Analysis of Car Accident Severity

1. Introduction

1.1 Background

It is no surprise that one of the leading causes of death in the United States is accidents, including those caused by motor vehicles. In fact, there are, on average, 6 million car accidents in the United States every year, of which 3 million people are injured. Additionally, more than 90 people are estimated to die in car accidents every day. While the largest portion of car accidents are caused by driver error, weather and road conditions still contribute a reasonable amount to these numbers. These conditions include environmental factors such as weather, road condition, time of day, and location.

1.2 Problem

Knowing which factors and how much they affect car accidents will provide important information that can be used to caution drivers during certain road conditions through road signs or warnings. In this project, we will focus on seeing which environmental and road factors affect car accidents the most and building a model that best helps predict the severity of a car accident under certain conditions.

1.3 Audience

Government officials or organizations in charge of road traffic and regulation should use this information to increase the safety of everyone on the road. This is also informational for the general public as well.

2. Data Acquisition and Cleaning

2.1 Data Source

The data we will be using is the Seattle city data provided by SPD and recorded by Traffic Records. It includes all collisions that occurred at the intersection or mid-block of a segment from years 2004 to 2020.

2.2 Data Cleaning

The original data consists of 37 attributes and 221,006 entries. However, we only need to keep the attributes we are interested in for this analysis. Such attributes are described in the next section. After subsetting the data to only include the variables of interest, we drop all entries containing NaN and unknown values. The resultant table consists of 5 attributes and 174,570 entries.

2.3 Feature Selection

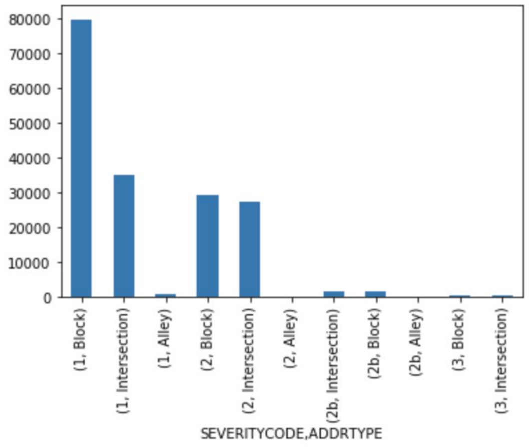
Since we are trying to see how environmental factors affect the severity of car accidents, we will primarily look at variables related to weather, road conditions, and location. These variables are ADDRTYPE, WEATHER, ROADCOND, and LIGHTCOND. ADDRTYPE describes the collision address type. WEATHER describes the weather condition during the time of the collision. ROADCOND describes the condition of the road during the collision. LIGHTCOND describes the light conditions during the collision. We will use these variables to develop a model that determines the severity of the accident (SEVERITYCODE). This variable is a code that corresponds to the severity of the collision (0-unknown, 1-prop damage, 2-injury, 2b-serious injury, 3-fatality).

3. Methodology

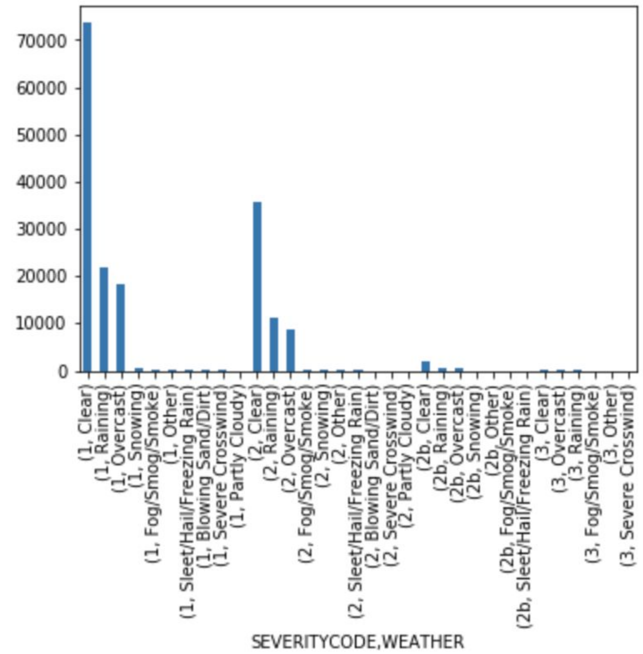
3.1 Exploratory Data Analysis

The following plots show each attribute and its frequency based on the severity code. They are to help visualize the most common road or weather conditions for each type of accident.

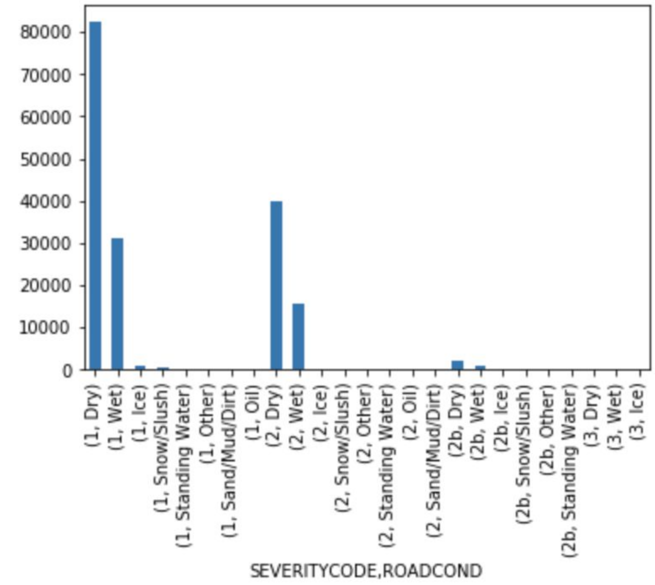
3.1.1 Bar Plot of ADDRTYPE grouped by SEVERITYCODE



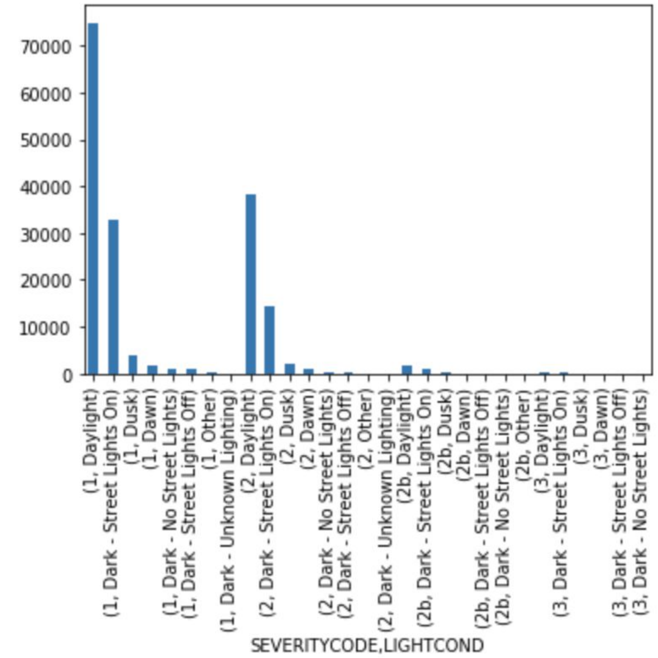
3.1.2 Bar Plot of WEATHER grouped by SEVERITYCODE



3.1.3 Bar Plot of ROADCOND grouped by SEVERITYCODE



3.1.4 Bar Plot of LIGHTCOND grouped by SEVERITYCODE



3.2 Modeling and Evaluation

Since our variables in interest are all categorical variables, we need to label encode them and assign them numerical labels. This way we are able to properly fit models using our data. Then we split our dataset into training and test sets. For these models, we randomly split the data into 70% training and 30% test. Since we want to use machine learning models that can be used to predict a certain class or group based on given conditions, we fit models using SVM (Support Vector Machines), K-Nearest Neighbors, Logistic Regression, and Decision Trees. After fitting each model, we calculate its accuracy using various methods such as f1-score, jaccard similarity score, and classification report.

Classification Model	Accuracy	F1-Score	Precision	Recall	Log Loss
SVM	Train set: 0.6597271663434234 Test set: 0.6572912489736686	0.5213709856308176	0.43203178597736525	0.6572912489736686	
K-Nearest Neighbors (k=7)	Train set: 0.6358235337441387 Test set: 0.6317236638597697	0.5395379811233175	0.5286711151648776	0.6317236638597697	
Logistic Regression	Train set: 0.6597271663434234 Test set: 0.6572912489736686	0.5213709856308176	0.43203178597736525	0.6572912489736686	0.7087700067484126
Decision Tree	Train set: 0.6597598998355142 Test set: 0.6572912489736686	0.5213709856308176	0.43203178597736525	0.6572912489736686	

4. Results

Through our exploratory analysis we see that the majority of accidents from this dataset resulted in prop damage. In all accident groups, it appears that the accident occurred the most during daylight, when the weather was clear, and road condition was dry. The second most likely conditions that cause accidents are rainy weather, wet road conditions, and being in the dark with street lights on. Most of the accidents that resulted in prop damage occurred at a block, while accidents that resulted in injury occurred almost equally at blocks and intersections. It seems that accidents that resulted in fatality only occurred at blocks and intersections.

The four models we built are all very similar in terms of prediction and accuracy. The prediction accuracy is around 63-65%. The f1-score is around 52-54%. The jaccard similarity score is around 63-65%. The log loss for the logistic regression model is around 70.8%. It is pretty difficult to choose the most accurate model since the test set accuracy is the same for the SVM, Logistic Regression, and Decision Tree models. However, the Decision Tree train set accuracy is slightly higher so we can say that this model is more accurate.

5. Discussion

Overall it seems that the model accuracy for all models can be greatly improved. The highest prediction accuracy is only around 65.73%. The classification reports for all the models (except K-Nearest Neighbors) also show that the model could only predict accidents with prop damage. This could likely be due to eliminating many variables that are more significant predictors or having significantly more data for prop damage accidents than other accidents. We only kept the "environmental" variables so our model had very limited data and predictors.

In terms of recommendations, based on our observations and analysis, we should pay more attention on the road when faced with conditions that do not seem to bring much risk. Road signs and warnings should be put up to caution drivers and pedestrians, especially at blocks and intersections. More caution and speed limits should be enforced during rainy and wet conditions since they are the second lead cause of car accidents (in this model). Installing signs and lights that light up in the dark will be very helpful in cautioning drivers at night. If we want to predict the severity of car accidents solely based on these environmental and road factors, the decision tree would be a good model to use.

6. Conclusion

In this project, we focused on finding the major environmental factors and road conditions that affect car accidents, as well as building a model that can help predict the severity of car accidents based on these conditions. We cleaned our data and prepared it for exploratory data analysis and model building. We fit four machine learning models on our data and determined which model produces the most accurate predictions. Based on our analysis and results, we made some recommendations to improve the safety of drivers and those on the road during certain road and weather conditions. Finally, we suggested a model that produces the best results for further analysis of car accident severity based on the same predictors.