Car Accident Severity in Seattle

DATA ANALYSIS

Daniel Christman | Coursera Capstone | 9/29/20

Introduction

BACKGROUND

Seattle is home to many, with a large amount of people commuting into the city from outside suburbs. Many large corporations, like Amazon, Nordstrom, Starbucks, and Microsoft, are headquartered in and around the city, drawing a immense number of residents into the area. With Seattle being known for its unpredictable and often rainy weather, there has been a high number of accidents on the roadways.

PROBLEM

In Seattle and across the globe, there exists a constant variable for all commuters, that being weather and road conditions. On a given day where road conditions are poor and weather conditions are creating a challenging environment to drive safely, a commuter will face a difficult decision in determining whether it is safe to drive. Currently, a commuter would have no way of predicting the likelihood of getting into a severe accident. Even if the commuter is confident in their ability to handle the weather, their selected route may take them down a road where conditions will greatly increase their risk of a severe accident.

Luckily, Seattle has made available the data surrounding accidents to allow data scientists to analyze and extract trends to better predict the main environments and scenarios where accidents may be more common to occur.

Data

DATA SUMMARY

This dataset breaks down severity into two categories: 1 meaning that there was no injury sustained in the accident, just property damage; 2 meaning that there was an injury sustained. This dataset also contains data regarding the location of the accident, description of the severity, amount of persons involved in the accident, amount of vehicles involved in the accident, the type of junction in which the accident occurred, whether the driver was under the influence, weather, road condition, light condition, and whether the driver was speeding.

DATA SELECTION

This model is mainly concerned with the data surrounding accident severity, weather conditions at the time of the accident, road conditions at the time of the accident, and light conditions.

DATA PREPARATION & CLEANING

In order to make better use of the data I wanted to work with, I first began by adjusting the severity codes in the dataset from 1 and 2 to 0 and 1. Following this, I changed the weather conditions from string attributes to numeric values. The reasoning for this change in data was to make it possible to run machine learning models on the data. I also condensed some of the attributes in the weather column to condense those that were similar. For example, I put together "Severe Crosswind" and "Blowing Sand/Dirt" into one category labeled "Windy" set to value 1. The complete list of weather criteria are as follows:

- 0 = Clear
- 1 = Cloudy
- 2 = Windy
- 3 = Foggy
- 4 = Precipitation

Next, I completed the same process on the light condition column of data. I converted the string attributes that were included in the dataset into numeric entries to better allow for manipulation and analysis. I also condensed similar lighting conditions into common identifiers, with those identifiers being set to a numeric value. For example "Dawn" and "Dusk" were both set to "Dim" lighting, receiving a numeric value of 1. The rest of the classifications are as follows:

- 0 = Bright
- 1 = Dim

- 2 = Dark

Next, I once again completed this process on the road condition data. I took the preset attributes and converted them into numeric values while also condensing similar attributes. For example, "Wet" and "Standing Water" were both set to represent "Slick" and were each given a common numeric value of 2. The rest of the classifications are as follows:

- 0 = Dry
- 1 = Coarse
- 2 = Slick

Following this, I created a new dataframe containing only the variables I was using in my analysis, that being the unique identifying key, the severity code, weather conditions, light conditions, and road conditions.

Because some indices in the dataset contained values such as "Unknown" or "Other", I decided to drop these rows from the table, as well as dropping rows where the values in my selected columns were null. This would allow me to work with data that was all of the same integer type.

	INCKEY	SEVERITYCODE	WEATHER	LIGHTCOND	ROADCOND
0	1307	1	1	0	2
1	52200	0	4	1	2
2	26700	0	1	0	0
3	1144	0	0	0	0
4	17700	1	4	0	2

Methodology

EXPLORATORY DATA ANALYSIS

In analysis of the data, I found that there were a much higher number of instances in which an accident only caused property damage, represented as 0 in the dataset, in comparison to accidents that caused physical injury, represented as 1. In further analysis of the data, I found that the vast majority of cases of accident were linked to either road conditions, weather conditions, or lighting conditions.

My main goal in analysis was to understand if the combination of these three factors was leading to a higher likelihood of accident in which physical injury was present.

MACHINE LEARNING MODEL

I decided to implement a decision tree classification model, as I felt it was the best way to determine whether the combination of inclement conditions would lead to a higher likelihood of accident containing a physical injury vs an accident only containing property damage.

Results

DECISION TREE CLASSIFICAITON

The decision tree classifier was the chosen method for predicting the likelihood of a severe accident based on the given criteria. Entropy was the chosen criterion for the classifier, and the maximum depth of the model was set to 5.

```
Decision Tree Classifier accuracy score:  0.6728642033419628


              precision    recall  f1-score   support

           0       1.00      0.67      0.80     50988
           1       0.00      0.00      0.00         0

   micro avg       0.67      0.67      0.67     50988
   macro avg       0.50      0.34      0.40     50988
weighted avg       1.00      0.67      0.80     50988
```

Discussion

A noted observation of the study is that there is a more probable chance of an accident that will cause injury in the presence of one or more of the factors considered. The model was able to predict the likelihood of this outcome with a .67 recall accuracy and .80 f1-score.

RECOMMENDATION

I would recommend that travel advisories be provided to residents of Seattle in the instance where inclement weather is active or predicted to be active in the near future. If possible, advisories should be sent to users of common mapping applications to inform them if the route they select takes them through a route with inclement road conditions, weather conditions, or lighting conditions, with higher alert advisories sent if there is a combination of factors present.

Conclusion

In conclusion, this report discusses the main factors that increase the potential of accidents that result in injury, as well as how the likelihood of these accidents can be affected by the presence of one or more of these factors. This report then discusses a model that can be used to predict the likelihood of an accident being one that causes injury based on the number of inclement factors present.

With inclement weather always posing a threat to residents of Seattle, it is necessary for the city to take measures to provide advisories to its residents when the likelihood of accidents is high. Exploring the data of these accidents provides an excellent insight into the main causes and ways in which these causes can be predicted and mitigated.

References

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv