Zachary Strain
Applied Data Science Capstone
Capstone Project
September 25, 2020

Predicting Traffic Accident Severity

I. Introduction and Business Problem

In the United States, traffic accidents involving motor vehicles cause over 100 deaths per day, and are a leading cause of death in the US, as reported by the Center for Disease Control and Prevention (CDC)[1]. In addition to the cost of human life, there is an enormous economic impact as well, which exceeds $75 billion for the cost of related productivity losses and medical care.

This project seeks to explore a range of factors that may affect the severity of automobile accidents. For the purpose of this project, and given the 3rd-party data used for analysis, "severity" is to be understood as the impact on the traffic as measured by the length of the traffic delay (in terms of time) caused by the accident. Insights from this project would be particularly useful to city transportation, safety, and zoning departments.

By understanding what factors contribute to more severe accidents, city officials would be able to plan better and safer cities, and more effectively deploy city resources in response to traffic accidents. If there are patterns to location, weather condition, and points of interest around traffic accidents, then city officials would be better able to address this and related issues.

“economic impact […] exceeds $75 billion for the costs of related productivity losses and medical care.”

II. Data

The data[2] used for this project is from a dataset made available through the research from the following papers:

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, arXiv preprint arXiv:1906.05409 (2019).

[1] https://www.cdc.gov/motorvehiclesafety/costs/index.html

[2] https://www.kaggle.com/sobhanmoosavi/us-accidents

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. “Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights.” In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

Collected between February 2016 and June 2020, the data includes approximately 3.5 million traffic accidents, and was collected from 49 states in the US. The dataset includes attributes that detail:

- weather conditions (i.e. temperature, precipitation, wind speed, etc.),
- location information (i.e., coordinates, street address, city, state),
- points of interest nearby the traffic accident (i.e., crossing, speed bump, station, railway, stop, traffic signal),

and other attributes pertaining to the recorded traffic accidents.

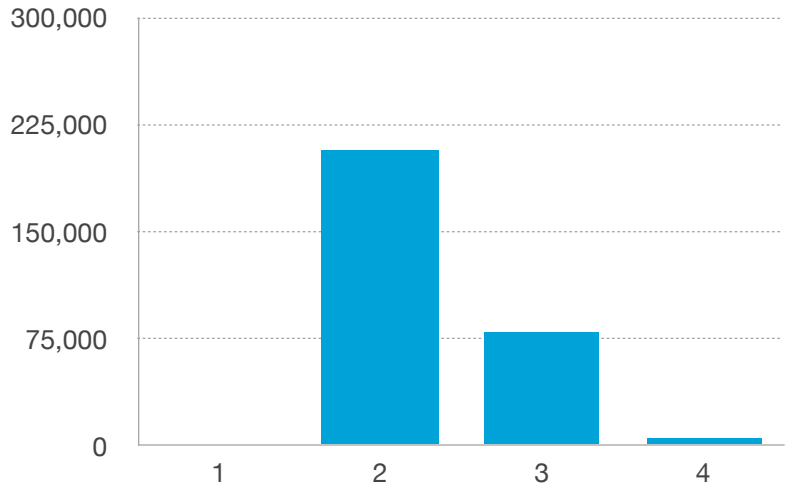
The variable of interest for this paper is ‘Severity’, which ranks the severity of an accident on a scale of 1 (least severe) to 4 (most severe). Severity is given by the impact of the accident on traffic in terms of the length of delay it causes, i.e., an accident with a severity ranking of 4 caused a longer delay to traffic than an accident with a severity ranking of 3. Using the described dataset, I will build a model that could be used to predict the severity of accidents.

III. Methodology

Outcome Variable

Firstly, let’s begin with an overview of the dataset. Initially, it included 3,513,617 rows and 42 columns[3]. To simplify the analysis, I decided to focus on one state, Texas, which accounts for 290,975. *Figure 1* illustrates the outcome variable for this report, `Severity`, in terms of value counts.

Figure 1: Number of Accidents by Severity Value

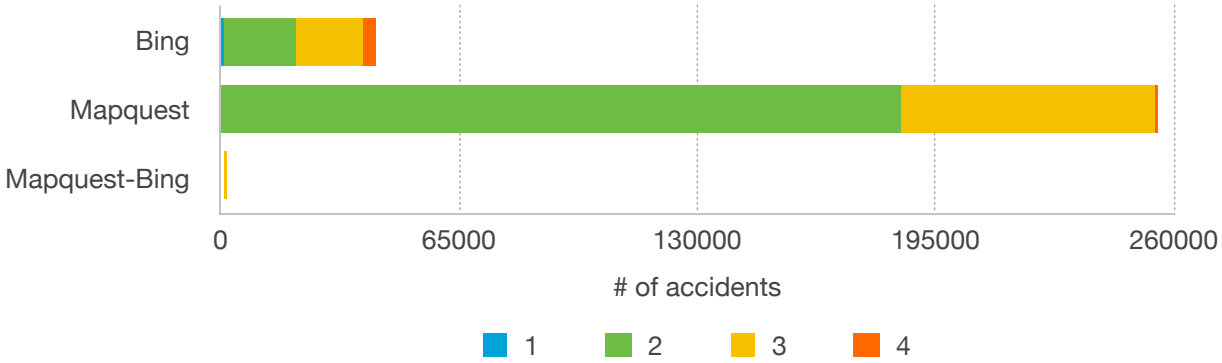


[3] Descriptions of the columns can be found at the following link: https://smoosavi.org/datasets/us_accidents

The most common accident `Severity` was 2, with over 206,326 observations. Accidents with a value of 2 account for more accidents in the dataset than the other three categories combined.

However, the data comes from 3 different sources, as seen in *Figure 2* below.

Figure 2: Severity of Traffic Accidents by reporting source



As illustrated above, there's a massive difference in distribution of `Severity` levels between Bing and Mapquest. For that reason, I'll focus on the Bing data, which accounted for 42,101 rows left in the dataset.

In order to better understand the outcome variable, `Severity`, which is supposed to be positively correlated with the length of the delay caused by the traffic accident, I calculated the actual delay of each accident using the `Start_Time` and `End_Time` variables.

Table 1

	Delay in minutes
mean	225.920416
Std	360.638261
min	6.600000
25%	29.716667
50%	360.000000
75%	360.000000
max	46334.783333

As seen in the table above, most accidents caused delays of about 6 hours, with the shortest being just over six minutes, and the longest stretching to over 32 days. As this last observation is an outlier, I created a max threshold to eliminate such observations, after doing so left 42,096 rows.

Even though the `Severity` variable is supposed to positively correlate with the delay caused by an accident, the table below suggests this is not the case, at least in the context of Bing data from Texas. Starting at a `Severity` of 2, the average delay decreases as `Severity` increases.

Table 2

Severity	delay_minutes (avg.)
1	49.694777
2	260.646633
3	199.714251
4	179.972504

Since `Severity` is an ordinal rating, I converted it to a binary so that I could run logistic regression instead of ordinal logistic regression. The values of the new `Severity` variable is to be understood as the following:

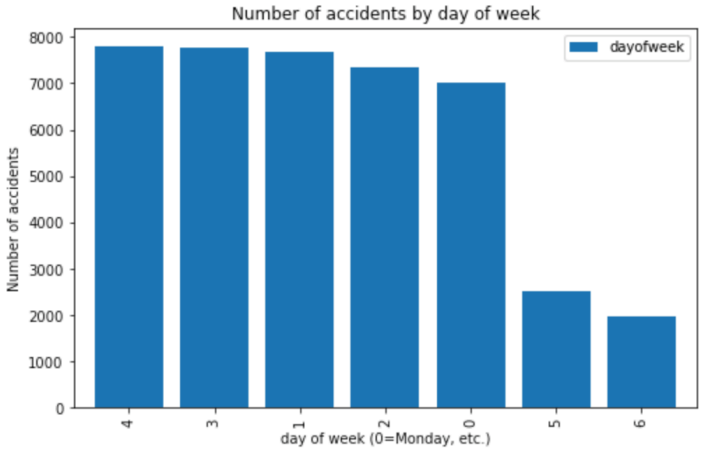
```
0 = initial Severity value of 1 or 2
1 = initial Severity value of 3 or 4
```

This transformation simplifies the model and provides a more even distribution of outcomes; `Severity` of 0 has 21,315 observations; 1 has 20,781.

Frequency and severity of accidents by hour and day of week

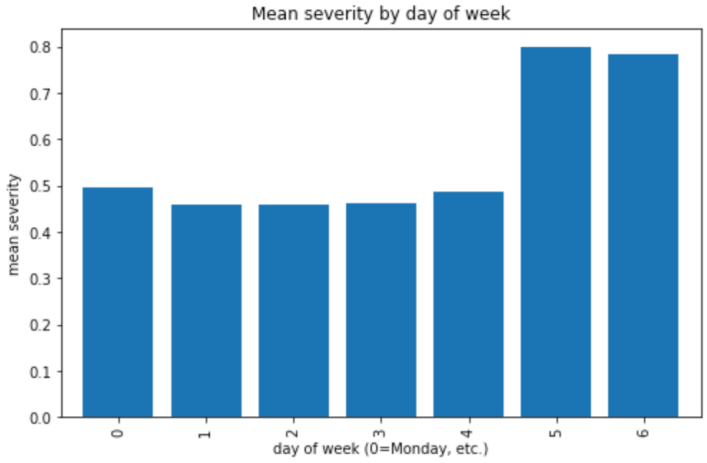
As seen in the figure below, most accidents happen during weekdays, with Friday having the most.

Figure 3



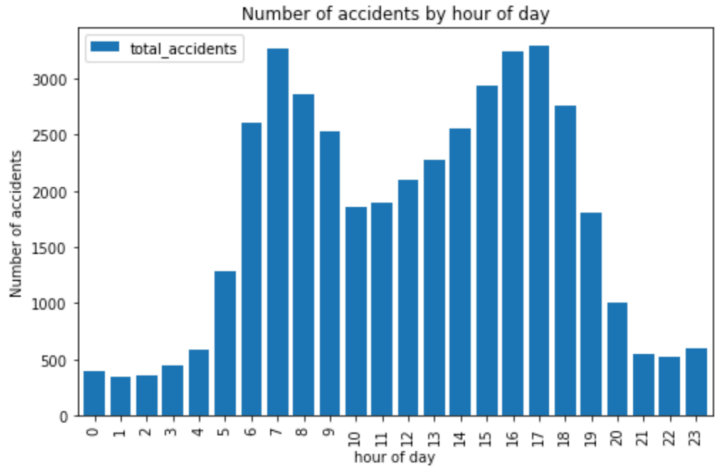
However, while weekdays (Monday-Friday) account for most of the accidents, Saturday and Sunday have a higher average of accident `Severity`, suggesting that weekends are more dangerous than weekdays.

Figure 4



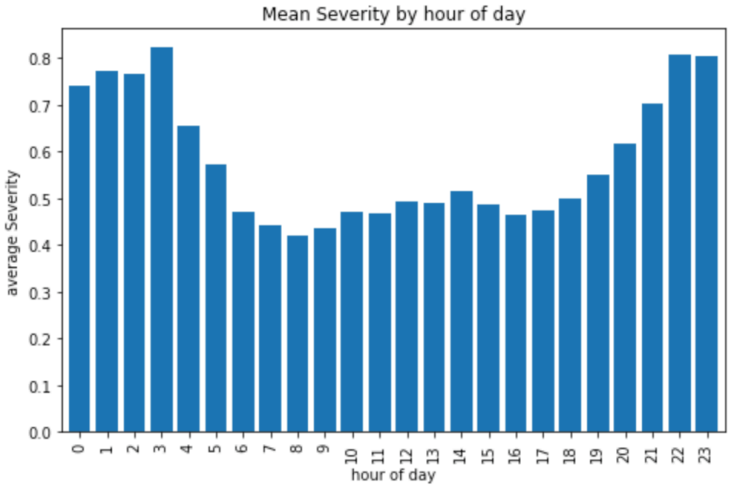
As seen in the figure below, most accidents happen during commuting hours, implying that more cars on the road could equate to more accidents

Figure 5



Yet while most accidents happen during daylight hours, accidents at night are more severe (on average), as seen in the figure below. It is almost an inverse of the previous figure, in that accidents tend to be more severe when they are fewer.

Figure 6



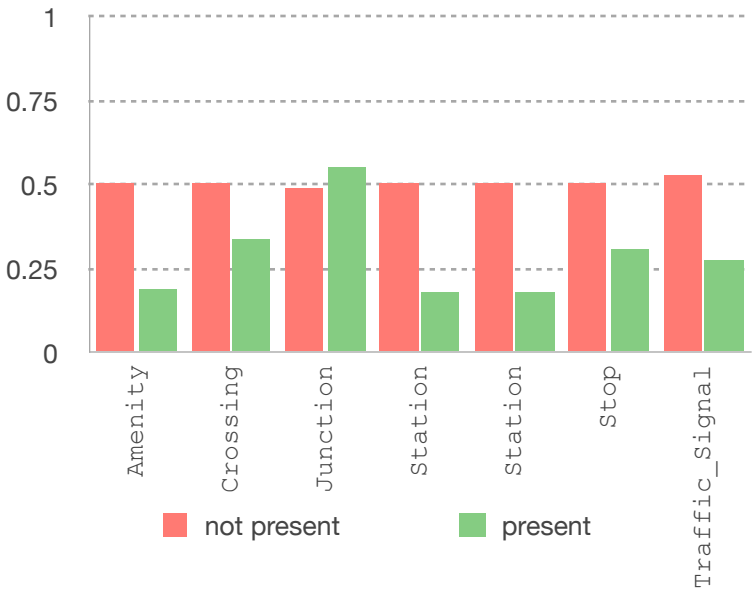
Weather

Weather variables explored, namely, `Temperature(F)`, `Precipitation(in)`, and `Visibility(mi)`, all had weak correlations (-, +, -, respectively) with `Severity`, but were each statistically significant. See this project's notebook for more details.

Points of Interest

Compared to the weather variables, there is typically a more noticeable difference in the mean severity based on the presence of the points of interest (POIs) variables. For most of the POIs, their very presence results in a lower mean `Severity`. While there were more POIs in the dataset than seen in the figure below, some did not appear in Texas, or were too few to be statistically significant.

Figure 7: Mean Severity by presence of POI



Constructing the Model

Based on what was learned from the exploratory analysis of the data and the resulting transformation of the outcome variable, the machine learning model employed in this paper is Logistic Regression. Using statistically significant variables, the model will predict the probability of an accident having a `Severity` rating of 0 or 1.

The model uses the following feature set:

```
X = [['weekend', 'start_hour', 'delay_minutes', 'Temperature(F)',
'Visibility(mi)', 'Precipitation', 'Amenity', 'Crossing', 'Junction',
'Station', 'Stop', 'Traffic_Signal']],
```

with a binary target variable of:

```
y = ['Severity'].
```

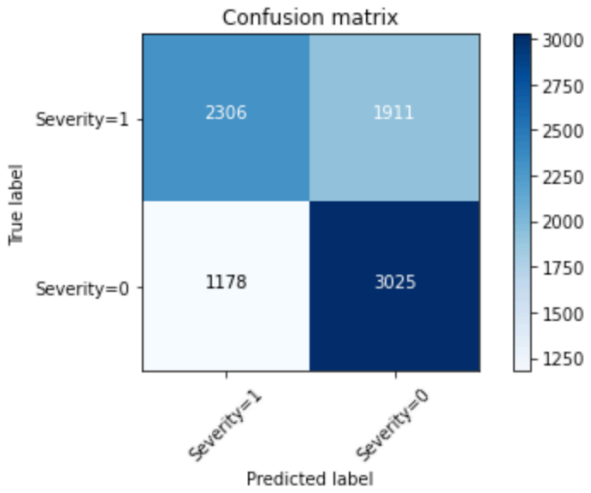
IV. Results

The classification report of the logistic regression model is reproduced in the table below.

	precision	recall	f1-score	support
0	0.61	0.72	0.66	4203
1	0.66	0.55	0.60	4217
accuracy			0.63	8420
macro avg	0.64	0.63	0.63	8420
weighted avg	0.64	0.63	0.63	8420

Based on the classification report above, the logistic regression model was more accurate at predicting less severe accidents (where Severity = 0) than those that were more severe (where Severity = 0), for the test data. The true positive rate (given by recall) of the less severe accidents was higher, and the f-1 score is slightly higher for the less severe accidents, but overall, the model was successful at predicting over half of the outcomes successfully. A breakdown of the number of true labels versus predicted labels is depicted in the confusion matrix figure below. The log-loss score of the model was 0.63483 (rounded).

Figure 8



V. Discussion

Based on the exploratory analysis, most of the variables included in the model had very weak correlation with the outcome variable. This resulted in a model with considerable room for improvement. Therefore, there are potentially more predictive variables that were not included in this paper's feature set. Most notably, geographic variables.

That being said, the exploratory analysis suggests that more severe accidents are more likely to occur where and when there's less traffic, especially at night.

VI. Conclusion

As the results of this paper show, the severity of traffic accidents, as random as they might seem, are influenced by a number of factors. This model looked specifically at environmental variables like weather condition and infrastructure features near the site of the accidents. Results from the logistic regression model suggest that these factors can, to a limited degree, help predict how severe traffic accidents can be. Infrastructural elements are certainly the ones most within a realm of human control, and for those that have a significant impact on traffic accident severity, city planners/developers, transportation officials, etc., would find more robust results of interest.

It is essential to acknowledge the limitations of this model, which are quite significant. Firstly, this model only analyzed data from one source (Bing), and, secondly, focus on one US state (Texas). To conduct a more rigorous analysis, data from multiple sources could be used, but not before establishing an understanding of the differences between sources and how their collection of the data affects it, and therefore the model and results. Additionally, US states can be categorically different on a number of levels when it comes to driving in them, from population, density, the number of commuters, driving laws, and infrastructure design. These are all factors that potentially have an effect on traffic accidents and their severity.