Predicting Accident Severity From Road and Lighting Conditions

Ethan Feldman

September 1, 2020

1. Introduction

1.1: Background

As we have increased the number of drivers on the road over the course of the last century we have also seen an unfortunate increase in the number of accidents. As car manufacturers and city planners have worked diligently to not only decrease the likelihood of accidents but also the resultant severity of accidents when they do unfortunately occur, there is still work to be done to better predict the occurrence and level of severity.

1.2 Business Use Case

In this project we seek to answer a question on the behalf of the Seattle Department of Transportation and the drivers of the city of Seattle regarding car accident severity. Namely, we want to be able to predict if accident severity, in particular if an accident results in soley damage to property or personal injury, based on the conditions of the road and the lighting at the time of the accident. If we can answer this question well enough, can we use it to change our education of drivers or send alerts during particularly treacherous conditions?

2. Data

2.1 Data Sources

For the purposes of this project we are using a dataset of all accidents recorded in Seattle by the Seattle Police Department from 2004 to present. We have isolated only the data pertaining to road conditions based on factors such as rain, standing water, mud, oil, or other factors and those for lighting conditions based on time of day and street lights. We have coded our severity scores as either accidents that result in damage to property or those that also include injury to those in any vehicle involved.

2.2 Data Cleaning

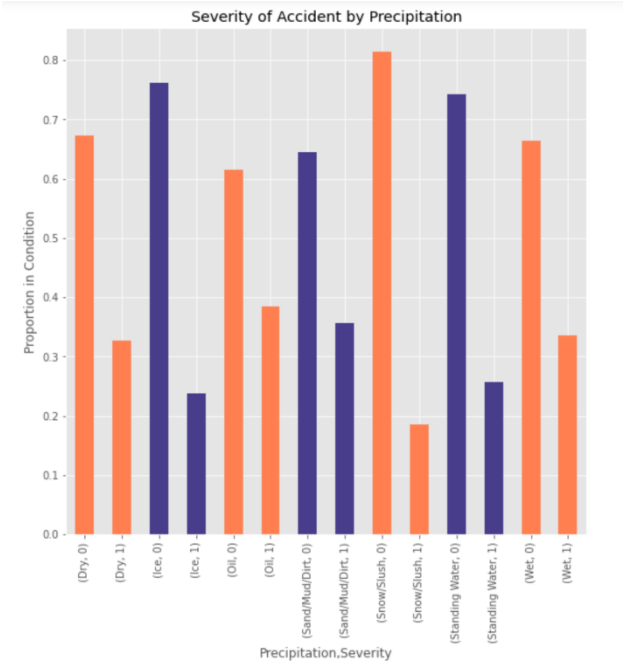
Data was read from a CSV file and into a Pandas dataframe for cleaning, structuring, and analysis. There were a number of pieces of data that were included that were deemed not pertinent to our analysis. Included in this were items such as descriptions of the accident, categories of locations, counts of pedestrians and passengers, and the data, among others. These items were not related to the analysis of the impact light and road conditions on accident severity.

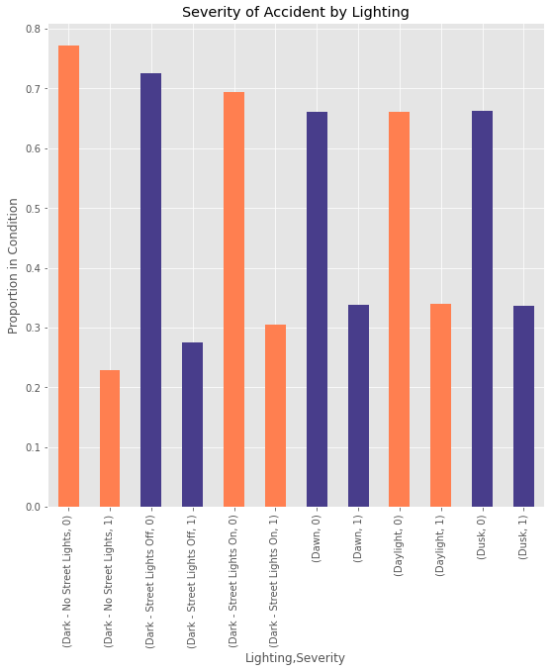
Once the data had been reduced to pertinent items, we identified and dropped any accidents which were missing data from any column, as it would be impossible to know accident severity or conditions without them being given. Furthermore, many accidents listed lighting or road conditions as "Other" or "Unknown." For the same reasons as missing data, these items were removed from our data.

Then, we moved on to dealing with the imbalance in our target of severity, with many more accidents being damage to property but not injuries to passengers. If left unchecked, we would expect our model to run with relatively high accuracy by always predicting solely property damage, but not actually providing any kind of prediction based on lighting and road conditions. This imbalance is not the result of improper sampling and should not be ignored, as it points to the truth that most accidents do not result in injuries to passengers, but still may result in costly damage for drivers. We have chosen to downsample the accidents with property damage to represent a set of the same size as the accidents with personal injury. In this way, our models can analyze for structure between the target and the feature set, and not get tricked by the imbalance of real outcomes in our target.

3. Exploratory Data Analysis

We can see a quick visual of our data broken down into two graphs below. The figure on the left is of severity, across the feature of precipitation and road conditions. The figure on the right is of severity across the feature of lighting. In this table, low severity is marked as 0 and high severity is marked as 1.





It is immediately clear from both of these diagrams that the bias in our target variable is very clear. In each instance, the low severity outcome is very heavily weighted. This will be especially meaningful in our logistic regression model, so we will rely on the downsampled data for more meaning there. We can also see that the disparity between low and high severity does not appear constant across the differing road and lighting conditions.

Most striking in these graphs is under which conditions do we find the highest proportion of high severity. While it may seem like common sense that snow, rain, or darkness would increase the rate of higher severity accidents, we can see that the accidents with the lowest relative frequency of high severity are, in fact, just those cases! Surprising indeed.

4. Predictive Modeling

There were four types of models that we used to predict accident severity: logistic regression, decision trees and random forests, and k-Nearest Neighbor classifier.

4.1 Logistic Regression

We applied our logistic regression model to both our original and downsampled data set, attempting to provide predictions between the two classes of high and low severity accidents. As you can see in the tables below, there is an interesting interplay between the different metrics for accuracy. While using the original data yields us our highest accuracy (f1) score, there is a clear issue when the f1 score is broken down between predicting low and high severity.

The model based on the original data only predicts low severity! Just like a model that predicts tails on ever coin flip, it's going to be right a certain percentage of the time, it just isn't a model that is making predictions on anything other than the original bias in the data. The second model, completed based on the downsampled data that removes the bias in the target data set (notice the balance in the support column) reports an overall lower accuracy score, but the score comes from truly predicting between the two outcomes based on our feature set. These results are certainly interesting, but the accuracy seems too low in the downsampled case to be our preferred method and without enough meaning in the original case.

Downsampled

	precision	recall	f1-score	support
0	0.53	0.31	0.39	10931
1	0.52	0.73	0.60	11069
accuracy			0.52	22000
macro avg	0.52	0.52	0.50	22000
weighted avg	0.52	0.52	0.50	22000

Original

	precision	recall	f1-score	support
0	0.67	1.00	0.80	22505
1	0.00	0.00	0.00	10985
accuracy			0.67	33490
macro avg	0.34	0.50	0.40	33490
weighted avg	0.45	0.67	0.54	33490

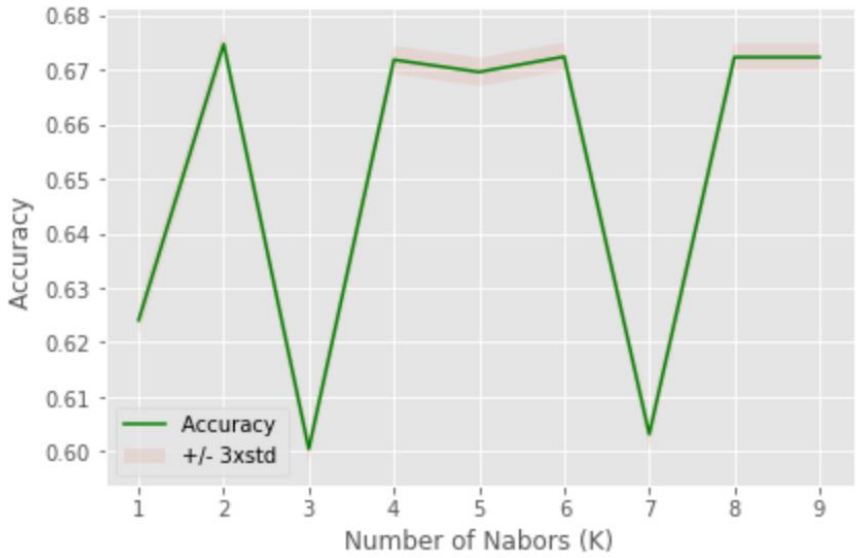
4.2 Decision Tree and Random Forest

The applications of these models are each more straightforward. The need for down sampling is less important, which can be observed by the down sampling actually hurting the accuracy metrics across the board. These models do not have the same issue as the logistic model, in that they are able to predict outcomes as low or high severity, without needing to modify the data to remove the bias in the target data set.

Our decision tree classifier returns an overall accuracy of 67.4%, while the Random Forest Classifier had an accuracy of 67.19%. These two methods provide a similar level of accuracy

4.3 k-Nearest Neighbor

Our final model is the k-Nearest Neighbor, which seems to find similar accidents to label as "nearby" based on their similarity as a measure of distance. The k, in this model, is an attempt to determine how many neighbors are used at a time for a measurement of similarity for predicting where future data points may be classified. In our case, the best accuracy found was using k=2, giving as accuracy of 67.46%, as seen in the table below.



5. Conclusions

In this project, we have created and evaluated a number of models for predicting accident severity based on lighting and road conditions. These can be of great use to first responders in allocating resources, cities in planning for increased lighting or advisories for upcoming weather, or for driver's education about potentially dangerous driving situations to avoid if possible. We can see from our metrics that the two models that will provide our most accurate predictions going forward would be the 2-Nearest Neighbors and the Decision Tree Classifier created from the original data set.

6. Future Directions

This data set provided a wealth of information, but missed a elements that we know exist as well. For example, no data points were for accidents with highest levels of severity in accidents (accidents no just with injury to passengers but more severe cases involving fatalities).

Models in this study do not explain how or why these accidents occur, simply predicting correlation to particular weather and lighting conditions. They have not included detail into how the lighting conditions might impact a driver different based on time of year, if the drivers are local or not, or how drivers might have been impacted by choices of signage, speed limits, or past experiences.