Forecasting US Road Accidents

Aaron Xiao, Kiran Sidhu, Paul Flemming, Rozi Hagos, Tara Singh

Executive Summary

Trudeau Consulting has been contracted for the development of urban plans to build future interstate highways. The main objective is to analyze US accident data over the last few years for trends. In order to achieve this objective, we needed to develop forecasts for the next ten years by state and an aggregate forecast for the US. We also utilized the Arima model to create the forecasts at the national level and by state. The results show that there is an overall increase in accidents in the United States as a country and the majority of the individual states show an increase as well.

Business Objectives and Problem Formulation

Our firm was hired by Biden Engineering, a consulting firm with a government contract. They have been contracted for the development of urban plans to build future interstate highways. They have engaged us to analyze US accident data over the last few years for trends. Our overall goal was to develop forecasts for the next ten years by state and an aggregate forecast for the US.

Methodology

We will outline our preliminary analysis, the methods used to manipulate the data and provide a descriptive analysis. The Arima model was employed to develop the forecasts which will be provided at the national level and segmented by state.

Preliminary Analysis

This dataset contains car accident data for the 49 continental states of the US. It covers a period from February 2016 to June 2020. It was sourced from MapQuest and Bing, two APIs that provided streaming traffic incident (or event) data. It includes 3.5 million records with 49 features.

We discovered several interesting observations from running some basic descriptive plots in Python. Seventy percent of accidents didn't happen near a traffic object. Of the 30% of the remaining accidents, 44.7% were near a traffic signal, followed by 20.4% near a Junction and 19.7% near a Crossing. In terms of the severity level, 67.5% of accidents had a severity level of 2 and 28.4% were at severity level 3. This covers 96% of the accidents in the dataset. Most of the accidents take less than an hour's time to get resolved. However more than 15% of the accidents take 360 minutes to resolve.

Weekdays had more accidents than weekends, with peaks at 8:00am and 5:00pm. This would be during normal rush hour which is to be expected as most people are travelling to and from work or school at that time. On the weekend, the peak is around 1:00pm which is when people head out for errands and social activities.

Data Manipulation

To clean up the data, certain transformations were required. Start time and end time were changed to datetime variables from character variables. Logical features (Boolean) were changed to dummy variables. Records that had a negative accident duration or missing values were removed. Outliers were cleaned for distance, temperature, pressure, precipitation, visibility, and wind speed.

To prepare the data for the Arima model, records were grouped by month to create one data point for each month. There were seven states (Colorado, Montana, North Dakota, New Mexico, South Dakota, Utah and Wyoming) that didn't have enough records to be used in forecasting, so they were removed from the model.

Descriptive Analysis

In the Appendix (Figure 1), we can see the top 15 states by accident count, with California having the highest count, followed by Texas, and then Florida. Out of the remaining list, South Carolina has a disproportionate number of accidents as its population compared to the other 15 states. Oregon, Minnesota and Tennessee are also not in the top 15 for population but appear in the top state list for accident counts (Reference 1). This tells us that population is a factor that should be considered with the data.

We also reviewed the Time Zones (Figure 2). When we sort the accident percentages by time zones, the order is Eastern (42.3%), Pacific (28.1%), Central (23.9%), Mountain (5.7%). When done by population it's Eastern (47.6%), Central (29.1%), Pacific (16.6%), Mountain (6.7%). This shows us that Pacific Time Zone has a disproportionate number of accidents to their population (Reference 2).

Model Building and Evaluation

In order to develop annual forecasts (aggregate US and by states) from 2021-30 (ten years) and analyze forecasts by segments, we used an autoregressive integrated moving average (ARIMA) model. This is a statistical analysis model that uses time series data to better understand the data set under review or to predict future trends. ARIMA uses a number of lagged observations of time series to forecast observations; which makes it a suitable model to identify the trend for the next ten years. For this data set a "grid search" was used to find the optimal set of parameters that yields the best performance for the ARIMA model. We fit the ARIMA model and plotted the results.

A plot was produced to show the national trend and the trend by state. We plotted California separately and we grouped the states that had declining trends. Finally, we grouped together states with similar counts in groups of nine for the remaining plots.

Insights and Summarizing Results

We plotted time series decompositions for the US at a national level (Figure 3). This will allow you to decompose your data into seasonal trends and residual components. It also shows the accident count for the dataset. The overall trend for accidents is increasing, and the seasonality shows that most accidents occur in October and the lowest amount happen in July. The national trend started dropping in 2017, hit the lowest point in 2018 and started increasing again in 2019.

For the year 2020, we added the 2020 forecast data to the observed data as we only had the data for the first six months of the year (Figure 4). The confidence level for the US forecast is 95% which means that there is a 95% chance that the future values will fall in the shaded area. The confidence interval is a range of projected values where we expect our future results to land. In Figure 5, we changed the confidence level to 50%, which means that there is a 50% chance that the future values will fall in the shaded area.

We then conducted forecasts at the state level. For the following charts (Figures 6-11), the confidence level has been set at 50%, which means that there is a 50% chance that the future values are in the shaded area. Figure 6, shows the accident forecast for Texas through Illinois for the next ten years. As noted previously, there were four states with low populations, but high accidents (South Carolina, Minnesota, Oregon and Tennessee). We can see South Carolina has the highest forecasted amount of

accidents in its chart, putting it ahead of both Texas and Florida which have much higher population sizes and current amount of accidents. South Carolina's forecasted accidents amount is followed by Minnesota, one of the other states that have lower populations but still appear in the highest amount of accidents. The state with the third highest forecasted accidents is Oregon, another low population but high accident state. Texas and Florida are forecasted to rise in accidents. Florida is still expected to have a lower accident total than North Carolina which is below it at this point in time.

The accident forecast for Georgia through New Jersey for the next ten years is shown in Figure 7. The last state of the four with low populations but high accidents, Tennessee has a higher forecast than more populated states like Georgia and New Jersey. Another interesting point in this chart is that Alabama is forecasted to have more accidents than Tennessee, despite them having fewer at this point. Accidents in Oklahoma are expected to rise above Georgia and Arizona within the next ten years.

The accident forecast for Louisiana through Connecticut until 2030 are shown in Figure 8. We can see that Pennsylvania's forecast skyrockets and Louisiana, Missouri and Connecticut's forecasts are projected to fall. We also noticed that the trend for Indiana went down from 2018 to 2019, rose a bit in 2020, but is then expected to also skyrocket over the next ten years.

Figure 9, shows the accident forecast for Kentucky through Vermont until 2030. We can see that accidents increased significantly in Kentucky from 2017 through 2020. Kentucky's accidents are forecasted to drop over the next ten years but still has a much higher number of accidents than the other states. Arkansas and New Hampshire have a steeper increase rate than the others in this chart, but accidents Delaware are expected to increase above New Hampshire and Arkansas in 2030.

Since California has the largest amount of accidents, we created a separate forecast for it (Figure 10). There is a steady rise in accidents forecasted, but not alarmingly so. The slope will rise over the next few years, but not as steeply as in the past.

Figure 11, contains the forecast for Michigan through Nevada. After a peak in 2017, the total accidents have been rapidly dropping in Michigan. The trend for the future shows a continued rapid decline. The other states in this graph are Nebraska, Rhode Island, Mississippi, and Nevada. We can see that the

forecast drops below zero for all states in this graph. This is because there are declining trends for the last two years for all these states, this pushes the trendline into negative values. We can see by looking at the confidence interval shading, that there is still a possibility for the trend to move upwards for these states.

Recommendations

Effective urban planning has a direct impact on road safety and efficiency. As seen in the insights, the overall accident forecast for the US was trending upwards. In addition, 37 of the 49 states in the dataset were also on an upward trend for number of accidents. Our firm can provide a list of all states with an upward trend, this will be sorted by highest yearly average. Our recommendation to Biden Engineering is to leverage this list, start with the highest yearly average to target your urban planning. Value added benefits include cost savings because the resources will be used efficiently, saving time and increasing impact.

We suggest adding in population data to your data set to calculate more stats on the accidents per capita. This will be another strategic tool to target the most impactful projects. Including bridge data in the data set would be helpful to identify potential locations for future interstate highways. We recommend talking to the API developers to ensure that more traffic object data is included in the future (bridges, traffic signals etc.). This will help you cross reference data to see what kind of roads the accidents take place on (e.g. freeway). This can be done with more API support from different developers (Google Maps etc.) or enhancing the existing APIs used.

Conclusion

The main objective was to analyze US accident data for trends, to develop forecasts for the next ten years by state and an aggregate forecast for the US. Arima was used and created the required forecasts. The results show that there is an overall increase in accidents in the United States as a country and the majority of the individual states show an increase as well. It is recommended to incorporate more data but focus on the list our firm provided to allow for better planning on the government contract.

Appendix

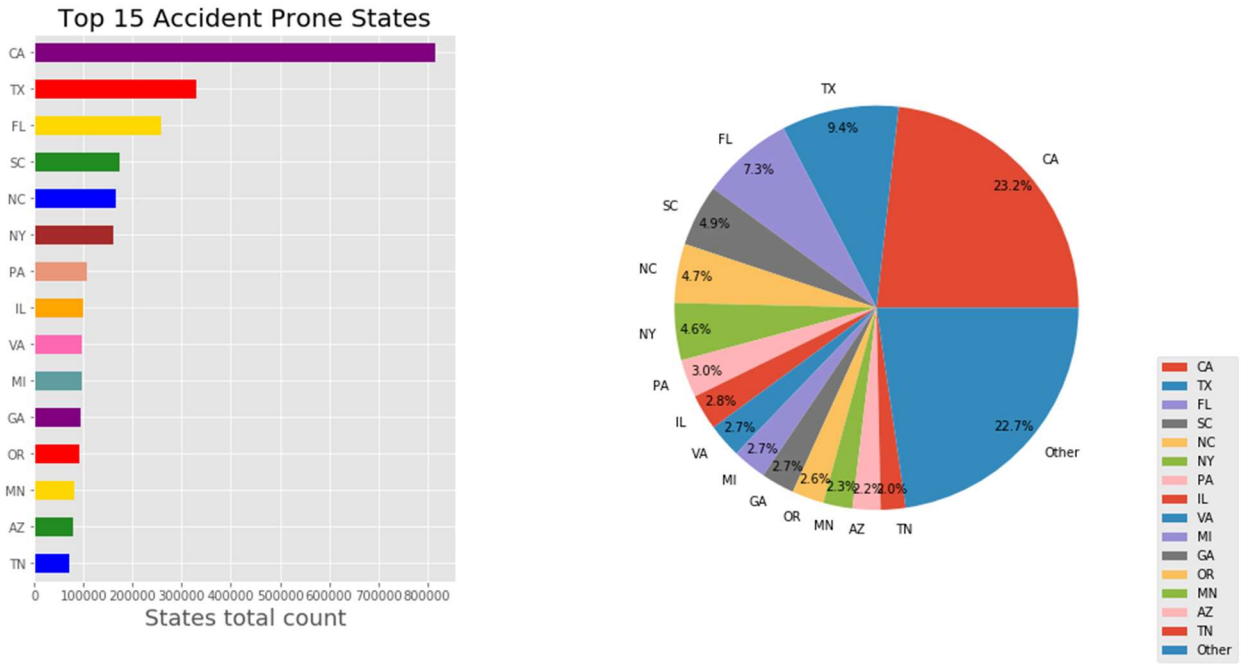

Figure 1: Top 15 States by Total Accident Count

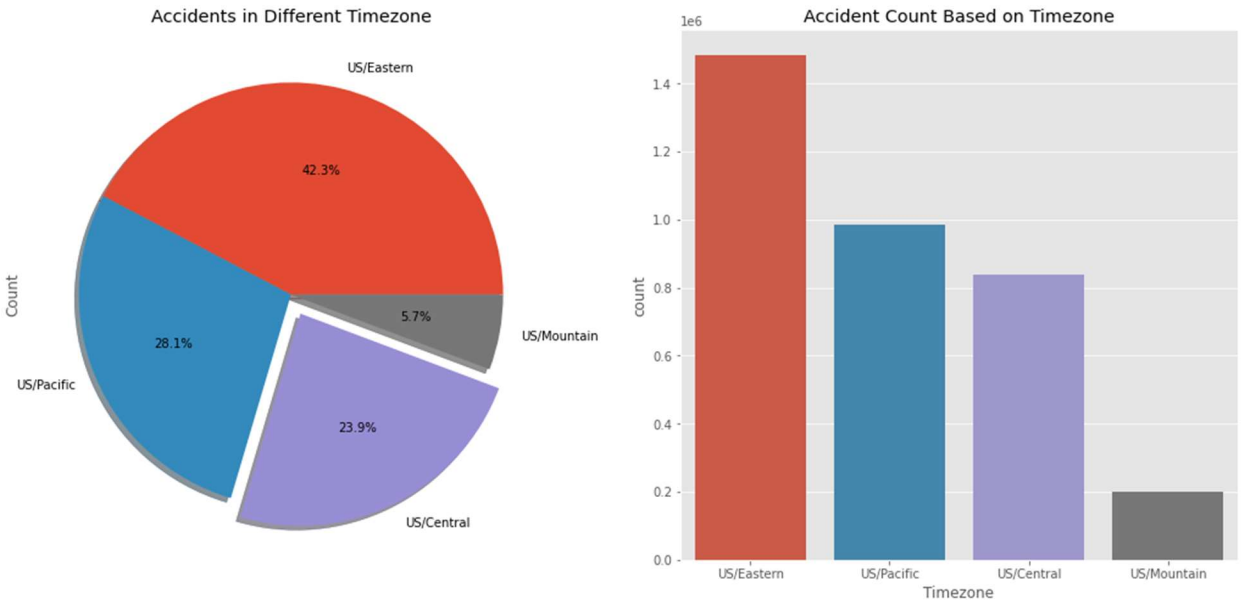

Figure 2: Accident Count Based on Timezone

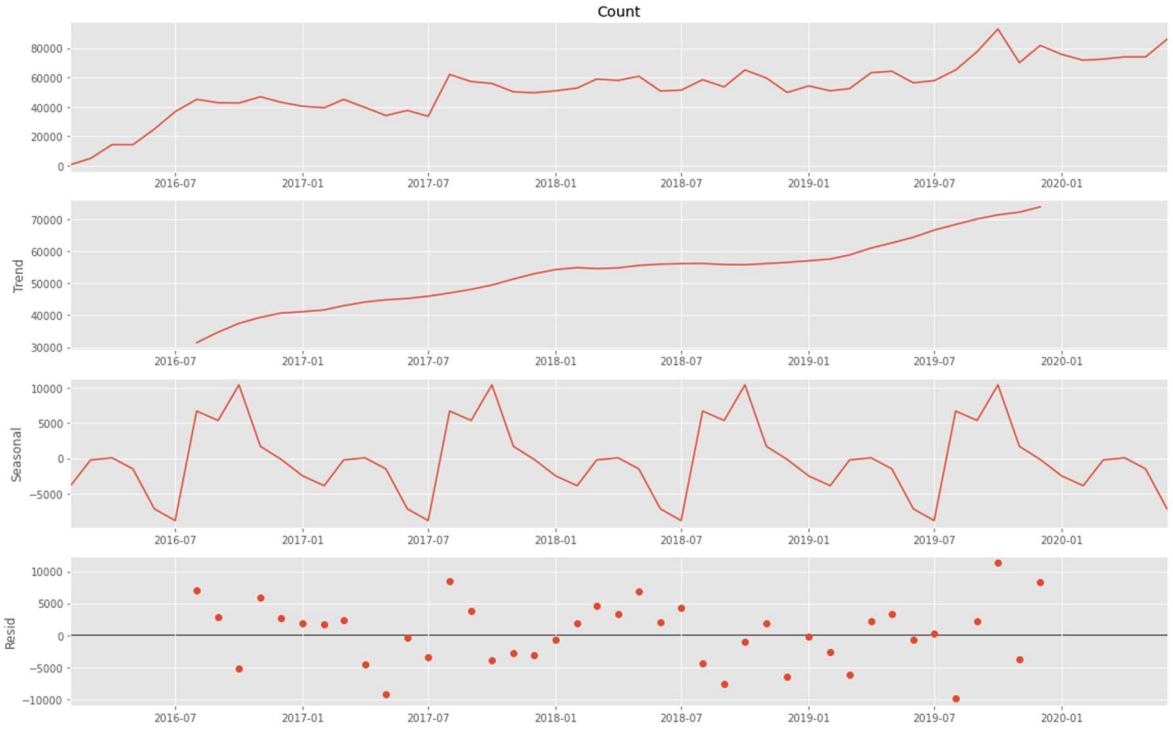

Figure 3: Seasonal Decomposition for US

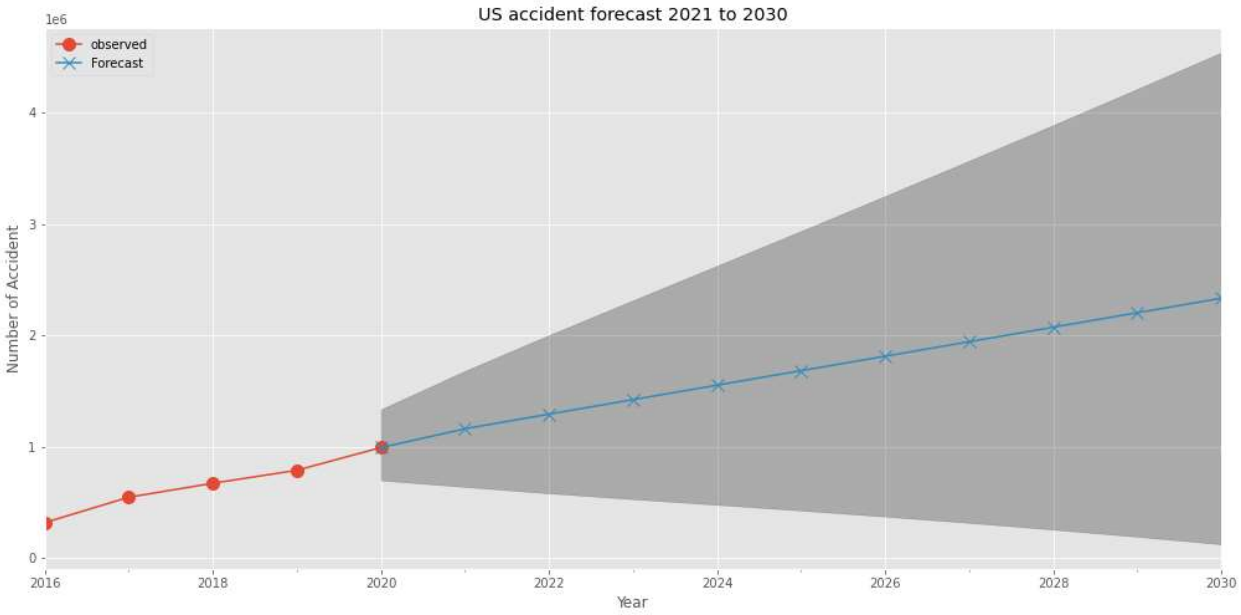

Figure 4: US Accident Forecast from 2021 to 2030 with a 95% Confidence Interval

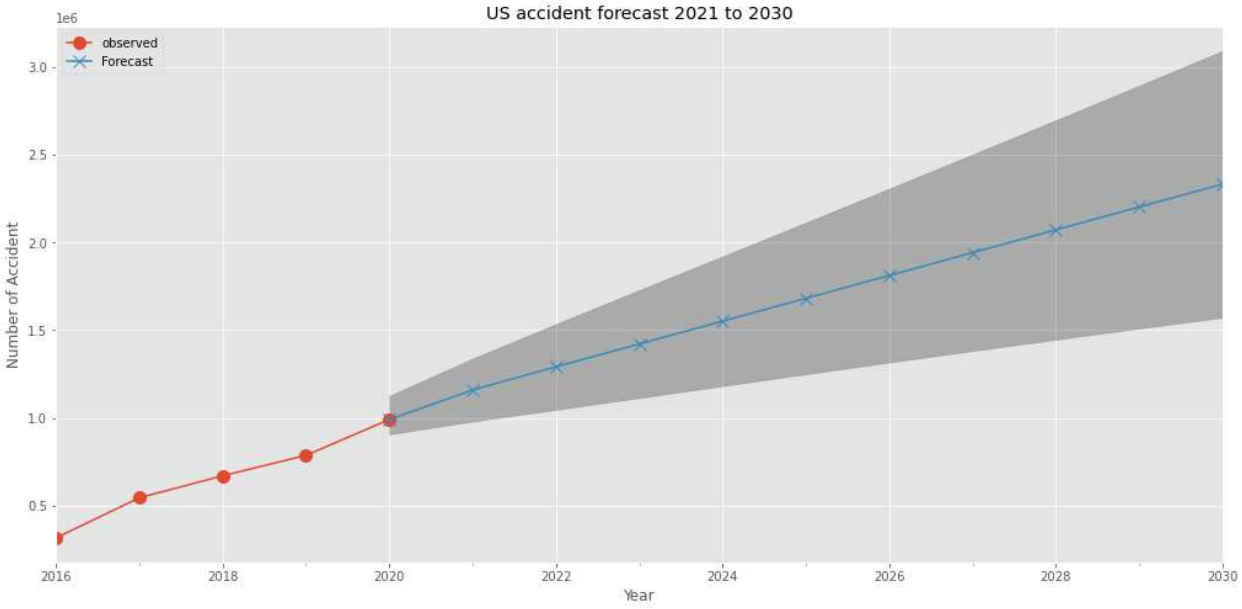

Figure 5: US Accident Forecast from 2021 to 2030 with a 50% Confidence Interval

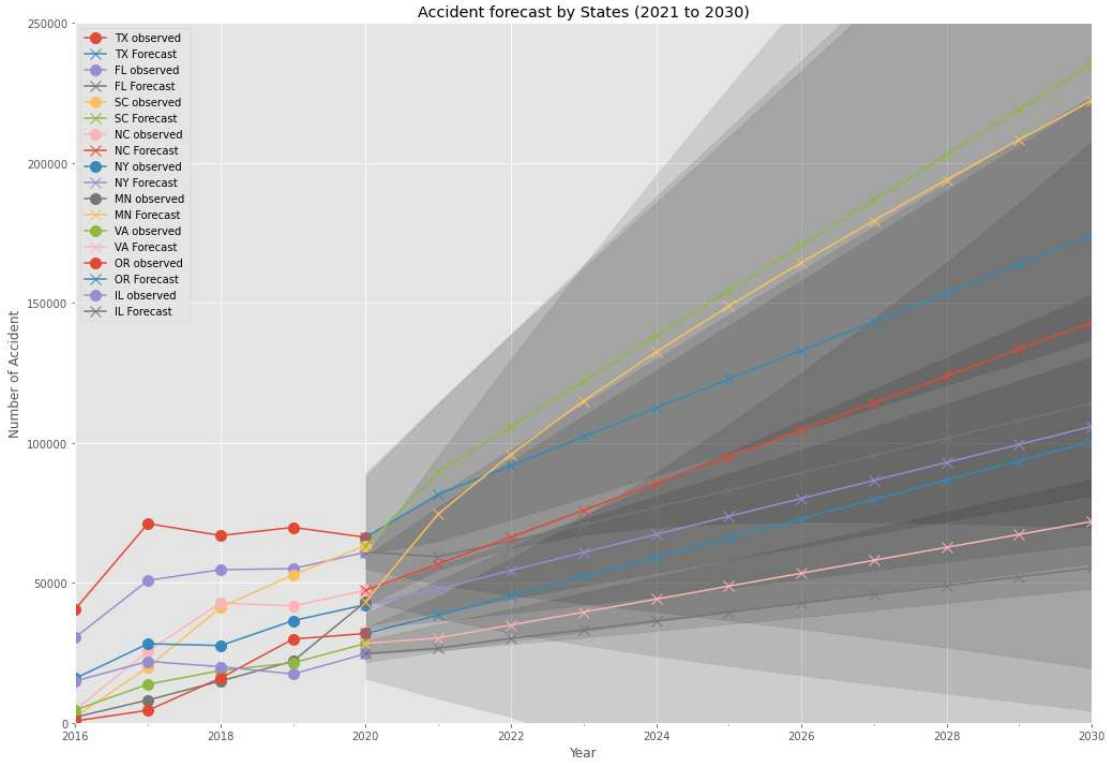

Figure 6: Accident forecast by States 2021 – 2030

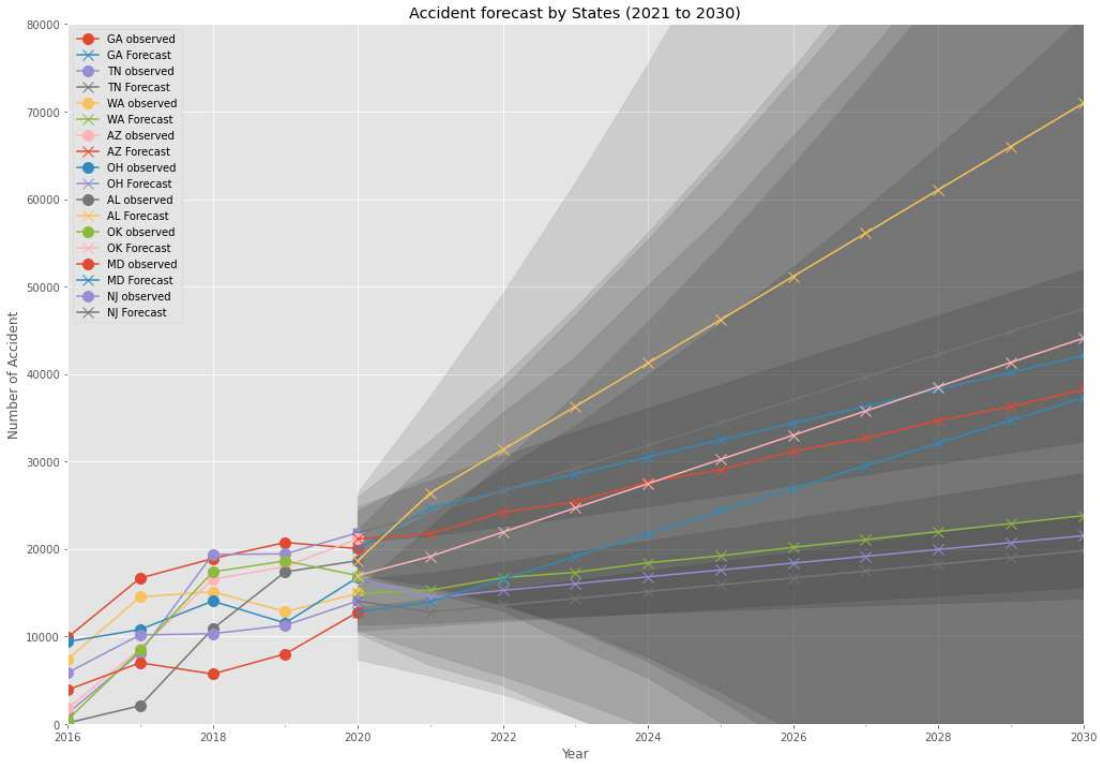

Figure 7: Accident forecast by States 2021 – 2030

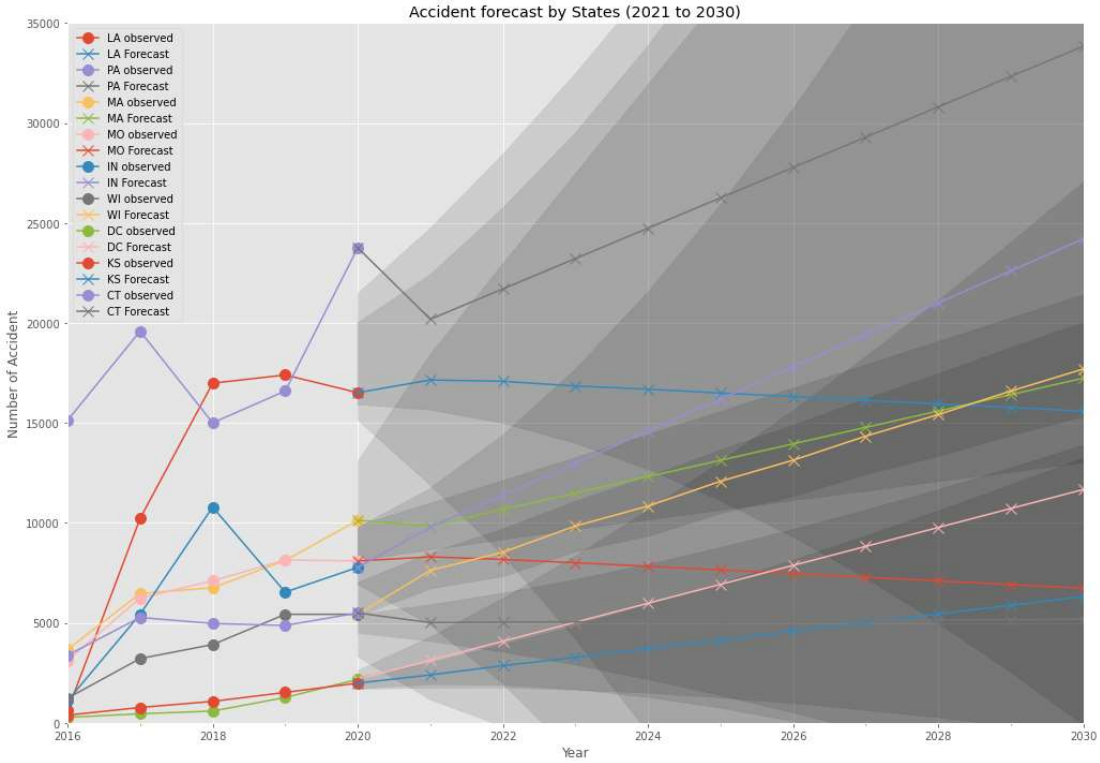

Figure 8: Accident forecast by States 2021 – 2030

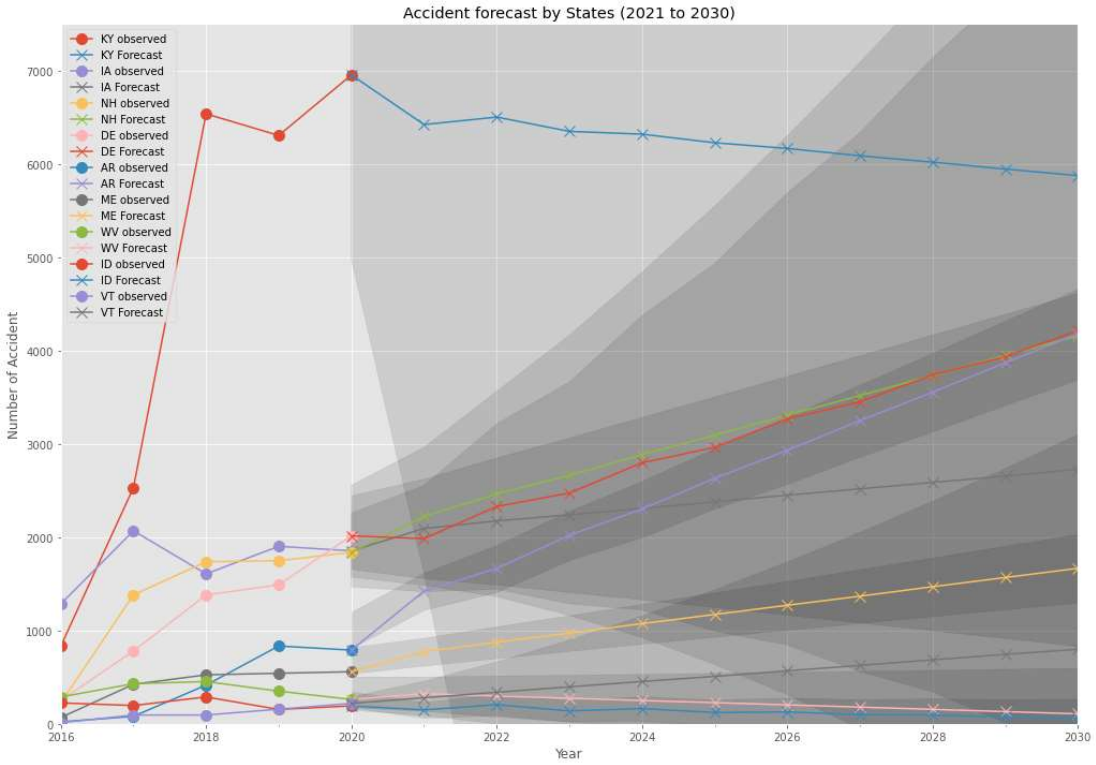

Figure 9: Accident forecast by States 2021 - 2030

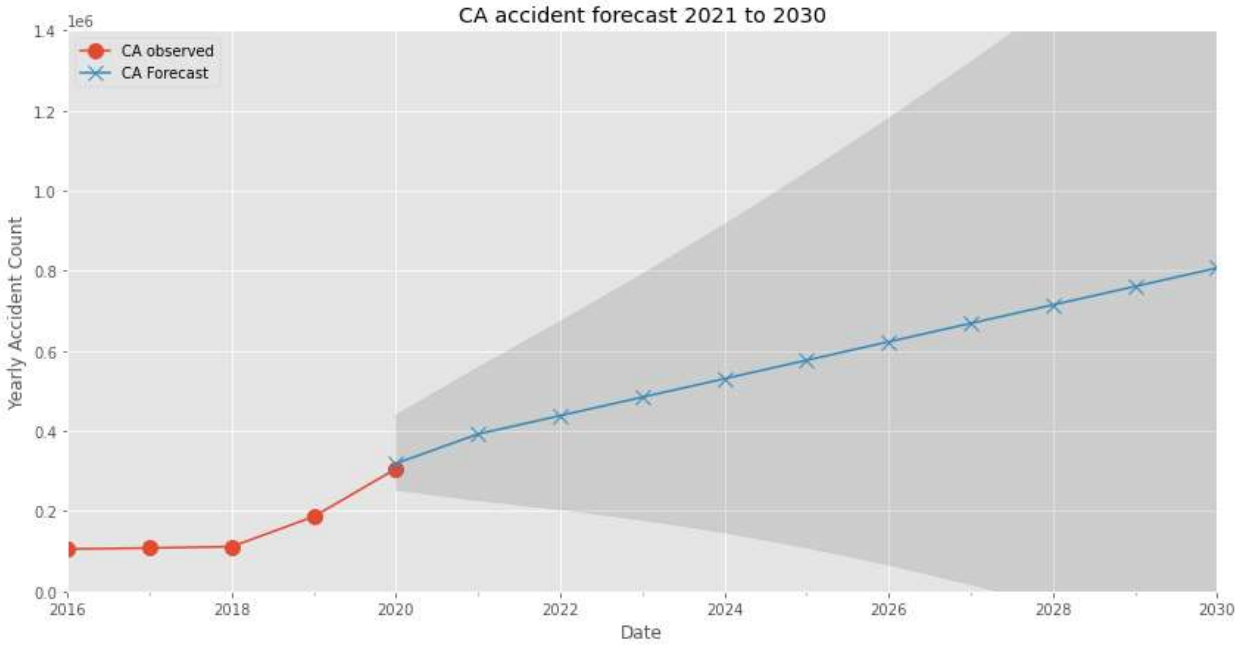

Figure 10: California Accident forecast - 2021 – 2030

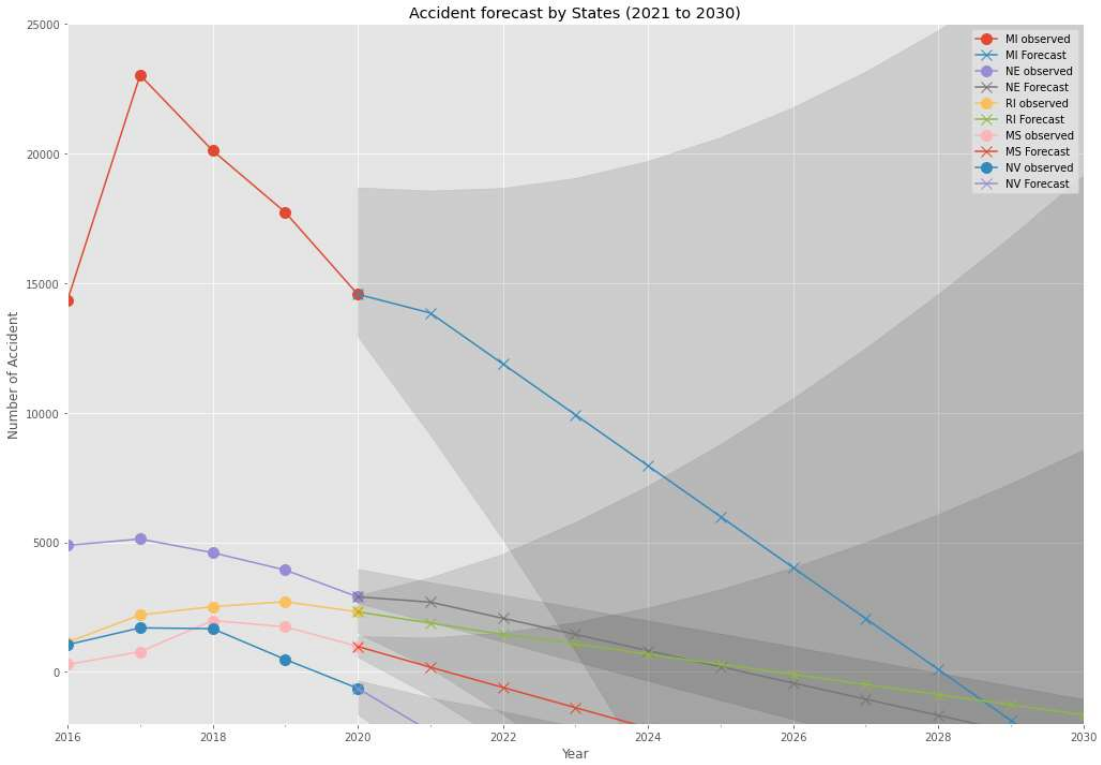

Figure 11: Accident forecast by States - 2021 - 2030

References

- 1.https://web.archive.org/web/20151223044815/http://www.census.gov/popest/data/national/totals/2015/index.html
- 2. https://www.rpsrelocation.com/blog/data-visualization/american-cities-by-time-zone/#:~:text=In%20terms%20of%20total%20population,Census%20estimates%20(via%20MetricMaps)
- 3. https://machinelearningmastery.com/calculate-feature-importance-with-python/
- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. "A Countrywide Traffic Accident Dataset.", 2019.
- 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.
- https://www.optimistdaily.com/2020/01/effective-urban-planning-doesnt-just-boost-efficiency-it-can-save-lives/