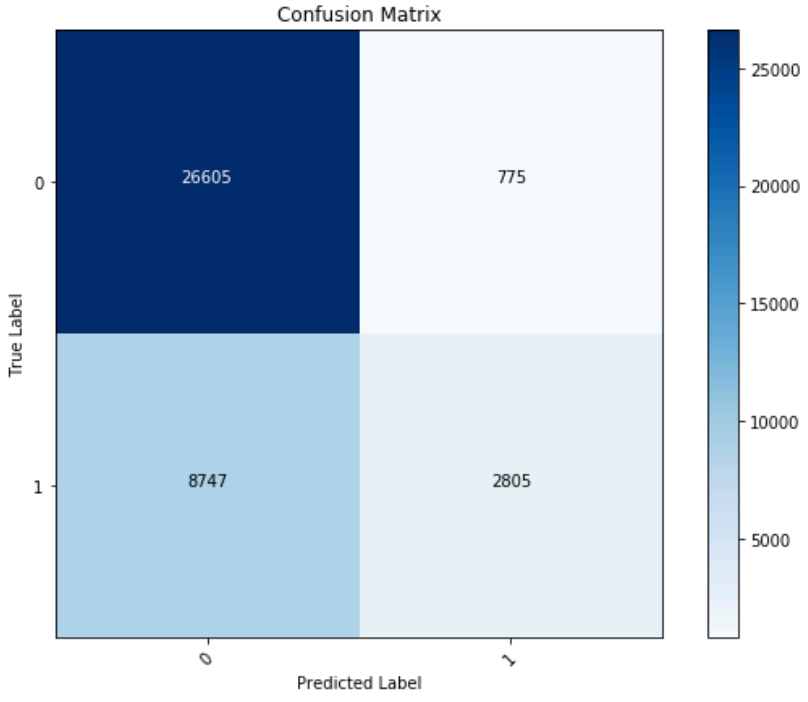

Figure 15 - Logistic Regression Confusion Matrix

5. Discussion

As one can see from Table 2, the Decision Tree was the most accurate model and therefore should be implemented. Though the accuracy scores were all very close, the Decision Tree model produced the highest Jaccard score (~0.758), F1 score (~0.850), and Subset Accuracy score (~0.758). Even though these scores are fairly accurate, there was still some variance that the models could not avoid based on the dataset given. Several relevant features contained data marked as 'Unknown' and also contained some missing data. More complete and robust data could definitely help improve these accuracy scores.

The confusion matrices are a good way to visualize both how many of the accidents were correctly predicted and how many false positives and false negatives resulted from each model. Between the three models, the Decision Tree model correctly predicted the most accidents which resulted in a severity score of zero (26,817), but correctly predicted the least amount of accidents which resulted in a severity score of one (2,675). In addition, the Decision Tree model incorrectly predicted the highest amount of severity code zero accidents (false negatives), but also incorrectly predicted the least amount of severity code one accidents (false positives). The Decision Tree model's overwhelming amount of false negatives is a bit worrisome in terms of real world applications, but once again, with a more robust dataset, the number of false negatives would surely go down.

6. Conclusion

The task of this project was to analyze a dataset of car accident data created by the city of Seattle, Washington and to produce a model that could accurately predict the severity of a car accident given various factors and conditions. The data was properly cleaned and the irrelevant data was dropped before exploratory data analysis was conducted. The initial data analysis helped to visualize the different features and show their importance to the severity of an accident. Three classification models were then created and their accuracies were found. The model with the highest accuracy scores, the Decision Tree model, is suggested to be implemented. With a more robust dataset, the accuracy of this model will only get better. Going forward, this model will be of high importance to several parties. This model would be of great use to the government to help increase safety measures where accidents may be more severe and also to help station emergency personnel in areas where these features present a higher chance of being a severe accident. Insurance companies would be able to use this model to properly adjust premiums for drivers who spend more or less time driving in areas which could have a higher chance of a severe accident. Lastly, automakers and the people driving vehicles themselves would have great use for this model. The model would be able to warn drivers when to pay more attention in situations where a severe accident is more likely to occur. In conclusion, this project was a success in the fact that a model was created with relatively high accuracy scores.