TRAFFIC ACCIDENTS
IN THE US

An analysis of traffic data, and building an ML model to predict the severity of traffic incidents.

Biplob Gauli
MSDS 696
Regis University

April 2022

Abstract

Since Covid-19, people have started working remotely across several industries and thus, broken the norm that only people working in the technology field can do so. However, now that things are getter better as more and more people are getting vaccinated, employees are slowly being asked to return to the office. Although there is a fair share of people who would enjoy moving back to their office, some people still feel reluctant and want to continue working remotely. One of the factors causing this reluctance is long commute times, which is often exacerbated by traffic accidents.

This paper delves into questions pertaining to traffic accidents, like what times and days of the week are worse than others, is there any weather or feature that can predict the severity of accidents, and what traffic measures are associated with less severe accidents.

PROBLEM DESCRIPTION

PER ASIRT (ASIRT, 2022), road crashes are the leading cause of death in the U.S. for people aged 1-54, causing 38,000 deaths annually, and 4.4 million serious injuries. In addition to that, accidents cause traffic congestions and people spend more time on the road than needed.

DATA

The US traffic dataset includes details on accidents collected in over 49 states dating from 2016 to 2021. The data was collected from various entities like traffic camera, sensors, law enforcement agencies, and state departments. There are close to 3 million rows and 47 different features including date, time, location, weather conditions, road features and traffic features near the place of the accident (Moosavi, 2021).

RESEARCH TOPICS

- Does visibility impact the severity of the accidents?
- What hours of the day and week has the most and least number of accidents?
- Is traffic light better at reducing the severity compared to a roundabout?

METHODOLOGY

Here is a rough outline of the process for this project:

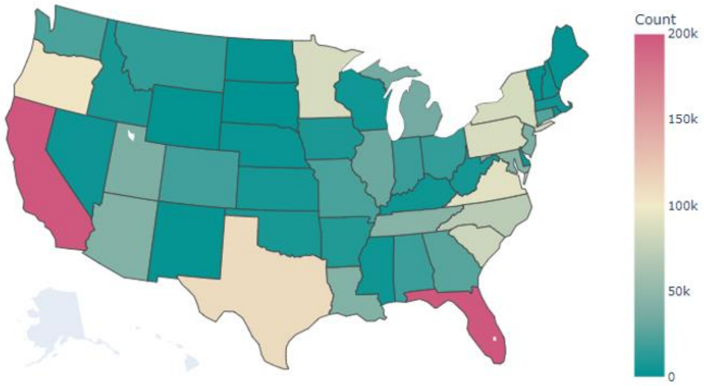
- Data extraction, cleaning, and preprocessing
- Explanatory data analysis
- Correlation analysis and feature selection
- Answers to the research question
- Machine learning models to predict the severity of an accident
- Compare KNN, SVM, Random Forest and Neural Network and provide a recommendation

DATA CLEANSING

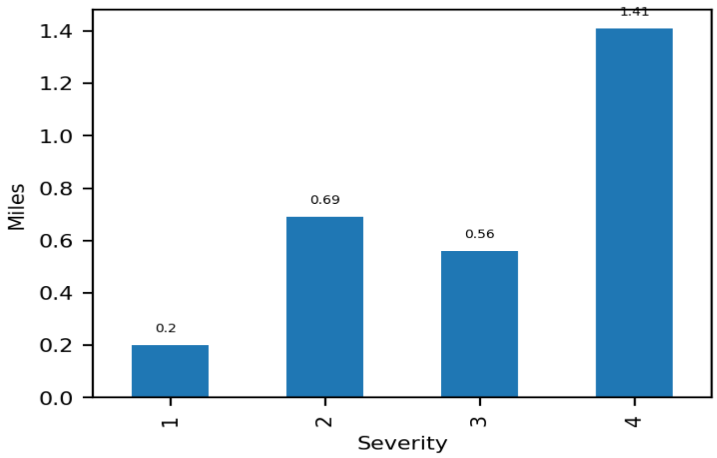
The dataset has 47 columns with close to 3 million rows that amounted to over a gigabyte in file size. To reduce the complexity and runtime of this project, 22 columns that were irrelevant/redundant were removed. Some of these columns indicated sunset or sun rise, type of twilight, direction of wind, and air pressure, to name a few. Records with missing values and duplicate entries were also removed. This new dataset is less than half the original size and was used as the new baseline for further analysis.

EXPLORATORY DATA ANALYSIS

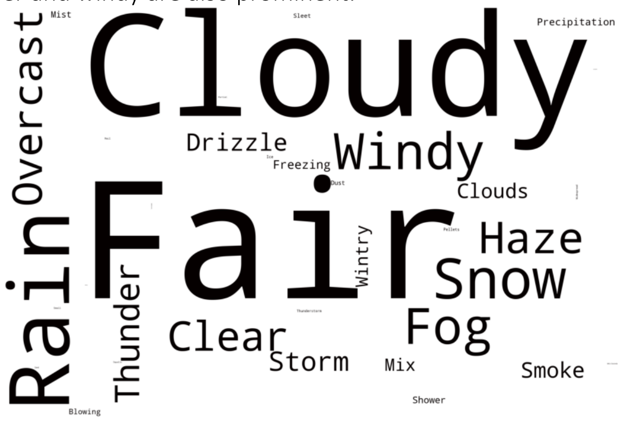
Based on the new dataset we have, we will do some exploratory data analysis. The following diagram shows the volume of accidents across different states. California and Florida have the maximum number of accidents. Texas is next, followed by quite a few states on the east coast.



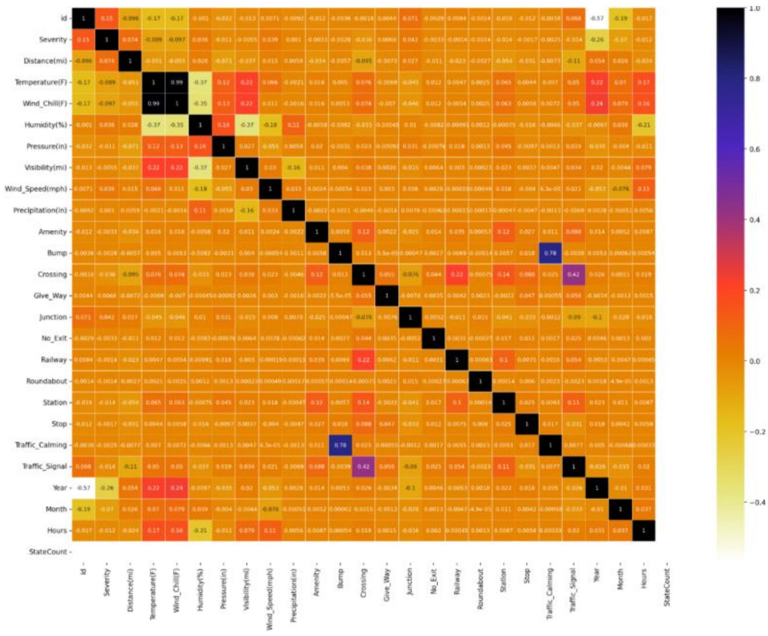
The following chart shows the four categories of accidents based on the severity, 1 being the lowest and 4 being the highest. Severity 4 produces significantly longer traffic at 1.41 miles, while the rest are below a mile.



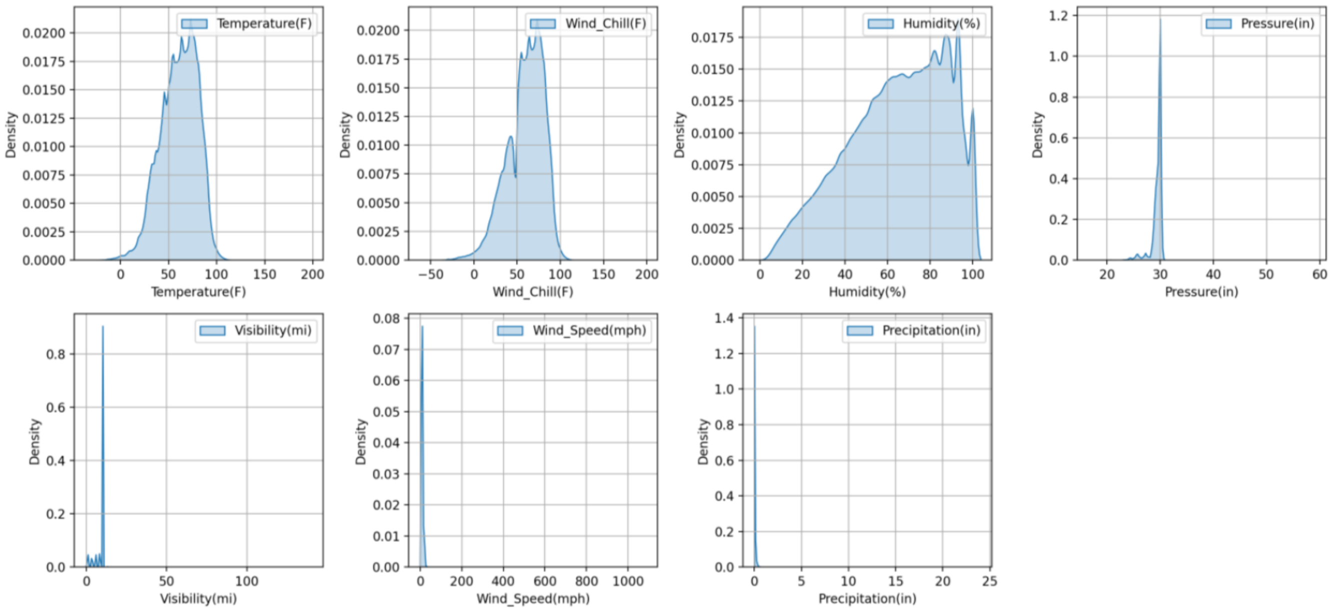
Below is the word cloud of the most common weather conditions during accidents. Majority of the accidents happen during fair or cloudy conditions. Rain, snow, thunder and windy are also prominent.



Similarly, this correlation plot depicts the relationship between the different features we have on our baseline dataset. Although there isn't a strong correlation between severity and the rest of the factors, we do see some interesting observations. For instance, higher humidity is linked with lower visibility and lower windspeed.



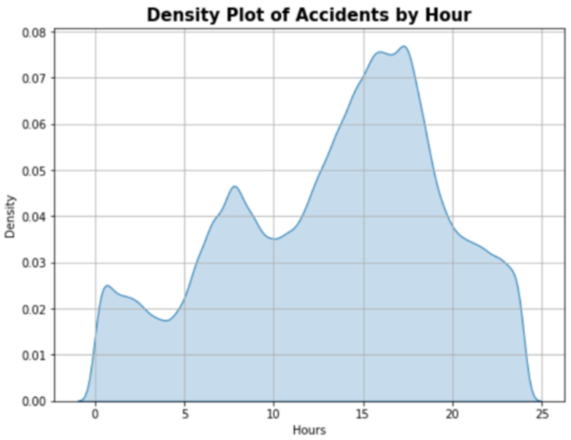
Similarly, here are the details on the different weather conditions during the time of accident like the temperature, windchill humidity levels, pressure, visibility, wind speed and precipitation.

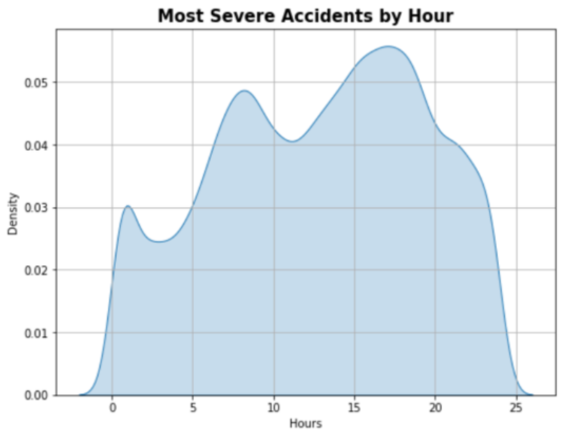


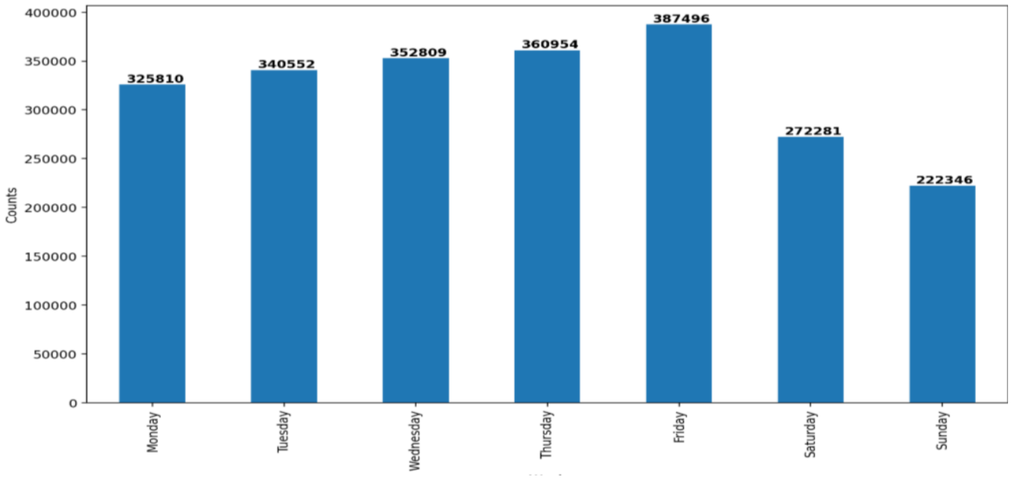
October to December have the highest number of accidents. This can be because of winter conditions as well as holiday travel followed by end of school. However, it is interesting to see that this number drastically drops down as we move to January. The summer months seems very mediocre and even below average.

What hours of the day and week has the most and least number of accidents?

Based on the charts below, most accidents and most severe accidents happen during 3 to 6 PM. It gets progressively worse from Monday to Friday, and below average on weekends.







Is traffic light better at reducing the severity compared to a roundabout?

- H0: both have same severity
- H1: they have different severity
- Based on the t-test and the p value of 0.168 we fail to reject the null hypothesis.
- Can't conclude traffic light is better than roundabout or vice versa.

Does visibility impact the severity of the accidents?

- H0: all classes of severity have same visibility
- H1: different severity has different visibility
- Based on the hypothesis testing, the P value is less than 0.05, so we reject the null hypothesis.
- We can infer that visibility does impact the severity of an accident

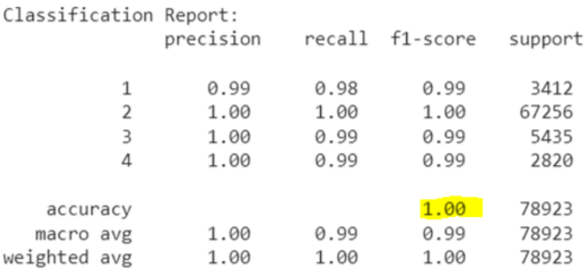
Machine Learning

We will build three different machine learning models that will predict the severity of the accidents. First, we shuffle the data and convert all categorical and boolean columns into integer columns using one-hot encoding. We also use principal component analysis (PCA) to limit the number of features we input on our models to simplify our model and reduce the run-time. From over 85+ features, we reduced it to 25. Finally, we will split the data into test and train at an 80:20 ratio. Along with that, we will also do hyperparameter tuning to get the best results.

Random Forest

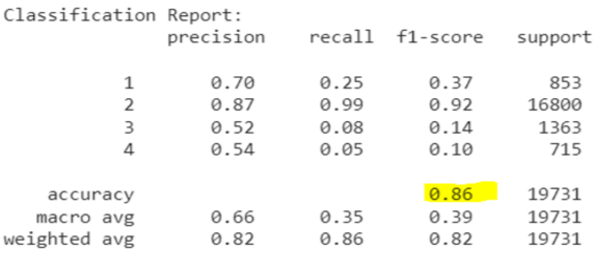
The f1 score for the overall model is 100% for our training data, which is good. The final score for our test data is 86%. This is slightly lower than the training model and this is because of lower scores on level 1,3 and 4, despite a high score of 92% on level 2.

Train



```
Classification Report:
              precision    recall  f1-score   support

           1       0.99      0.98      0.99      3412
           2       1.00      1.00      1.00     67256
           3       1.00      0.99      0.99      5435
           4       1.00      0.99      0.99      2820

    accuracy                           1.00     78923
   macro avg       1.00      0.99      0.99     78923
weighted avg       1.00      1.00      1.00     78923
```

Test

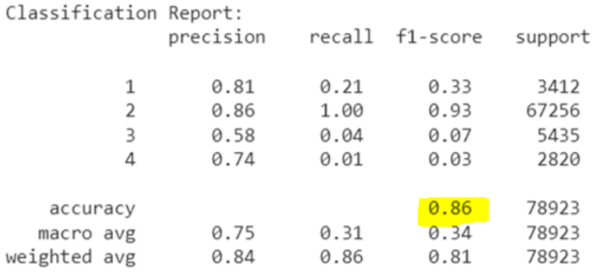


```
Classification Report:
              precision    recall  f1-score   support

           1       0.70      0.25      0.37       853
           2       0.87      0.99      0.92     16800
           3       0.52      0.08      0.14      1363
           4       0.54      0.05      0.10       715

    accuracy                           0.86     19731
   macro avg       0.66      0.35      0.39     19731
weighted avg       0.82      0.86      0.82     19731
```

K Nearest Neighbors

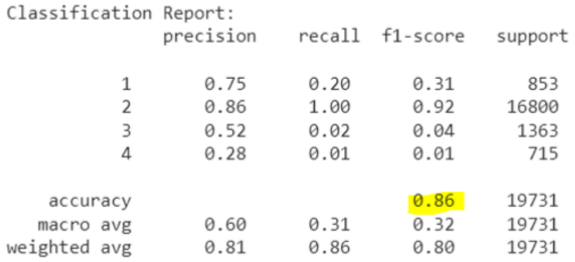
The f1 score for this method dropped down to 86% for our training model, as well as the test model. This is because unlike the random forest model, only level 2 has good scores.

Train



```
Classification Report:
              precision    recall  f1-score   support

           1       0.81      0.21      0.33      3412
           2       0.86      1.00      0.93     67256
           3       0.58      0.04      0.07      5435
           4       0.74      0.01      0.03      2820

    accuracy                           0.86     78923
   macro avg       0.75      0.31      0.34     78923
weighted avg       0.84      0.86      0.81     78923
```

Test

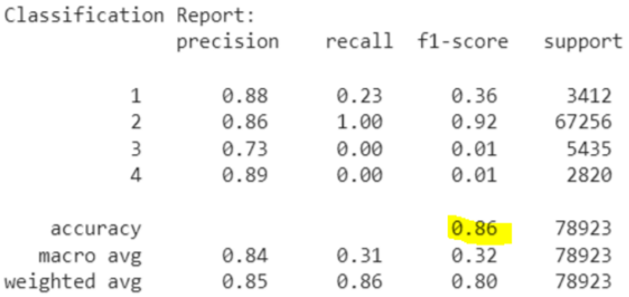


```
Classification Report:
              precision    recall  f1-score   support

           1       0.75      0.20      0.31       853
           2       0.86      1.00      0.92     16800
           3       0.52      0.02      0.04      1363
           4       0.28      0.01      0.01       715

    accuracy                           0.86     19731
   macro avg       0.60      0.31      0.32     19731
weighted avg       0.81      0.86      0.80     19731
```

Neural Network

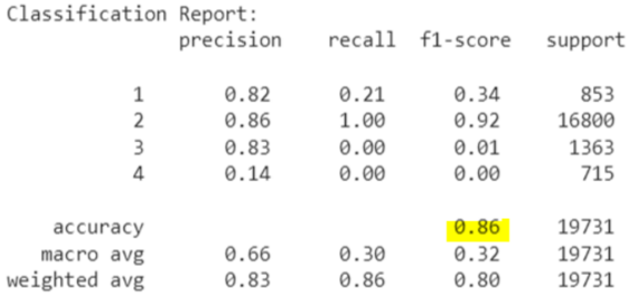
The result for this method is very comparable to the KNN method in both test and train models. However, this model took more than four times to run compared to the two methods.

Train



```
Classification Report:
              precision    recall  f1-score   support

           1       0.88      0.23      0.36      3412
           2       0.86      1.00      0.92     67256
           3       0.73      0.00      0.01      5435
           4       0.89      0.00      0.01      2820

    accuracy                           0.86     78923
   macro avg       0.84      0.31      0.32     78923
weighted avg       0.85      0.86      0.80     78923
```

Test



```
Classification Report:
              precision    recall  f1-score   support

           1       0.82      0.21      0.34       853
           2       0.86      1.00      0.92     16800
           3       0.83      0.00      0.01      1363
           4       0.14      0.00      0.00       715

    accuracy                           0.86     19731
   macro avg       0.66      0.30      0.32     19731
weighted avg       0.83      0.86      0.80     19731
```

Best Machine Learning Model

Determination for the best machine learning model is based on the following criteria.

- Accuracy/F1 Score
- Ease of implementation
- Runtime

Although All 3 methods gave a similar accuracy of ~86% for our testing, random forest is the winner by a slim margin. Although the scores for levels 1, 3 and 4 are not acceptable, it is much higher compared to the KNN and Neural Network method. It was also simple to implement and took less than 30 minutes to run.

Conclusion

There were multiple outcomes for this project. First, we found out that it is best to avoid traffic from 3-6 PM for least impact with accidents. Similarly, the traffic gets progressively worse from Monday to Friday, so the first few days of the week are ideal for commute days.

We then found out that traffic light and roundabout are equally effective in controlling the severity of accidents based on hypothesis testing. Similarly, visibility does have an impact on the severity of accident. Although there isn't a linear correlation between visibility and severity of accident, mean severity is highest for class 2.

References

ASIRT. (2022, April 21). *Road safety facts*. Association for Safe International Road Travel. Retrieved April 27, 2022, from https://www.asirt.org/safe-travel/road-safety-facts/#:~:text=More%20than%2038%2C000%20people%20die,for%20people%20aged%201%2D54.

Moosavi, S. (2021, January 1). *US-accidents: A countrywide traffic accident dataset*. US-Accidents: A Countrywide Traffic Accident Dataset. Retrieved April 27, 2022, from https://smoosavi.org/datasets/us_accidents