Traffic from US Car Accidents

Zack DeNoto

Bellevue University

DSC 680: Applied Data Science

Dr. Fadi Alsaleem

May 30, 2021

Abstract

This paper will addresses if it is possible to determine how severe traffic will be impacted by a car accident in the US. It will also compare the traffic from accidents prior to Covid-19 and during Covid-19 to see if the pandemic had any impact on the traffic. This paper will demonstrate if it is possible to make such a prediction based on the factors provided in the dataset, as well as determining what factor(s) most affect the traffic. The objective of this project was to find a good dataset, clean the data, determine what factors affect traffic the most, determine the impact from Covid-19 and create the most accurate model(s) possible. The results showed that it is possible to predict the severity of traffic based on the fields in the dataset, with the Covid-19 group having the highest accuracies. Precipitation was the highest correlated field even though all correlations were low. This report also showed that the average traffic impact time from an accident went up during Covid-19 compared to before Covid-19.

Intro

Close your eyes and count to five; now open your eyes. In that short amount of time there was a car accident somewhere in the US. There are on average over 6 million car accidents in the US every year and 3 million people are injured from car accidents every year (Driverknowledge.com, 2021). In recent years, the number of accidents has been higher, with around 12 million vehicles involved in crashes in 2018 in the US due to the level of traffic. The US is one of the busiest countries in terms of road traffic with nearly 280 million vehicles in operation and more than 227.5 million drivers holding a valid driving license (Wagner, 2020). The impact of car accidents goes beyond the individuals in vehicles; it also impacts the other drivers on the road with traffic and with regular citizens paying taxes to fix areas with high amounts of accidents. Car accidents in 2010, for example, had an economic cost of an estimated $242 billion (Miller, 2015).

Though Covid-19 started in late 2019, March 2020 is when the large impact hit the US. This analysis considers March 2020 when Covid-19 hit and separates the data based on March 2020 or prior.

When the pandemic hit, there was a lot less traffic, as many jobs went remote. Last year in 2020 Americans drove 13% fewer miles than in 2019. In places like Minnesota, traffic volumes fell 60% (Apnews.com, 2021). With fewer cars on the road, one would think that there would be less accidents as well as less traffic from less accidents. However, a recent NHTSA report showed that vehicle speeds increased 22% in several metropolitan areas over pre-pandemic numbers (Apnews.com, 2021). Covid-19 also caused drivers to fail to stop at stop signs by an increase of 10%. Additionally, speeding during harsh turns has increased 15% during the pandemic (Teletracnavman.com, 2020). The number of DUI's and speed racing stayed the same or were slightly higher even though there were fewer cars on the road (Pichon, 2020). With more dangerous driving patterns coming from Covid-19, this analysis looks to see if it is possible to predict the severity of traffic impacted from a car accident with data before Covid-19 and during Covid-19.

Data

The dataset used for this analysis was called US Accidents from Kaggle.com, retrieved from the link https://www.kaggle.com/sobhanmoosavi/us-accidents. This dataset contains information on traffic produced from car accidents from the years 2016-2020. The data came from multiple traffic APIs and includes data from 49 states. The csv file had 47 columns and almost 3 million rows with fields such as start time, end time, weather, description, side, precipitation, and wind speed.

Methodology

The first step that needed to be done was to examine and clean the dataset. After using R to explore the dataset, it seemed there were quite a few columns that were unnecessary to the analysis and so they were deleted. After the columns were deleted, I looked for any missing values or null values that could negatively impact the analysis. The dataset contained no missing data, but contained a lot of null values. Depending on what column the null value was in, it was either replaced with a 0 or the row was removed if the column was needed.

Using histograms to view some fields in the dataset, it was clear there were very few outliers and most factors seemed to have a fairly normal distribution, as seen in Table 1. I did not remove the outliers due to the size of dataset; the outlier's effect would have been negligent. The next step was to convert the data that was in text format into integers or floats for modeling. Over ten fields needed to be converted to the correct format. The start time and end time fields were in text format but the data was in datetime format. After converting the fields into datetime format, I created a new column called time to get the difference between end time and start time for the accident.

Once the data was cleaned up, I examined the correlation between some of the fields compared to the severity. After correlations were looked at, the goal was to create models for determining the severity to see if I could accurately predict the severity of the traffic from the car accident. The data was split into two additional subsets based on the date of the accident. If the date of the accident was prior to March 2020, it was labeled as not being in the Covid-19 group, and if it was in March 2020 or after it was put in a group for Covid-19. Due to the amount of rows and the computational limitations, I took a sample of 50,000 rows for each of the groups to use for modeling. Lazy Predict was used initially to get a better understanding of what models would be likely to be more accurate. Then I tested several of the models Lazy Predict showed such as Random Forest, Logistic Regression, K-Nearest Neighbor, Stochastic Gradient Descent (SGD), and XGBoost. I used this process on the non Covid-19, the Covid-19, and combined datasets.

Results

As mentioned in the methodology section, the first modeling step was to use Lazy Predict to get a better sense of what models to then test out. As seen in figure 1 below, it seemed like there were many fairly accurate models for determining Severity for the Covid-19 group.

Figure 1 l An Image from Lazy Predict showing preliminary model results for the Covid-19 group.

```
clf = LazyClassifier(verbose=0,ignore_warnings=True, custom_metric=None)

models, predictions = clf.fit(X_traincov, X_testcov, y_traincov, y_testcov)
models
```

```
100%|██████████| 29/29 [03:19<00:00, 6.88s/it]
```

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
DecisionTreeClassifier	0.78	0.44	None	0.78	0.26
NearestCentroid	0.41	0.42	None	0.50	0.06
BaggingClassifier	0.83	0.42	None	0.80	1.43
XGBClassifier	0.84	0.40	None	0.81	10.51
LGBMClassifier	0.84	0.38	None	0.80	0.75
RandomForestClassifier	0.84	0.37	None	0.80	3.75
ExtraTreeClassifier	0.75	0.36	None	0.75	0.08
PassiveAggressiveClassifier	0.70	0.34	None	0.72	0.26
ExtraTreesClassifier	0.83	0.33	None	0.79	3.06
KNeighborsClassifier	0.82	0.31	None	0.77	12.20
BernoulliNB	0.82	0.30	None	0.76	0.07
LinearDiscriminantAnalysis	0.82	0.30	None	0.76	0.63
AdaBoostClassifier	0.83	0.30	None	0.77	1.20
GaussianNB	0.04	0.28	None	0.01	0.07
Perceptron	0.70	0.28	None	0.71	0.15
SVC	0.83	0.28	None	0.76	61.37
LogisticRegression	0.83	0.26	None	0.76	1.02

The accuracies were fairly high, with most models being above 80% accurate. There was a large range of accuracies from low accuracies with GaussianNB and NearestCentroid being 40% or lower and high accuracies with XGBClassifier and LGBMClassifier achieving 84% accuracy. However, Lazy Predict is just a quick tool to get an idea of how accurate models may be, not how accurate they truly are. I then modeled for Severity with Random Forest, Logistic Regression, XGBoost, Stochastic Gradient Descent (SGD), and K-Nearest Neighbor. The accuracies for the models were 84.3%, 83.0%, 84.4%, 83.1%, and 81.4%, respectively for the Covid-19 group.

Next, I ran the Lazy Predict model and the same five models as tested with the Covid-19 group.

Figure 2 l An Image from Lazy Predict showing preliminary model results for the non Covid-19 group.

```
clf = LazyClassifier(verbose=0,ignore_warnings=True, custom_metric=Non
e)

models, predictions = clf.fit(X_trainnon, X_testnon, y_trainnon, y_test
non)
models
```

```
100%|██████████████████████████████████████| 29/29 [03:19<00:00,
6.89s/it]
```

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
NearestCentroid	0.41	0.45	None	0.47	0.06
GaussianNB	0.05	0.40	None	0.01	0.07
DecisionTreeClassifier	0.65	0.39	None	0.65	0.25
BaggingClassifier	0.71	0.39	None	0.68	1.35
PassiveAggressiveClassifier	0.54	0.38	None	0.58	0.14
XGBClassifier	0.74	0.37	None	0.68	3.30
LGBMClassifier	0.74	0.36	None	0.68	0.74
RandomForestClassifier	0.73	0.36	None	0.68	3.73
ExtraTreeClassifier	0.63	0.34	None	0.63	0.09
ExtraTreesClassifier	0.73	0.32	None	0.67	3.45
LinearDiscriminantAnalysis	0.72	0.30	None	0.62	0.16
KNeighborsClassifier	0.69	0.29	None	0.65	12.87
BernoulliNB	0.72	0.28	None	0.62	0.07
AdaBoostClassifier	0.72	0.28	None	0.64	1.19
Perceptron	0.71	0.27	None	0.62	0.13
LogisticRegression	0.73	0.26	None	0.62	0.93

The Lazy Predict results from the non Covid-19 group as seen in Figure 2 above showed a lower range and lower overall accuracies compared to the accuracies of the Covid-19 group. When modeling the same five models from the non Covid-19 group, the accuracies were 73.2%, 73.0%, 73.7%, 69.6%, and 69.2% in the same order. The last group to model was the combined group. As seen in Figure 3 below, the model accuracies from the Lazy Predict models were very similar to that of the non Covid-19 group. The accuracies for the five tested models used with the other groups were 73.3%, 73.4%, 73.8%, 59.8%, and 69.3% respectively.

Figure 3 l An Image from Lazy Predict showing preliminary model results for the combined Covid-19 group.

```
clf = LazyClassifier(verbose=0,ignore_warnings=True, custom_metric=Non
e)

models, predictions = clf.fit(X_train, X_test, y_train, y_test)
models
```

```
100%|██████████████████████████████████| 29/29 [03:17<00:00,
6.82s/it]
```

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
NearestCentroid	0.40	0.48	None	0.46	0.06
DecisionTreeClassifier	0.66	0.44	None	0.66	0.25
BaggingClassifier	0.72	0.42	None	0.69	1.34
XGBClassifier	0.74	0.40	None	0.69	3.28
LGBMClassifier	0.74	0.39	None	0.68	0.70
RandomForestClassifier	0.74	0.37	None	0.69	3.72
ExtraTreeClassifier	0.64	0.35	None	0.64	0.09
ExtraTreesClassifier	0.73	0.34	None	0.68	3.47
Perceptron	0.66	0.33	None	0.63	0.15
KNeighborsClassifier	0.70	0.31	None	0.66	13.47
GaussianNB	0.06	0.30	None	0.05	0.07
AdaBoostClassifier	0.73	0.30	None	0.63	1.19
LinearDiscriminantAnalysis	0.73	0.29	None	0.63	0.15
BernoulliNB	0.73	0.28	None	0.63	0.07
LogisticRegression	0.73	0.25	None	0.62	0.94

These results show that you cannot always take the results of Lazy Predict to be the true model accuracy, but that it is a great tool to help determine which models are likely to produce. The Covid-19 group and the combined group had several Lazy Predict models in the 70%-79% accuracy range and of the models tested, most accuracies were in or close to that range. The non Covid-19 group had mostly Lazy Predict accuracies in the 80%-89% accuracy range, and when testing out the five models, was also within that range. These results show that based on the dataset you can predict how severe the traffic will be from an accident with a fairly high accuracy depending on the model used, especially since the Covid-19 pandemic started. When looking at the correlations between various fields and severity, the highest correlation was only 0.09572 for the precipitation field. All the correlations were less than 0.10 when looking at the relationship to severity. This analysis shows that even with many low correlated fields, you can still accurately predict the targeted field when the fields are all working together.

Conclusion

Every year in the US, commuters on average spend 54 hours a year wasted in traffic, as seen in a report from Texas A&M Transportation Institute. That time is the average, but several larger cities such

as Boston, New York City, and San Francisco have even more time wasted- as much as 103 hours per year (Willingham, 2019). Some traffic can come from congestion, but traffic also comes from car accidents. Traffic has only been increasing year by year, up until 2020. A recent study found that traffic delays have fallen 50% in major cities due to the Covid-19 pandemic. The study found that the average American driver only spent 26 hours in traffic in 2020 (Autonews.com, 2021). With less cars being on the road from the pandemic, one would assume that there were fewer accidents in 2020. However, it seems that drivers were more aggressive with less cars on the road. The monthly mileage death rate jumped to 26.1% in July 2020, with an increase of 11% over July 2019 for number of motor vehicle deaths. The deaths in 2020 were up 2%, despite it going down year after year in previous years (Injurylawer.com, 2020). These more deadly car accidents are big contributing factors towards the higher traffic times during Covid-19 compared to prior to Covid-19. When looking at the average traffic impact from a car accident in the datasets, the Covid-19 group had the highest average time of 3 hours and 10 minutes compared to the non Covid-19 group average time of 2 hours and 42 minutes. When it comes to the traffic impact or the Severity field in the data, the Covid-19 group had a lower average of 2.135 while the non Covid-19 group had a higher average of 2.358. This indicates there may have been some accidents that may have been mislabeled or the traffic impact from the accident was greater than anticipated, as the severity should correlate very highly with the time of the accident, at least in theory. With Covid-19 causing more reckless driving, hopefully as things go back to normal, including driving patterns, people will drive slower and cause the traffic to go back down to pre Covid-19 times. Though this analysis was eye opening, it is heartbreaking to hear about more accidents and more deaths caused from dangerous driving, even when there were fewer cars on the road.

References

Driverknowledge.com. (2021). Car Accident Statistics in the U.S. Retrieved from https://www.driverknowledge.com/car-accident-statistics/

Miller, Blincoe, L. J., Miller, T. R., Zaloshnja, E., & Lawrence, B. A. (2015, May). The economic and societal impact of motor vehicle crashes, 2010. (Revised) (Report No. DOT HS 812 013). Washington, DC: National Highway Traffic Safety Administration.

Wagner, I. (2020). Road accidents in the United States - Statistics & Facts. Retrieved from https://www.statista.com/topics/3708/road-accidents-in-the-us/#dossierSummary

Willingham, A. (2019). Commuters waste an average of 54 hours a year stalled in traffic, study says. Retrieved from https://www.cnn.com/2019/08/22/us/traffic-commute-gridlock-transportation-studytrnd/index.html#:~:text=In%20the%20report%20from%20the,two%20and%20a%20half%20days.

Autonews.com. (2021). Average time spent In U.S. traffic jams cut by 73 hours during pandemic, study finds. Retrieved from https://www.autonews.com/mobility-report-newsletter/average-time-spent-us-traffic-jams-cut-73-hours-during-pandemic-study

Injurylawer.com. (2020). Has the COVID-19 pandemic reduced the number of car accident fatalities? Retrieved from https://injurylawyer.com/car-accidents/has-the-covid-19-pandemic-reduced-the-number-of-car-accident-fatalities/

Apnews.com. (2021). Risky driving: US traffic deaths up despite virus lockdowns. Retrieved from https://apnews.com/article/pandemics-health-traffic-coronavirus-pandemic-799e455b73902b1638cc2ac46a172972

Apnews.com. (2021). US traffic deaths spike even as pandemic cuts miles traveled. Retrieved from https://apnews.com/article/pandemics-us-news-traffic-coronavirus-pandemic-307ab70e3336288bf38df73e9dba348e

Teletracnavman.com. (2020). The Impact of the COVID-19 Crisis on Driving Behaviors. Retrieved from https://www.teletracnavman.com/resources/resource-library/infographics/covid-19-crisis-on-driving-behavior-infographic

Pichon, A. (2020). Staying Safe on the Roads in the Time of Covid-19. Retrieved from https://blogs.lexisnexis.com/insurance-insights/2020/12/covid-19-negative-impact-on-traffic-safety/