Predicting Vehicle Accident Severity

Zain Javaid

October 4, 2020

1. Introduction

1.1 Background

Car accidents are one of the leading causes of death worldwide. With over 1.25 million people perishing and 20-50 million sustaining life-changing injuries. With that many afflicted with this problem, it is important that we try as a society to limit how frequently these events take place. There are many factors that either increase or decrease the likelihood of fatal accidents taking place. For example, when it is raining, when visibility is limited, when roads are curved, and traffic congestion. The objective of this project is to help use these factors to show how severe an accident would be if it were to take place on a given day. This information could be used by commuters and vacationers to help drive safer and better.

1.2 Problem

The problem I'm looking to try to solve is predicting how bad an accident would be if it were to take place on that day.

1.3 Interest

This model has a broad interest group with anybody looking to drive could stand to benefit from this model. It would allow any commuter, vacationer, or errand runner to check how bad an accident may be, and to be able to adjust their driving accordingly. Other groups such as law enforcement, EMT workers, and firefighters could use the model to help be ready for bad accidents.

2. Data

2.1 Data Description

The data I'm working with is a 194,674-row dataset with 38 attributes to work with. Many of these attributes are made to help create very specific groupings of accidents, while others describe the location of the accident, the weather conditions on the day of the accident, the type of vehicles involved in the accident, etc.

2.2 Data Source

The data used to train this model originated from the Seattle Police Department and was recorded by Traffic Records. The data goes from 2004-present and is updated weekly.

2.3 Attribute Selection

For this model, I will only be using attributes relating to weather conditions on the day of the accident. As such, all other attributes will be dropped.

2.4 Data Cleaning

In this dataset, many of the values are strings that are no good for the model. Therefore, I had to go through and replace words such as ‘Clear’ and ‘Rainy’ with integer values such as 0 and 1. If I didn’t do this I would be dealing with a Logistical Regression problem rather than a linear regression problem. I also had to deal with over 15,000 instances of NaN values which was very easy to fix with a simple fillna(0).

3. Methodology

3.1 Machine Learning Usage

Within this project, I used Linear Regression to help get predicted values. While I am not as familiar with Linear Regression as I am with other Machine Learning model types such as Binary Trees, KNN, or clustering, I figured I might as well give it a try.

3.2 Statistical Testing

In order to help understand what correlations exist, I decided to create a correlation chart.

	WEATHER	ROADCOND	LIGHTCOND
WEATHER	1.000000	0.786270	0.354345
ROADCOND	0.786270	1.000000	0.401874
LIGHTCOND	0.354345	0.401874	1.000000

This chart depicts the correlations I had between my 3 chosen attributes. As you can see, the column LIGHTCOND has very little correlation to the other 2 attributes. Therefore, it will not be used to train the model I’m working on.

4. Results

4.1 Model Performance

Ultimately, it showed in my model performance how inexperienced I am with Linear Regression. While I was able to get the model to work and was able to clean the data, the results produced were underwhelming. I believe the model would have performed significantly better if it weren’t for the formatting of the data. Ultimately converting string values to integer values without it having a significant impact on the model’s results is difficult. While I was determined to perform Linear Regression on this dataset, the results would have probably been better if Logistical Regression was used.

5. Recommendations

After working on this project, I can hopefully give some recommendations that will be helpful to you. The first one I can give is to know what types of Machine Learning will produce the best results for your dataset. I made the mistake of attempting to perform linear regression on a dataset that would be much better suited for a Logistical Regression or Binary Tree model. Additionally, when working with a large dataset like the one I was using, make sure to write the most efficient code you can. This will avoid lofty runtimes and will keep you from pulling your hair out.

6. Future Directions

Within this model, I was not able to achieve high accuracy in predicting accident severity. There is a lot of room for improvement in terms of replacing the type of Machine Learning model done on this dataset. I also think that doing single linear regression didn't fully capture the nuances of the factors that may contribute to how bad a vehicle accident could be. While this iteration may not have yielded ideal results, with more optimization, this could become a highly useful tool for travelers and commuters alike.