Predicting Car Collision Severity

Ariella Goldman
Coursera IBM Data Science Capstone Project

September 11, 2020

1 Introduction

1.1 Background

In 2019, there were 16,000 car crashes per day in the United States. Collisions between motor vehicles often result in physical damage or injury (or both). In 2013, the average auto liability claim for property damage was $3,231. In contrast, the average auto liability claim for bodily injury was $15,433, according to Verisk Analytics.

3 million people are hurt in car crashes in the US every year. According to the National Highway Traffic Safety Administration (NHTSA) from a 2014 study, US motor vehicle crashes in 2010 cost almost $1 trillion in loss of productivity and loss of life. Private insurers pay approximately 50% of all motor vehicle crash costs. The amount of money spent by insurance companies as a consequence of car collisions is tremendous. However, there's a large contrast between the money spent on property damage vs. physical injury. Therefore, it is advantageous to predict the severity of a car collision. This information can be used by insurance companies to help empower drivers to drive more safely. If something could warn a driver, given for instance, the weather and the road conditions, about how severe a potential accident would be, a driver might drive more carefully or perhaps change travel plans.

1.2 Problem

Data that might contribute to car collision severity might include his weather road conditions, light conditions, hour, day of week, and driving under the influence. This project aims to predict the severity of a car collision based on these data.

1.3 Interest

Obviously, private insurance would be very interested in accurate prediction of car collision, in order to reduce severity and spend less money. Drivers themselves would also be very interested, in addition to the government, who has a vested interested in human life.

2 Data acquisition and cleaning

2.1 Data sources

The dataset regarding car collisions was provided by coursera. It can be downloaded here. The data dates weekly from 2004 to present. The data has been collected from the Seattle Department

of Transportation. The data set contains information about car collisions concerning both external and internal aspects. External factors include weather, road conditions, light conditions, day of week and time of day. Internal factors include collision address type (i.e. alley, block, intersection), whether a driver was speeding, whether the collision was due to inattention, whether a driver was under the influence of alcohol or drugs.

This data set contains 194,673 rows. There are 37 attributes, some numeric and some categorical. Some attributes have missing data. The data set is labeled with the severity of the accident - "1" or "2" indicating property damage or injury, respectively. Severity level 2 "injury" is considered more extreme than severity level 1 "property damage". This is not a balanced labeled dataset. Metadata about the dataset can be found at here.

2.2 Data cleaning

The car collision data from the Seattle Department of Transportation contains 194,673 rows and 38 columns (including the target label). About 70% of the data was labeled as 1 and 30% labeled as 2. In order to achieve a balanced data set, I downsampled the majority class so that the entire data set was equally distributed among the two labels.

With 38 columns, I needed to do some organization and exploration. I renamed every column for ease of reading. In my original explorations, I noticed that a few attributes were identical to other attributes. This results in me dropping 4 attributes immediately. I then double-checked the meaning of each attribute and decided whether the column contained useful information. Some of these attributes contained unique data for each and every row, which I decided was not helpful for the purposes of finding patterns in the data for classification. I dropped these attributes from the pandas table. In total, I dropped 17 columns from the table before investigating more deeply. Before selecting features, I checked the distinct values for every attribute in order to best understand the data.

While looking at individual attributes, I noticed that some numeric attributes had extraneous detail. For instance, the attribute that counts the amount of people involved in the collision had values from 0 through 82. However, the total amount of rows decreased as people involved increased. For simplicity, I grouped everything 5 and up together and replaced all numbers greater than 5 with 5. I cleaned up other attributes similarly. The attribute about light conditions had categorical values but I did group some together due to low frequency and commonality of value. I grouped together "Dark - Street Lights On", "Dark - No Street Lights", "Dark - Street Lights Off", "Dark - Unknown Lighting" as just "Dark". I decided it was appropriate to do this because each of the "Dark" values had similar target class label percentages.

One of the attributes contained the date-time of the collision. I created new columns for year, month, day, day of week, time, and hour. I noticed that the most collisions occurred in the mid afternoon.

Some seemingly-binary attributes have only one distinct value in addition to null values. I replaced those missing values with the other binary option. For example, the column about driving under the influence only had "Yes". I decided that "No"s just were not filled in and I filled in the rest of the data with "No." (In this case, I really filled it in with 0). Some attributes had categorical data that could easily have been represented by numeric data. I cleaned up the attribute about speeding from "Yes" and "No" to 1 and 0. I dealt with several other attributes similarly.

2.3 Feature Selection

After discarding redundant features and data cleaning, I inspected the correlation of independent variables, and found several pairs that were moderately correlated. After all, 3 features were selected. I decided to focus on weather, road conditions, and light conditions. These features are the most logical to have influence on the target variable. Additionally, these features are variables that are available as information to help a driver make a decision as to drive a car somewhere that day.

The features I chose were all categorical data. I converted them all to quantitative variables by using One Hot Encoding. I removed rows that had any null values for any of the features and ended up with 113,430 rows.

3 Exploratory Data Analysis

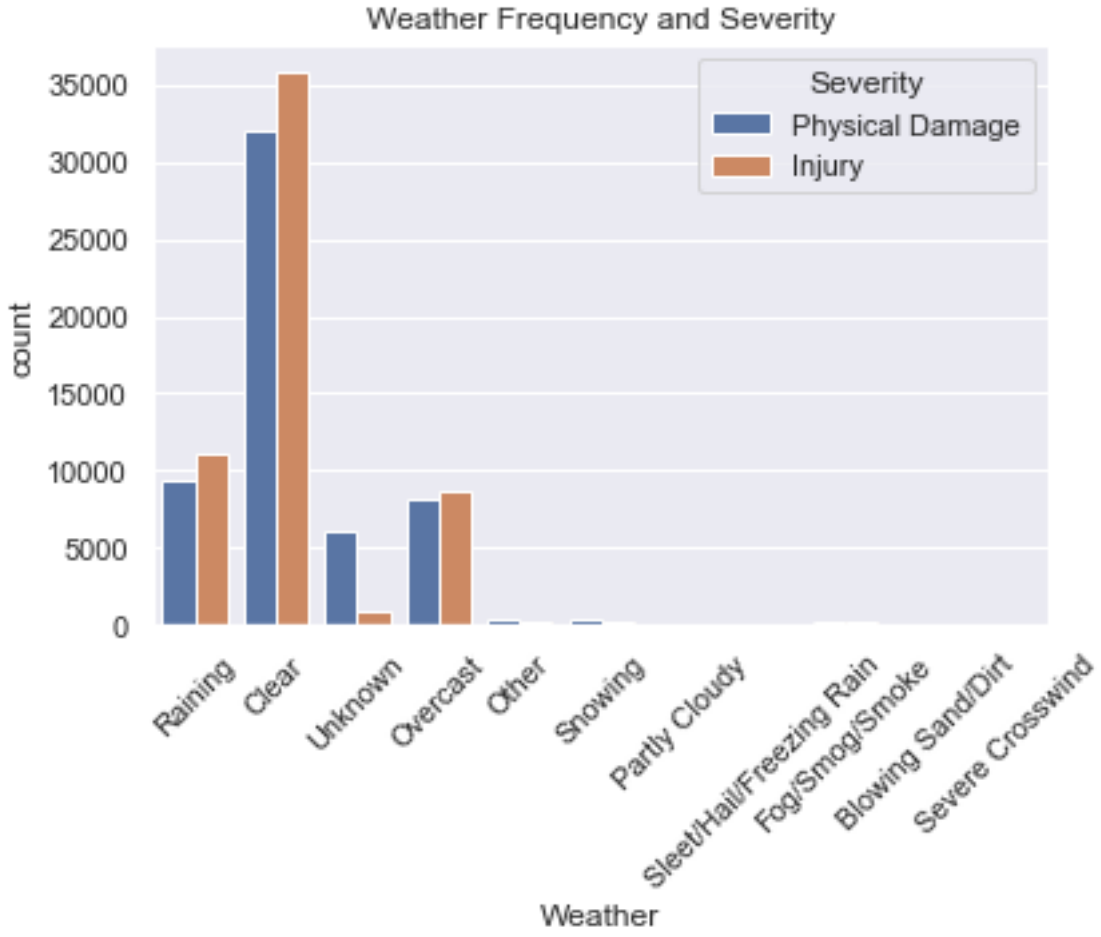
3.1 Calculation of target variable

I changed the labels of the target variable. The target variable originally had two classes: '1' and '2'. After converting the String class values into numeric values, I switched "1" to 0 and "2" to 1. No rows had null target values.

3.2 Relationship between severity and weather

Weather, categorical data, had 11 distinct values. After removing rows with nulls, I used One Hot Encoding to convert weather to quantitative data. The vast majority of car collisions with weather data occurred on clear days, with an appreciable amount of rainy days and overcast days. Several other weather types happened quite infrequently. I thought that rainy days would correlate with more severe car collisions but the data indicates that it is only moderately so (Figure 1). Only Weather = "Unknown" has more physical damage than injury, of the values that even make it onto the graph.

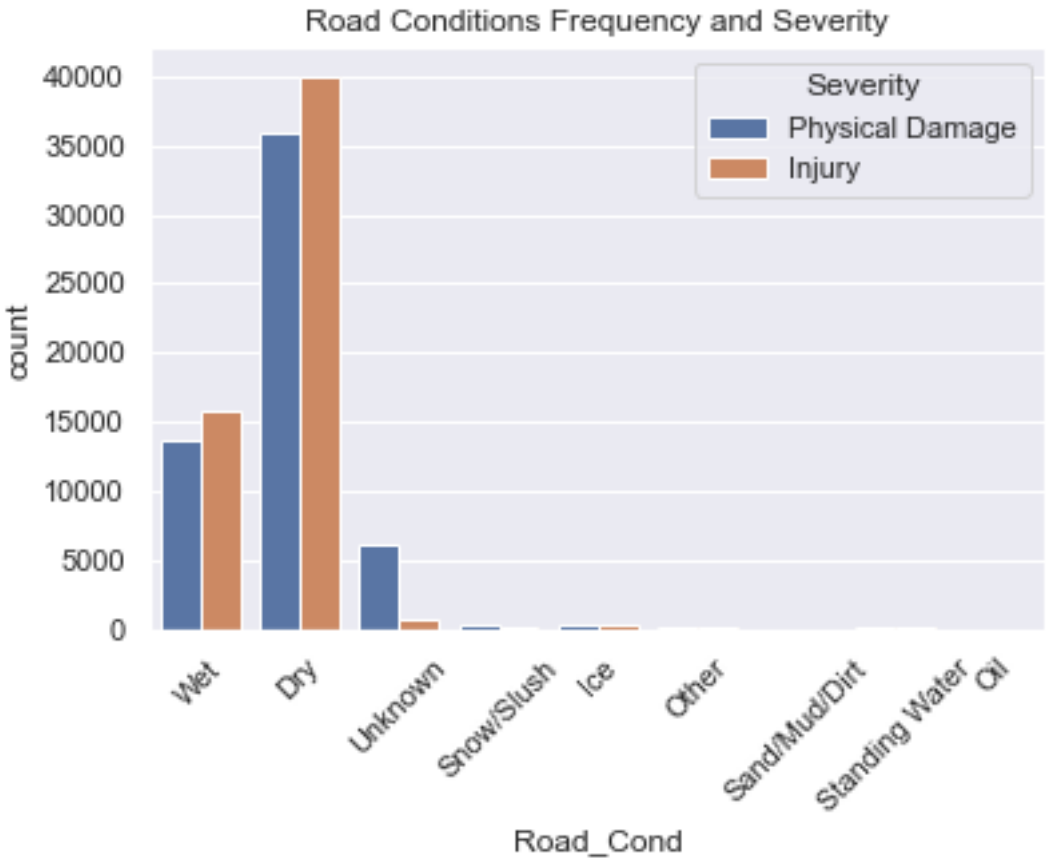
Figure 1: Relationship between Severity and Weather



3.3 Relationship between severity and road conditions

Road conditions, also categorical data, had 9 distinct values. After removing rows with nulls, I used One Hot Encoding to convert road conditions to quantitative data. The majority of car collisions with road conditions data occurred on dry roads, with an appreciable amount of wet roads and some unknowns. Several other road condition types happened quite infrequently. Both dry and wet roads and car collisions co-occur more often with injury than physical damage, but unknown road conditions much more highly correlate with physical damage (Figure 2).

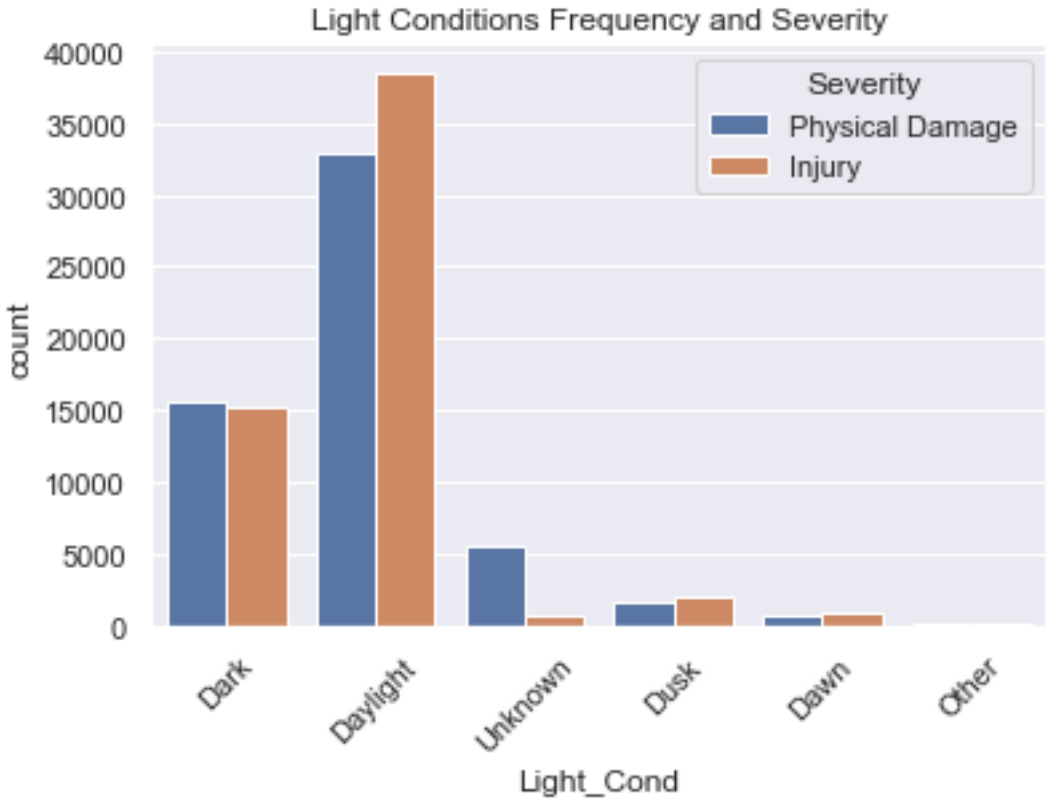
Figure 2: Relationship between Severity and Road Conditions



3.4 Relationship between severity and light conditions

Light conditions, also categorical data, had 6 distinct values, after adjusting in data preparation. After removing rows with nulls, I used One Hot Encoding to convert light conditions to quantitative data. The majority of car collisions with light conditions data occurred during daylight, with an appreciable amount during the dark and some unknown. Several other road light types happened quite infrequently. During daylight, injury is more common but in the dark, both physical damage and injury resulting from car collision are basically equally likely. (Figure 2).

Figure 3: Relationship between Severity and Light Conditions



4 Predictive Modeling

4.1 Classification models

The application of classification models was much more straightforward. I divided the samples into two classes (severity == 0 or severity == 1). The number of samples in each class were the same. I chose Jaccard accuracy score as the metric here. K-nearest Neighbors, decision tree, random forest, and logistic regression were tuned and built. Among the individual models, the random forest model performed the best (48% accuracy). All other models performed within 0.02 % in Jaccard accuracy.

Algorithm	Jaccard	F1-Score	Log-loss
KNN	0.46	0.54	N/A
Decision Tree	0.47	0.54	N/A
Random Forest	0.48	0.53	N/A
Logistic Regression	0.47	0.54	0.67

5 Conclusions

In this study, I analyzed the severity of car collisions. I identified features that affect severity of car accidents. I built classification models to predict severity based on the features. These models can be very useful in helping private insurance companies reduce money spent.

Further analysis on this data set should be done to determine the impact of other independent variables on severity level. For instance, including collision type as a feature might improve accuracy. However, the machine learning models can take a long time to run.