Predicting the Car Accident Severity in Seattle

Brent Arnold Apostol Basiano

August 24, 2020

1. Introduction

1.1. Background

Car accidents are a common yet avoidable problem in the United States. The Association for Safe International Road Travel (ASIRT) stated that more than 38,000 people die every year in crashes on US roadways with an economic impact of $871 billion. In Washington alone, car accidents occur every four minutes. Car accidents can be caused by, but not limited to, speeding, driving under influence, weather conditions, and road conditions. However, while these conditions are known to cause car accidents, can these conditions help determine the severity of each car accident? If these factors were correlated to the severity of car accidents, a predictive model can be made to help predict car accident severity.

1.2. Problem

The dataset contains multiple factors that could contribute to the car accident severity in Seattle. The report aims to predict the car accident severity based on past reports.

1.3. Business Interest

Transport organizations and the Seattle Department of Transportation can use these findings to help understand the common factors that influence the severity of car accidents. Law enforcement can also use this report to help them expect how severe a recent car accident is.

2. Data Collection

2.1. Data Source

The dataset is provided by the Seattle Police Department via IBM here; the metadata is also provided here. The dataset contains multiple factors along with the severity code that can help predict the severity of a car accident. While additional datasets can be used, the amount of information in this dataset suffices the task for this problem.

2.2 Data Cleaning

The dataset originally contained 194,673 samples with 38 features. There were multiple values that were missing in the dataset. I decided to drop any rows that were missing values or contained a label that states, “Not enough information.” Unusually, the feature “**SPEEDING**” contained 185,340(95% of the dataset) missing values making this feature not usable. Some features in the dataset were duplicates or described certain codes. For example, two columns labeled “**ADDRTYPE**” and “**JUNCTIONTYPE**” described where the accident took place either in an intersection or a block. I decided to drop the junction type and keep the address type.

Certain features were kept for exploratory purposes and not used for fitting in the machine learning model. The coordinates of the location were kept to see which part of Seattle

have seen a significant amount of car accidents. The coordinates were also color-labelled to show the severity of the car accident.

Additionally, I categorized the incident date based on the season. Keeping the entire incident date would be impractical since there are many samples. Therefore, categorizing them based on their season could be used in the model.

Last, I checked to see if there are any outliers in the data. I found there were outliers in the vehicle count and people count. For example, there would be 12 cars or 80 people in one accident. Fortunately, the dataset provided a state collision code number which can be used instead of the numerical data. The code helps describe the collision in a categorically manner. For example, code 45 stands for a vehicle collision with a bicycle. Therefore, the state collision code was kept while the numerical data were dropped. After cleaning, there were 184,146 samples and 10 features in the dataset for the machine learning model.

Table 1: Dropped Features

Dropped	Reason for Dropped
VEHCOUNT, PEDCOUNT, PEDCYLCOUNT, PEOPLECOUNT	ST_COLCODE describes the collision categorically. Prevents any outliers from numerical data.
OBJECTID, INCKEY, COLDETKEY, REPORTNO, STATUS, INTKEY, EXCEPTRSNCODE, EXCEPTRSNDESC	These features were used for record keeping.
SPEEDING	95% are missing values
SEGLANEKEY, CROSSWALKKEY	Not necessary
PEDROWNOTGRNT, SDOTCOLNUM, INATTENTIONIND	Described by ST_COLCODE
JUNCTIONTYPE, SEVERITYCODE.1, SEVVERITYDESC, ST_COLDESC, INCDTTM	Duplicates
LOCATION	Feature “COORDINATES” was used for data analysis

3. Exploratory Data Analysis

3.1 Target Variable

Since the purpose of this project was to predict the severity of car accident. The target variable is the severity code. The severity code contains 2 numbers: Code 1 is property damage and code 2 is that an injury was present. I first plotted the coordinates of the first 600 samples to see where they are occurring in Seattle. Next, I check the relationship between the target variables and each feature. When the relationship is insignificant, I dropped them from the dataset. There was also an imbalanced of data because there were more code 1 accidents than code 2 accident. This problem will be discussed during the model development.

3.2 Accident Location

To see where the accident took place, I used Folium to plot the first 600 samples. Looking the map, the samples are heavily plotted north of the International District. Code 1 is **Yellow** while code 2 is **Red**.

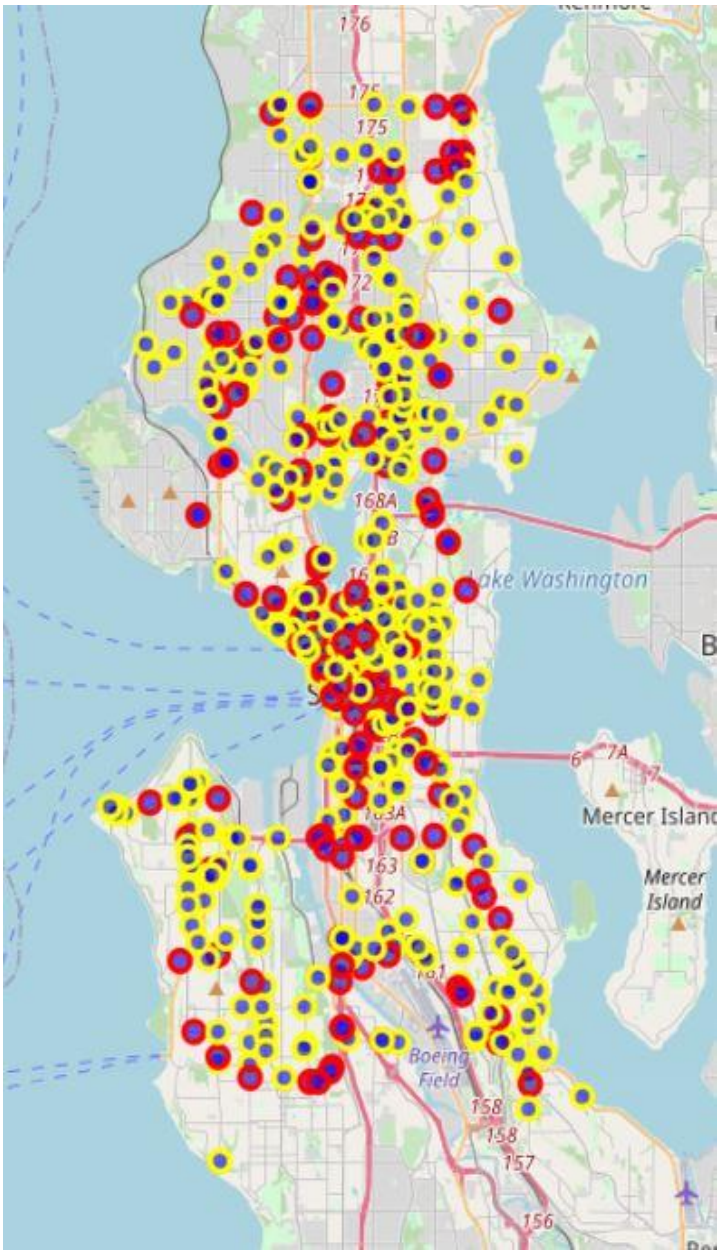

Figure 1: Map of Seattle with accidents plotted.

3.3 Relationship between Severity and Address Type

Accidents can occur either in a street block or intersection. It is more likely that there would be more block accidents than in intersections since a block or a grid covers more road than an intersection. Looking at Figure 2, there were more code 1 accidents in blocks than in intersections. Code 2 was almost equal in both types.

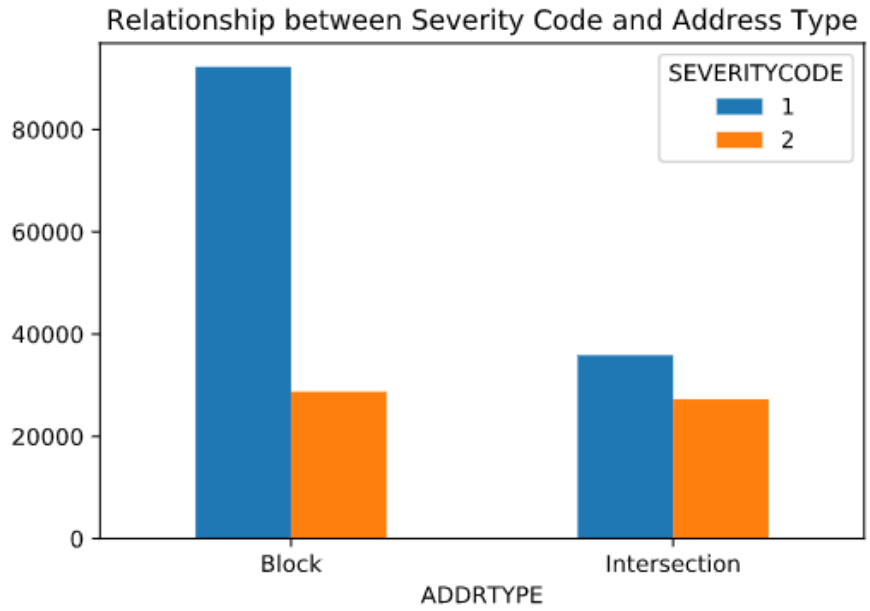

Figure 2. Address Type

3.4 Relationship between Severity Code and Driving Under Influence

The expectation is that driving under the influence of alcohol or drugs will most likely result in a car accident; however, the dataset shows otherwise. Looking at Figure 3, there were more accidents where the driver is not under the influence of alcohol nor drugs.

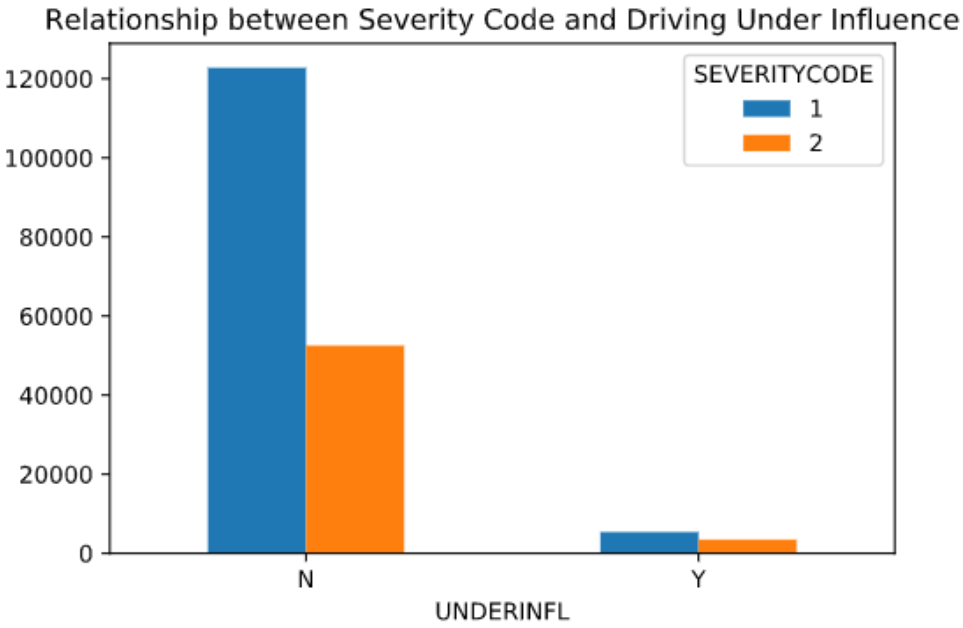

Figure 3. Under the influence of alcohol or drugs

To understand why there were more accidents while the driver is not under the influence of alcohol or drugs, I check the relationship between the light condition and driving under influence. It is expected that drivers tend to not drink during the day; the dataset shows that many accidents occurred during the day.

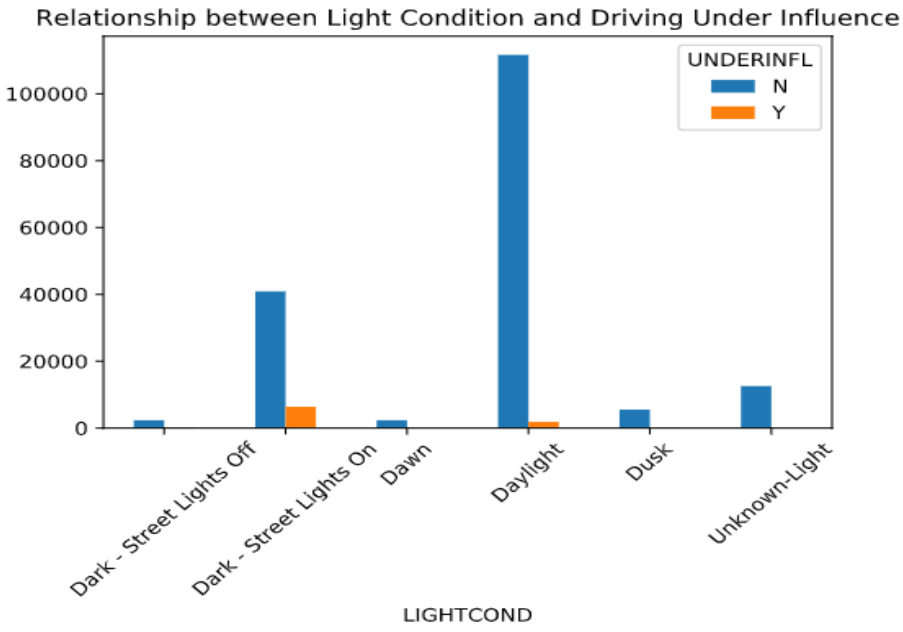

Figure 4. Light Condition and Driving Under Influence

3.5 Relationship between Severity Code and Light Condition

Since most people drive during the day, it is expected that there would be more accidents during the day. Looking at the map, accidents tend to concentrate north of the International District; therefore, during the night, recorded accidents will most likely occur where there are streetlights.

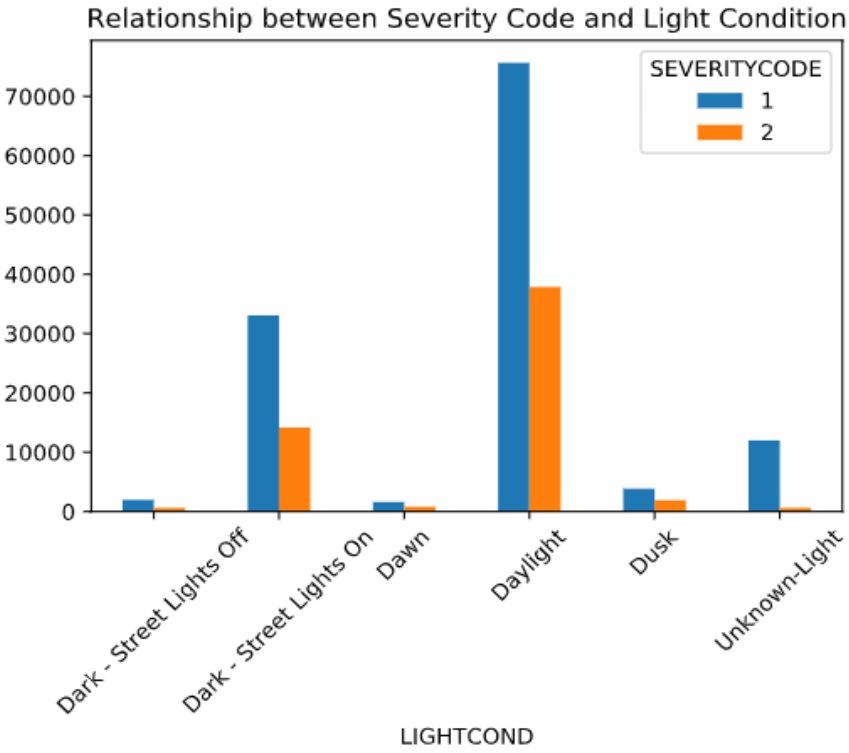

Figure 5. Light Condition and Severity Code

3.6 Relationship between Severity Code and Weather

It is expected that there would be more accidents in clear skies since drivers tend to be more cautious when driving in the rain. Based on Figure 6, more accidents are recorded while the weather was clear. Both severity codes were prevalent when the weather was clear with raining as the second most prevalent weather for car accidents.

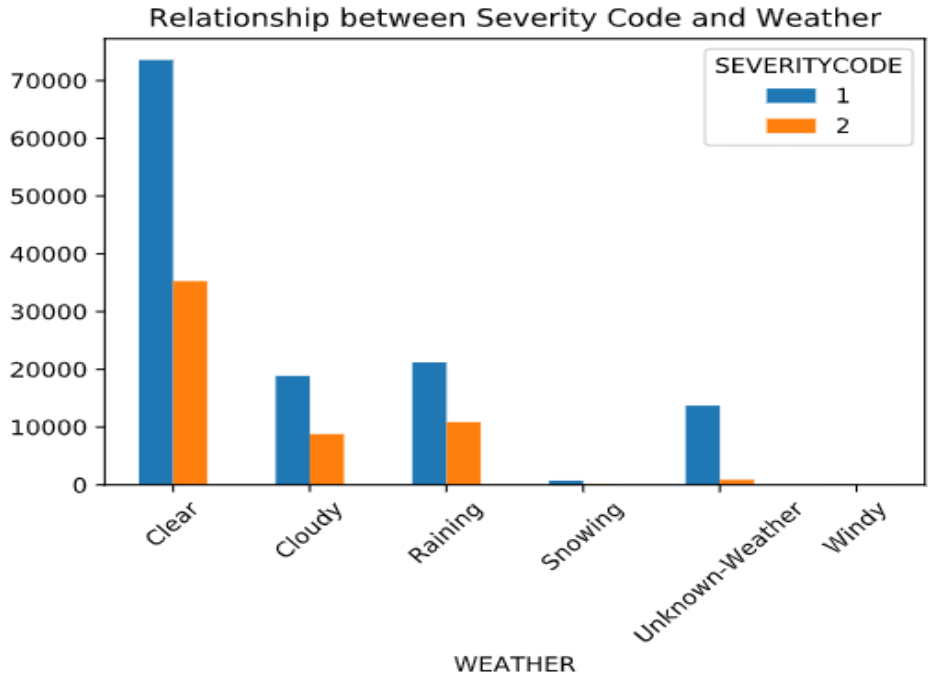

Figure 6. Weather Observations

3.7 Relationship between Severity Code and Road Condition

Since the weather shows that there are more accidents when the weather is clear, it is expected that there are more accidents while the road conditions are dry. Looking at Figure 7, there were more accidents when the road conditions were dry while wet roads was the second most prevalent. Dirt and oil had barely any car accidents.

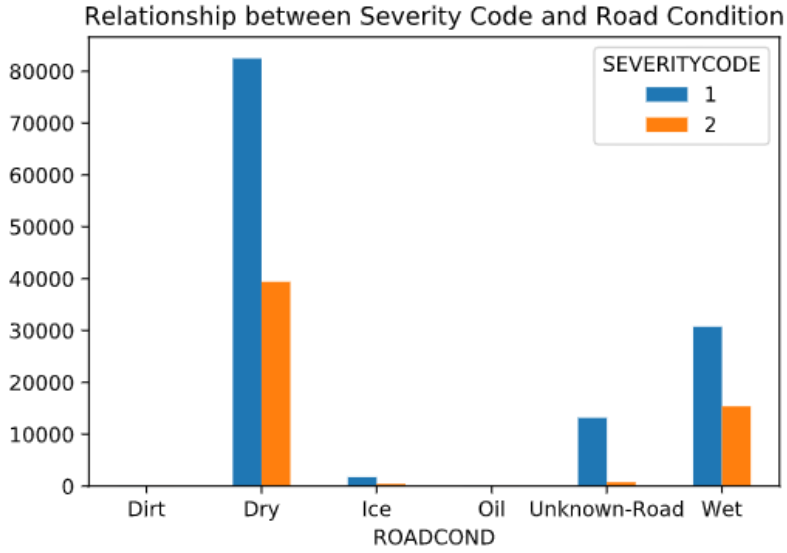

Figure 7. Road Conditions

3.8 Relationship between Severity Code and Hitting a Parked Car

Car accidents are expected to occur between two moving vehicles or property damage. There would be more accidents without a parked car. If there were a parked car, the severity code would most likely by Code 1 since it is an accident involving property damage.

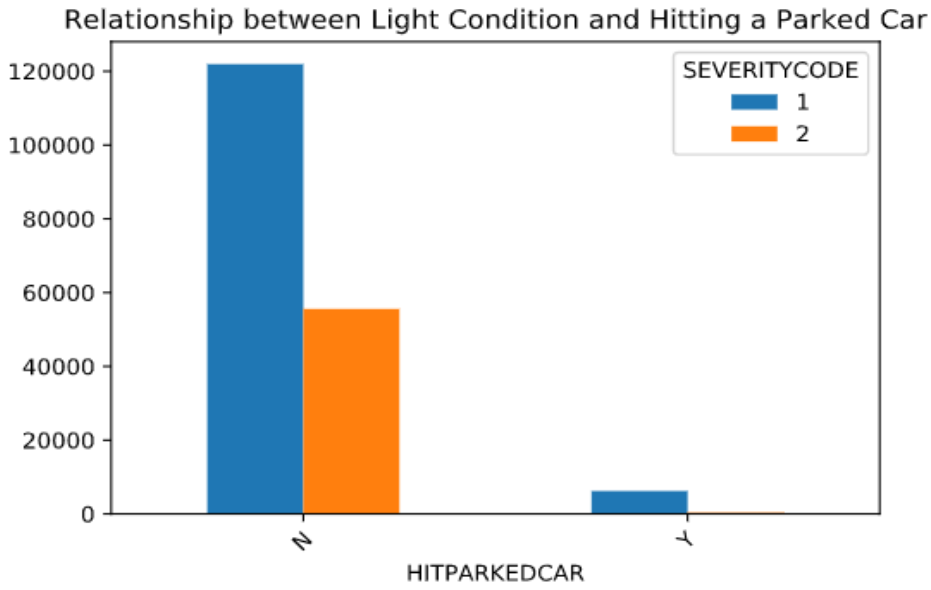

Figure 8. Parked Car Severity Code

3.9 Relationship between Severity Code and Season

Previously, the incident date was recategorized to incident season because going through each incident date is impractical. However, the expectation is that season might not have any correlation with severity code. Based on figure 8, the seasons showed that the amount of car accidents is almost similar to each other; therefore, the season will be dropped during the model evaluation.

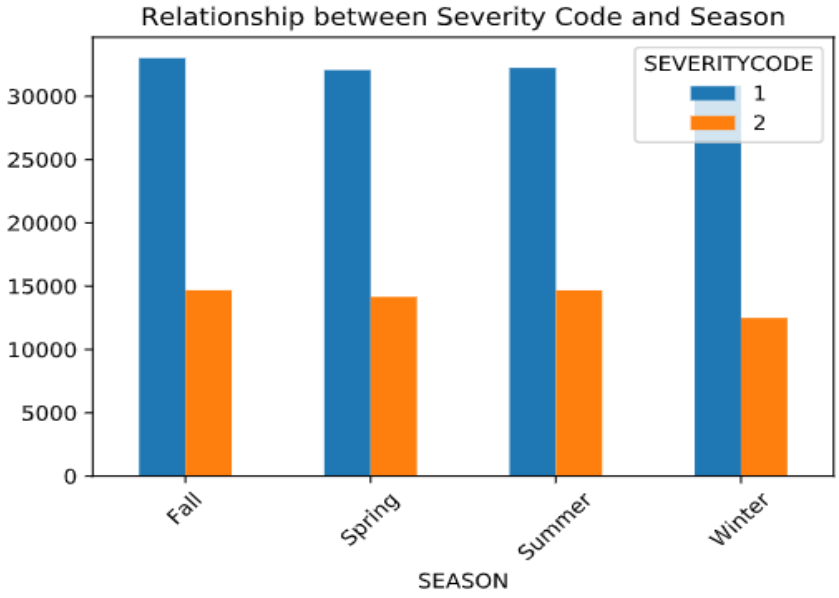

Figure 9. Severity Code and Seasons

3.10 State Collision Number and SDOT Collision Number

The state collision number is the collision code that describes a collision categorically between the numbers 0 to 84. While there are collision codes that are rarely used, we cannot ignore them since the code describes the collision; therefore, they cannot be labeled as outliers. Looking at Figure 10 and 11, code 32 is the most prominent and code 10 is the second most prominent. Code 32 and 10 states that a bicycle was damaged, and a car entered at an angle, respectively.

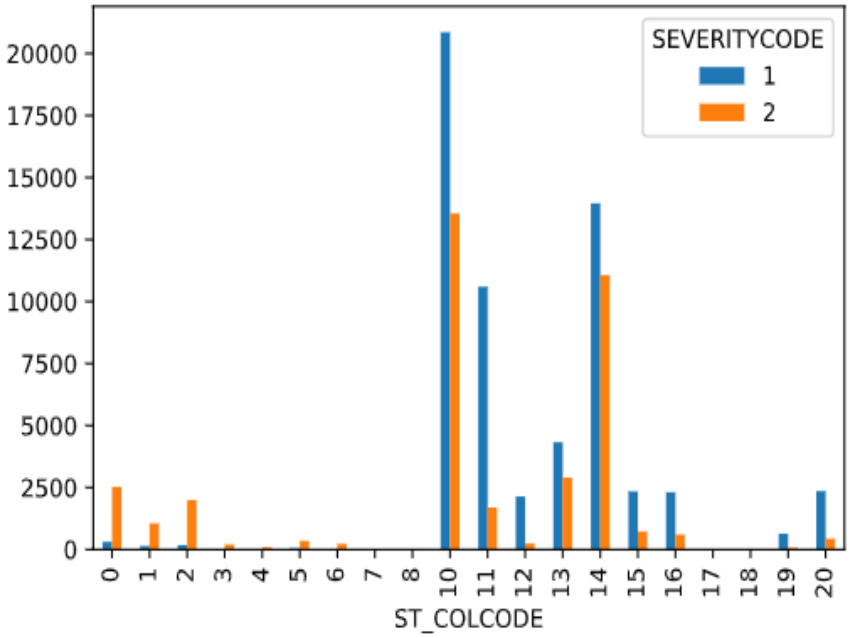

Figure 10. Code 0 to 20

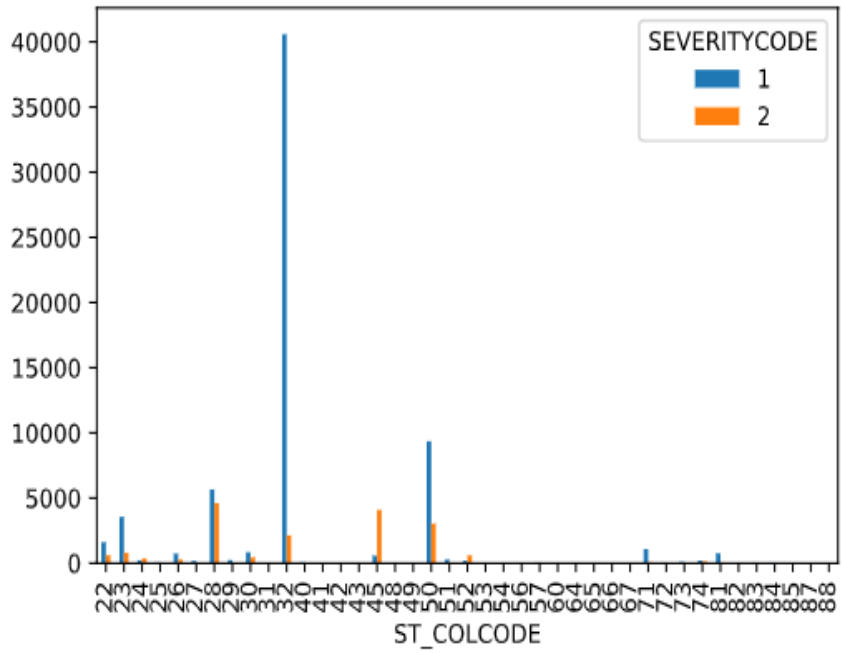

Figure 11 Code 22 to 84

The SDOT collision number describes the direction of the impact between a vehicle and a pedestrian (both walkers and cyclist). For example, if the code is 11, then the direction of travel is north before and after the collision. Looking at Figure 13, code 51 means a cyclist was hit head on which is a severity code 2 incident.

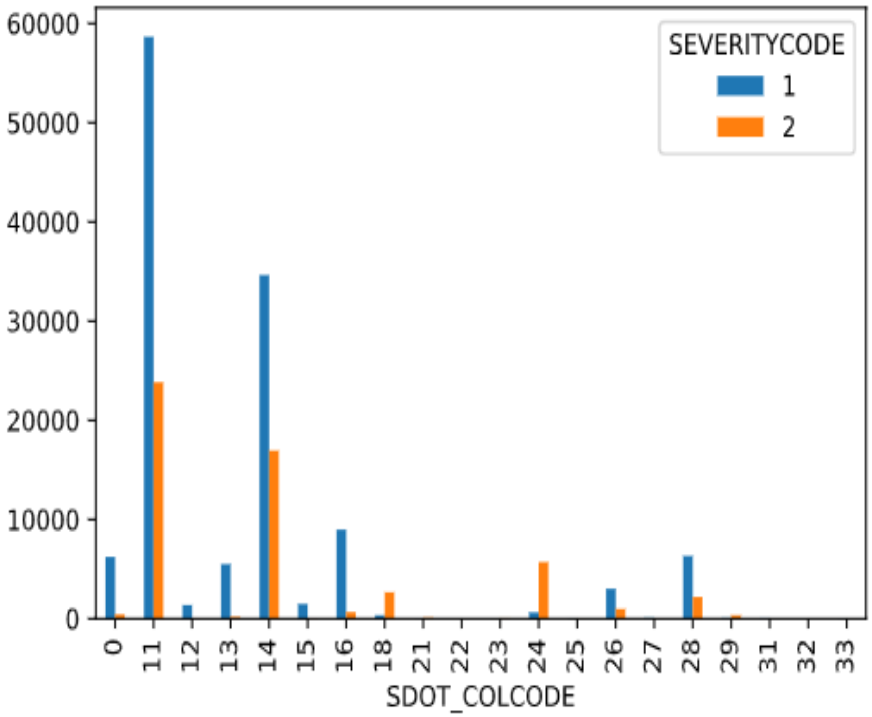

Figure 12. Code 11 to 33

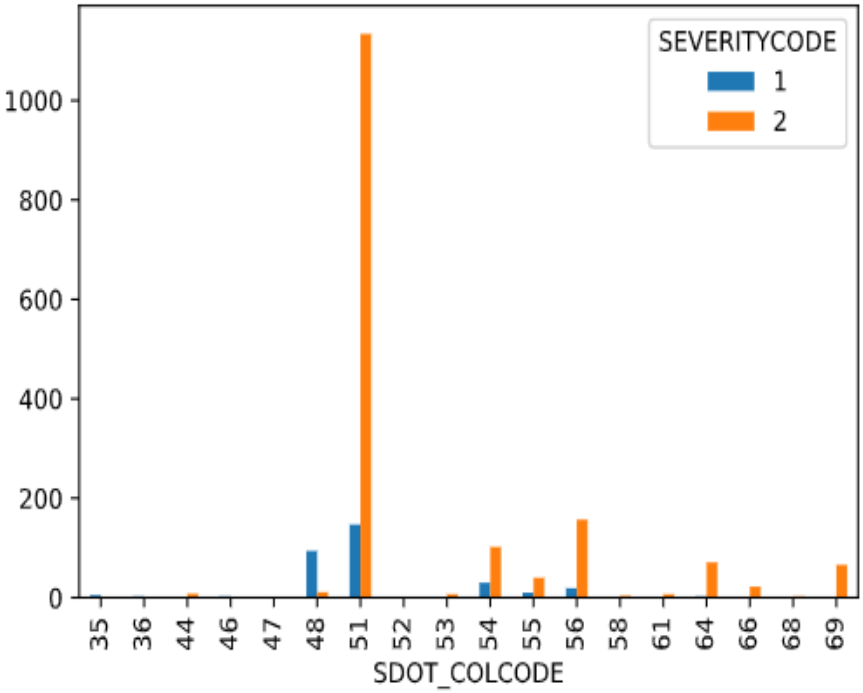

Figure 13. Code 35 to 69

4. Predictive Modeling

4.1 Feature Set

After analyzing the data, I decide to drop the seasonal feature because it did not show any correlation between severity code and season. The dataset now consists of 184,146 samples and 8 features with the severity code as the target variable.

Table 2 Feature Table

ADDRTYPE	SDOT_COLCODE	UNDERINFL	WEATHER
ROADCOND	LIGHTCOND	ST_COLCODE	HITPARKEDCAR

4.2 Imbalance Dataset and Machine Learning Model

Looking at the data analysis, the dataset shows evidence that there are more code 1 samples than code 2 samples. This dataset is known as an imbalance dataset. If a machine learning model were to be fitted by this dataset, then the model will almost always predict code 1. This problem can cause is evident of overfitting with another dataset is being used for evaluation.

To solve this problem, there are three common methods to use: downsample the majority class, upsample the minority class, and use a tree classifier. Downsampling the majority class reduces the size of the majority class to match the size of the minority class. Upsampling the minority class increases the size of the minority class to match the size of the majority class. Using a tree classifier considers all the weights of each feature to help accurately determine the target class.

For this project, I will compare the performance of a random tree classifier and logistic regression using a downsampled dataset. A random tree classifier is a bundle of decision tree classifiers that uses averaging to improve its accuracy and prevent overfitting. A logistic regression is a model that determines the probability of two classes like pass/fail or win/lose.

4.3 Feature Transformation for Model Fitting

Before fitting the dataset to the model, I needed to transform some variables into numerical values since the model cannot fit strings or text. For SDOT and state collision codes, I did not change their data type since it is already numerical. For the other features, I used one-hot encoding. One-hot encoding uses a binary variable for each unique integer variable. For example if a variable is green out of the following list: [red, green, blue], one-hot encoding will transform the variable into the following list; [0,1,0].

4.4 Model Development

To help ease the development of each model, I used sklearn's pipeline feature. Sklearn's pipeline sequentially applies a list of transformations which makes the development more organized. The final step is the fitting the final estimator which is either the random tree classifier or the logistic regressor to the training dataset.

4.5 Model Performance Results

As the metrics, I used the accuracy score, recall score, precision, F1-score, and support. The accuracy score is the accuracy between the true target class compared to the respective predicted class. The recall score is the performance score of finding the positive sample. The precision score is the performance score of not labeling a true negative sample a positive sample. The F1-score is the harmonic mean between precision and recall where 1 is the best. Last, support is the number of occurrences of each class in true class variable. The random tree classifier had a higher accurate score than logistic regression with a downsampled dataset.

Table 3. Cross Validation and Accuracy Score

Metrics	Random Tree Classifier	Logistic Regression
Cross-Validation Score	0.707	0.653
Accuracy Score	0.711	0.649

Table 4. Random Tree Classifier Classification Report

Metrics	Code 1	Code 2
Precision	0.74	0.77
Recall	0.97	0.23
F1-Score	0.84	0.35
Support	25701	11129

Table 5. Logistic Regression Classification Report

Metrics	Code 1	Code 2
Precision	0.66	0.64
Recall	0.61	0.69
F1-Score	0.63	0.66
Support	11221	11182

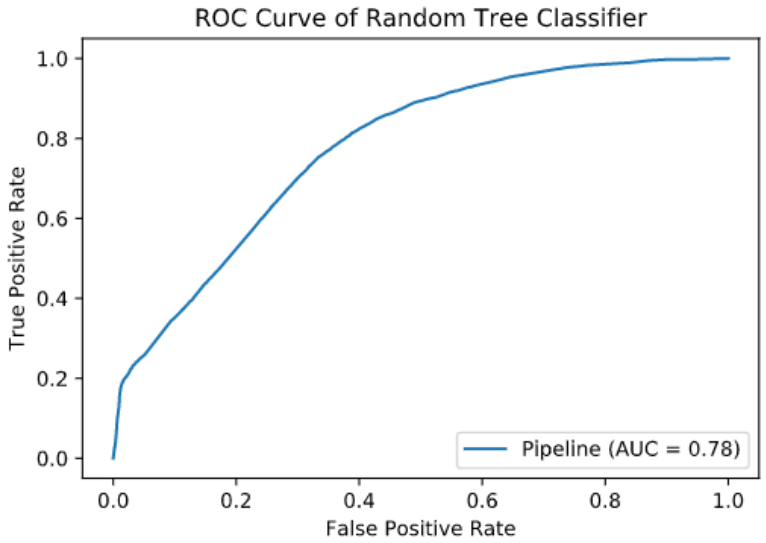

Figure 14. ROC Curve of Random Tree Classifier

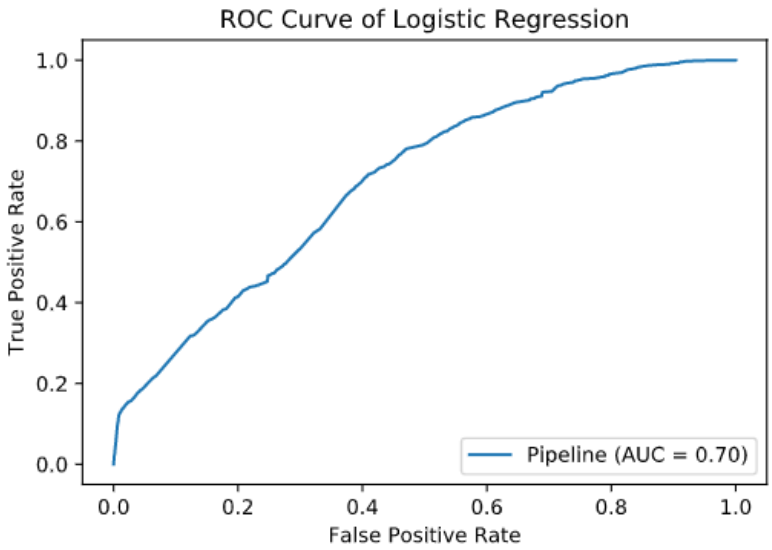

Figure 15. ROC Curve of Logistic Regression

4.6 Result Discussion

The random tree classifier contains interesting results, the low recall score for severity code 2 shows with an imbalanced data set, the tree was not able to label code 2 accurately but shows that the tree classifier can predict code 1 more accurately than code 2. Looking at the two ROC curves, the AUC for the random tree classifier is higher than the logistic regression meaning the random tree classifier is better in predicting the class variables than the logistic regression. This result can also be interpreted as the measure of separability for the random tree is better than the logistic regression. Fortunately, since the AUC in both models is above 0.5, the models have good separability.

The logistic regression model had a better classification report than the random tree classifier. Based on the classification report for code 2, the logistic regression can more accurately predict class 2 compared to the random tree classifier. This finding could be evidence that the random tree classifier has an overfitting problem, but more evaluation with a different dataset is required to confirm that finding.

5. Conclusion

In this project, I analyzed the relationship between the severity code and the features given in the dataset. I used address type, SDOT collision number, state collision number, alcohol or drug usage, weather, road condition, light condition, and parked car collision to determine the severity code of a car accident. I built a random tree classifier using an imbalanced dataset and a logistic regression model using a downsampled dataset. While the random tree classifier had a higher accuracy score than the logistic regression model, both models worked well. More improvements can be made by adding in more features that can determine the severity code of a car accident. For example, speeding is a common cause of car accidents which can be an influential feature for predicting severity code. Overall, the analysis of data and model evaluation showed that the predicting car accident severity is possible for real world use.