Predicting Accident Severity in the US

15.071 Project Report

Madeleine Golison, Cindy Heredia, Megha Maran, Adrianna Wojtyna

I. Introduction

Car accidents are a main cause of traffic slowdowns on roads. This leads us to the question - can we predict severity of accidents based on factors such as weather conditions, road type, and location? In this report, we create predictive models to anticipate the severity of traffic slowdowns due to accidents before they occur. This would allow police and highway patrol to be better prepared to manage traffic following an accident given certain conditions, significantly reducing the accident's impact on traffic. Furthermore, we analyze how COVID-19 affected traffic severity country-wide and by state, identifying if state lockdown policies during the pandemic had any effect on accidents. This would allow us to create a model for the NHTSA in which traffic controllers can input current conditions to stay ahead of potential accidents and traffic slowdowns. In summary, after consideration of different levels of granularity of model development as well as different classification techniques, we believe that XGBoost per-state models provide high average accuracy over all the states (86.6%) and at the same time, it's feasible to create a model within reasonable timespan. However, if only accuracies were to be taken into consideration, our recommendation would be to use either XGBoost for the whole US (0.95) or Random Forest models trained on the whole available data set, as they provide the highest accuracy out of all the considered models and segmentation (0.9556).

II. Data Analysis and Preprocessing

The data set used as part of our analysis comes from the Kaggle dataset US Accidents (2016-2021). The data considered in our analysis comes from A Countrywide Traffic Accident Dataset, which contains 47 variables and about 2.8M observations. We began with exploratory data analysis and visualization to understand the nature of our data. These visualizations can be seen in the Appendix. We can see that a large majority of the accidents fall under level 2 severity, and most of the accident data come from California. Additionally, we have more data from recent years (2020-2021).

Before building models, we had to clean and preprocess the original dataset. Firstly, we transformed multiple features into factors such that they are not treated as a set of characters or numbers. Next, we transformed timestamp into multiple features that correspond to hour, day, and month of a particular observation. In addition, we discovered multiple missing values. For example, we observed that for precipitation, wind speed and wind chill there are multiple rows with missing values. For all of these observations, corresponding weather conditions described good weather. Thus, we assumed that missing data for the corresponding values were the result of lack of precipitation and lack of wind, and we replaced missing values with zeros. Finally, we produced correlation matrices between all pairs of numerical variables (Exhibit 4). We produced a similar matrix between all pairs of categorical variables using Chi-Square test and Crammer V metrics (Exhibit 6). As a result of these we discovered multiple heavily correlated variables and decided to remove one of each of such correlated pairs. Start hour and start latitude are some examples of variables we decided to drop as a result of their high correlation with end hour and end latitude, respectively. As a result of this data preprocessing we reduced the number of considered variables from the initially available 47 to 41. After completing all preprocessing steps, we were left with 97% of the original sample. In order to validate this proportion, we also complete the same data cleaning process on the complete dataset, where the proportion of rows left after preprocessing was also close to 97%. After data cleaning and preprocessing, we divided the dataset into a training and testing set, with 70% of data allocated for training. Training and testing sets were created using createDataPartition in R in an attempt to balance the distributions of severity within the splits.

III. Models

A. Overview

After preprocessing, we ran our data through a multitude of models. Since severity that we were trying to predict is a categorical variable of the 4 levels, all the models were used for classification. Specifically, we used the following approaches to create predictive models: CART, Random Forest, and XGBoost. We analyzed the relationship between accident severity by state and by city to identify the strongest predictors. We did not use a linear model because a linear regression approach is not well suited for multi-class classification models. In our analysis, we tried multiple data segmentation to achieve the highest possible accuracy. We aggregated data using different levels of hierarchy: per region consisting of a set of states (for example Southwest, New England), per state, as well as the whole US. For each of these model types, we further segmented the data using 3 periods - the whole available time frame, 2016-2019 (pre-COVID) and 2020 (post-COVID). Given this data segmentation, we then trained and tested models using data from the particular segment, using our chosen training to test ratio (70%).

B. CART

We wanted to first incorporate CART into our analysis as the structure of the output improves the explainability; thus, it is more applicable in emergency situations, when there may not be access to the internet. Overall our CART model for the whole US did well on the whole dataset (0.9543 accuracy). We attribute this fact to the 10-fold cross validation we used to select the best tree. For most cases, models in each particular category provided the highest accuracy when the whole available time span was used for training. For example, as it's demonstrated on Exhibit 11, the CART model for the whole US provides the highest accuracy of 95.4% when all preprocessed data points were used. In comparison, once we divided data into "pre-COVID" (2016-2019) and "post-COVID" (2020+) intervals, models created on these data segment showed significantly lower accuracy (0.81 and 0.91, respectively) We attribute this observation to the larger amount of data available for training available for the model for the whole time span

(2016+). We observed an interesting outlier to that rule: models for the New England region. Models trained and tested on data from 2020 provided significantly higher accuracy than the corresponding model trained on the whole available time frame for that region (0.6968 vs 0.7257). We are not certain what attributes to these observations, but we hypothesize that with COVID lockdown and reduced traffic, behaviors that caused most traffic might be more apparent in a smaller dataset.

C. Random Forest

Next, we decided to verify the accuracy of Random Forest for severity classification. We created a random forest model for each of the the same subsets that we did with CART in an attempt to improve the accuracy of our models. What we found was similar to the above, but with slightly higher accuracy rates. For example, in Exhibit 11, we can see that random forest had very slightly higher accuracies on the Full USA model (0.9543 CART vs 0.9556RF), and significantly higher on the regional models where there was less data (0.89 average over considered regions for CART vs 0.87 for RF).This is also apparent when running the regional data on the Full USA model. We can see in Exhibit 13 that the Full USA CART models are not as accurate as random forest for regional data–this makes sense as the CART trees vary for regional vs Full USA models. The issue with the random forest model on the Full USA dataset is that it was just too slow to run. In order to improve the speed of model creation we decided to test out XGBoost.

D. XGBoost

The next type of model that was part of our analysis was XGBoost. There were multiple reasons why we decided to also verify the results of this model. Firstly, it provides a good extension to what can be achieved using CART. As XGBoost is based on the decision tree ensembles that consists of a set of CART trees, we hypothesized that we would be able to achieve higher accuracy using XGBoost in comparison to CART. In addition, since the infrastructure that supports' model backend is optimized in order to use large datasets, we assumed that we expected that we would be able to obtain useful models using larger data samples in a shorter amount of time. Indeed, XGBoost provided consistently higher accuracy than CART, based on almost all data segmentation methods (Exhibit 10), mentioned in Section IIIA (see Figure 1 above). Outliers to that observation were models created for Houston, TX, where XGBoost provided the lowest accuracy out of RF and CART, for all considered time frames. We attribute this observation to the fact that models created on a city level utilize less data than models created for county, state or all-country level. As XGBoost will build multiple trees while CART only one, the greater segmentation that occurs with multiple trees overfits the model based on training data. Overall, XGBoost performed well on the Full USA Model (0.9544 accuracy) and even at the state level (average per state accuracy of 0.86) for the whole available time period. We can see in Exhibit 10 that the smallest states have the lowest accuracy and we believe that this might be due to having much less data available for these states.

IV. Results

The classification models that we developed had the highest accuracy when built with a 500,000 row sample of all USA data for all years. Although the CART trees looked different for different cities/states vs the whole USA, the most important variables were shared. Distance (the length of the road affected by the accident) was the most important variable in roughly 90% of our models. Pressure and temperature were also in the top 5 most important variables in almost every model. Of the most important variables in the full USA/all years model, the top 5, distance, pressure, temperature, traffic signal and wind speed, were present in most of the regional models. This was a good validation for using the full USA model.

Finally, in an initial analysis, we built the models with data from all years to actually include the years as an independent factor variable. In these models, while different regions of the US had different important variables, in the local models that spanned data from all years there was an interesting, shared important variable, the year of the accident. The CART trees for these models actually had the first split on whether the year was 2020/2021 or not. This led us to believe COVID (and other large societal changes) could affect the severity of traffic accidents.

V. Conclusion and Recommendation

The classification models we developed performed well at the city and state level and allowed us to gain insight into the possible impacts of COVID on traffic accident severity. The economic and societal impact of traffic accidents cost states and cities across the U.S. amounts to hundreds of billions of dollars annually. A large part of these losses is driven, largely, by a small number of severe accidents. Building high-performing predictive models, similar to ours, can aid in helping public agencies reduce both the number and severity of traffic accidents. Utilizing random forests to build the main model that could be provided at a national level, by the National Highway Traffic Safety Administration, would be the most accurate; however, it is the slowest to create. If it does not need to be run frequently it would be the best, if it needs to be run often then XGBoost is significantly faster on the large data set and still quite accurate. For local traffic and emergency response groups, if they are concerned about access to technology or the internet, a CART tree might help them predict severity in real time without any additional technology.

VI. Future Considerations

To make our model more robust, if we continued with this project, we would like to consider the following relationships: rural vs. non-rural states and cities, republican vs. democratic states and cities, high-density vs. low-density cities.In addition, it would also be beneficial to expand horizontally through the addition of new variables that represent whether or not drivers were distracted, under the influence, experiencing a medical emergency, etc. This could improve the models performance and provide new insight into the drivers of accident severity. Further, as the amount of data available was often the limiting factor, if we were able to obtain data with more diversity in the severity level, as opposed to our dataset with a large majority of the data in severity level 2, we may be able to discover more insights and better models for predicting severity level.

VII. Appendix

Our code for data preprocessing, model implementations, as well as visualizations can be found in this folder. The original dataset we used in the code is here. The following Exhibits diagram our analysis.

Exhibit 1: Number of Accidents by Severity

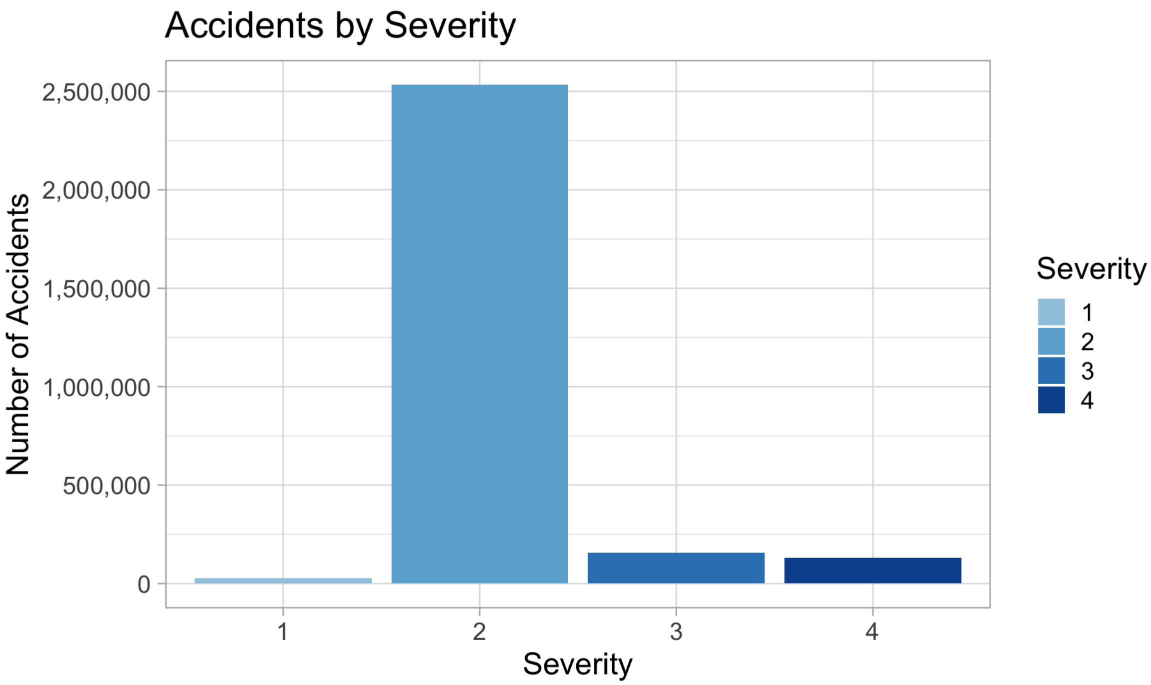


Exhibit 2: Number of Accidents by Year and Severity

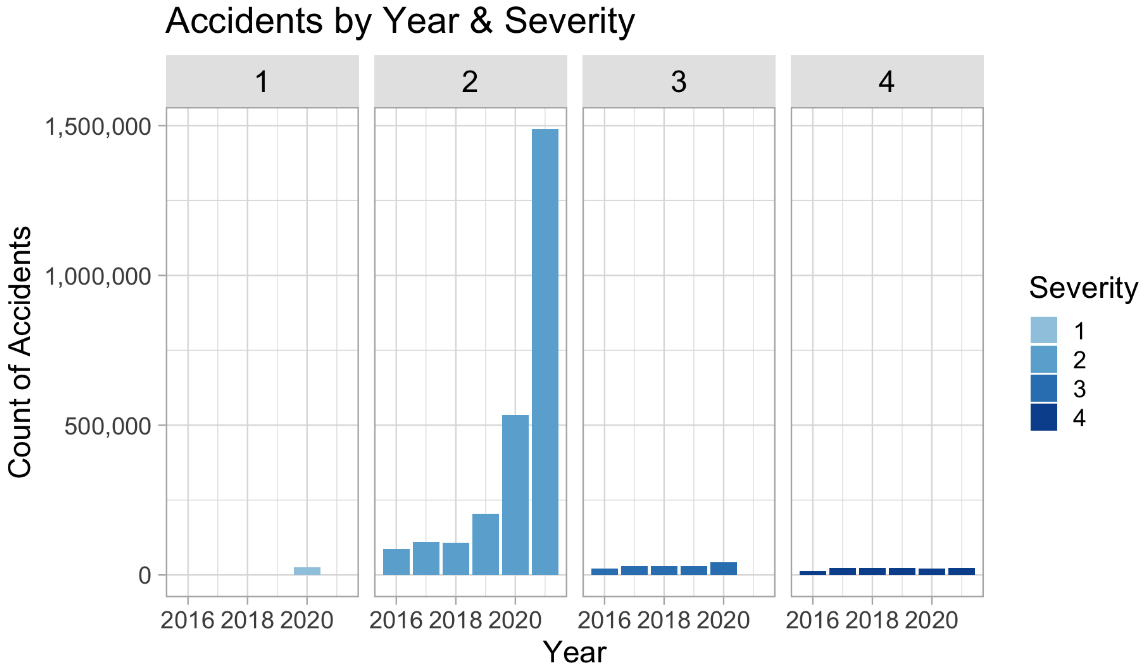


Exhibit 3: Number of Accidents by State (whole dataset 2016-2021)

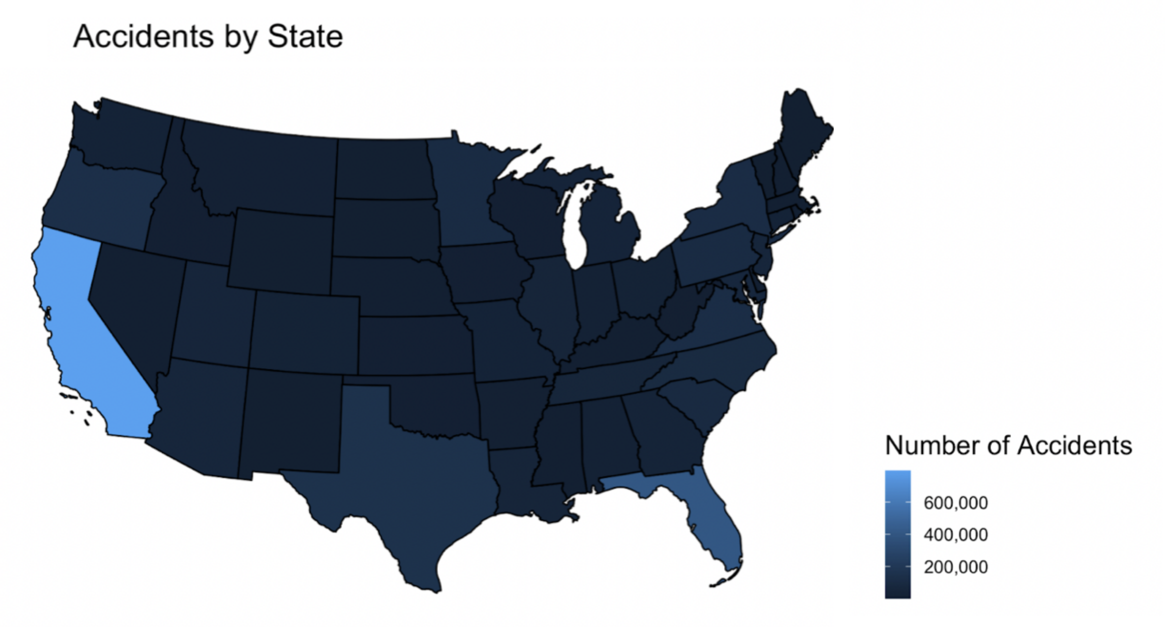


Exhibit 4: Correlation plot with selected preprocessed variable before removing correlated variables. Colors correspond to the correlation

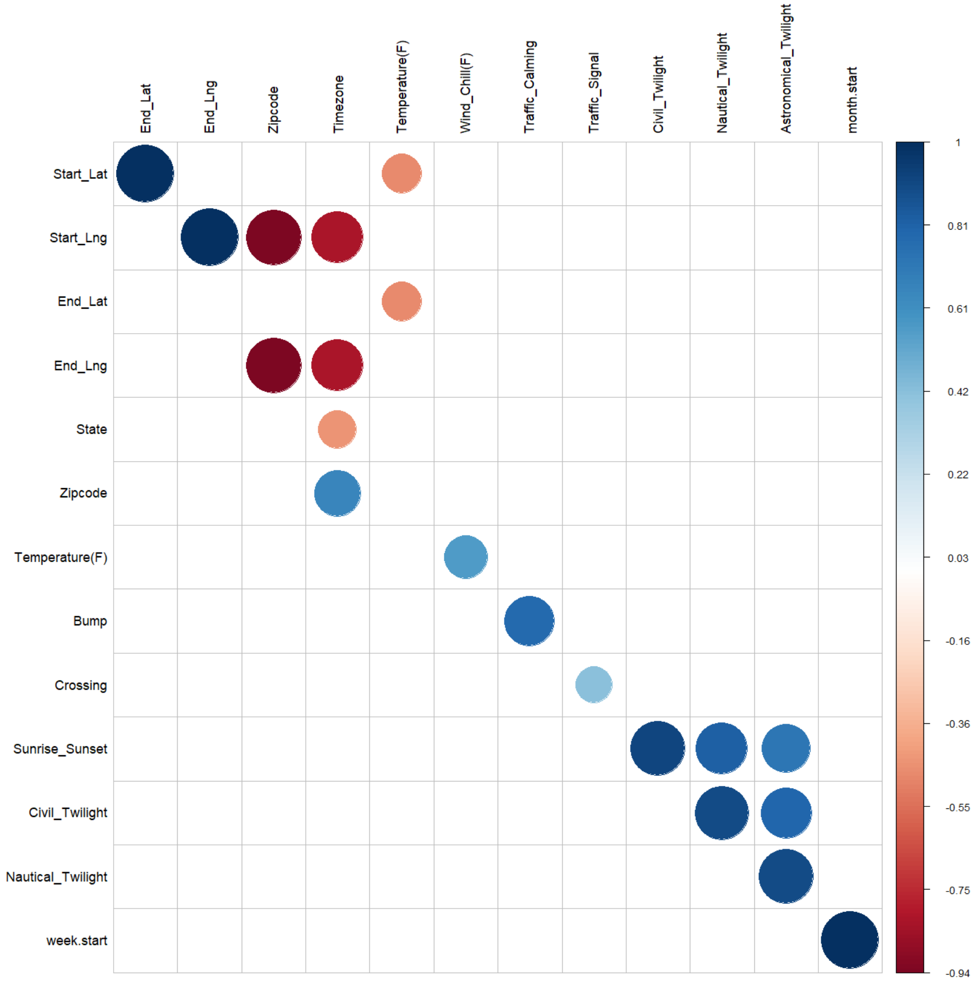


Exhibit 5: All variable pairs with |correlation score| > 0.4

```
                  Var1                  Var2       Freq
175          Start_Lng               End_Lng  0.9999992
131          Start_Lat               End_Lat  0.9999958
1804        week.start           month.start  0.9967559
476          Start_Lng               Zipcode -0.9457387
478            End_Lng               Zipcode -0.9457343
1496    Sunrise_Sunset        Civil_Twilight  0.9098572
1584 Nautical_Twilight Astronomical_Twilight  0.8993190
1540    Civil_Twilight     Nautical_Twilight  0.8919200
521            End_Lng              Timezone -0.8161654
519          Start_Lng              Timezone -0.8161571
1539    Sunrise_Sunset     Nautical_Twilight  0.8116021
1583    Civil_Twilight Astronomical_Twilight  0.8022109
1312              Bump       Traffic_Calming  0.7801172
1582    Sunrise_Sunset Astronomical_Twilight  0.7300879
528            Zipcode              Timezone  0.6658092
616     Temperature(F)         Wind_Chill(F)  0.5579491
563            End_Lat        Temperature(F) -0.4731741
561          Start_Lat        Temperature(F) -0.4731598
527              State              Timezone -0.4480216
1356          Crossing        Traffic_Signal  0.4125112
```

Exhibit 6: Correlation plots of preprocessed data before and after eliminating correlated variables

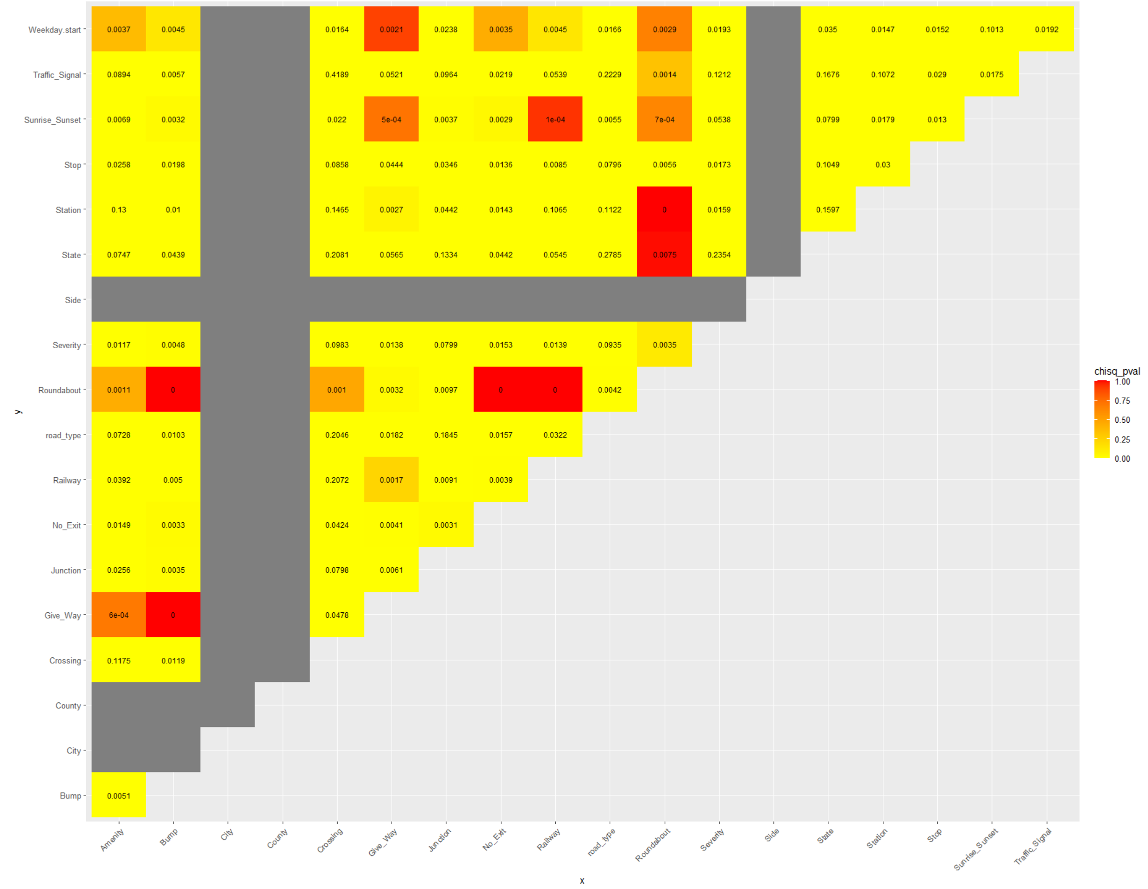


Exhibit 6 (con):

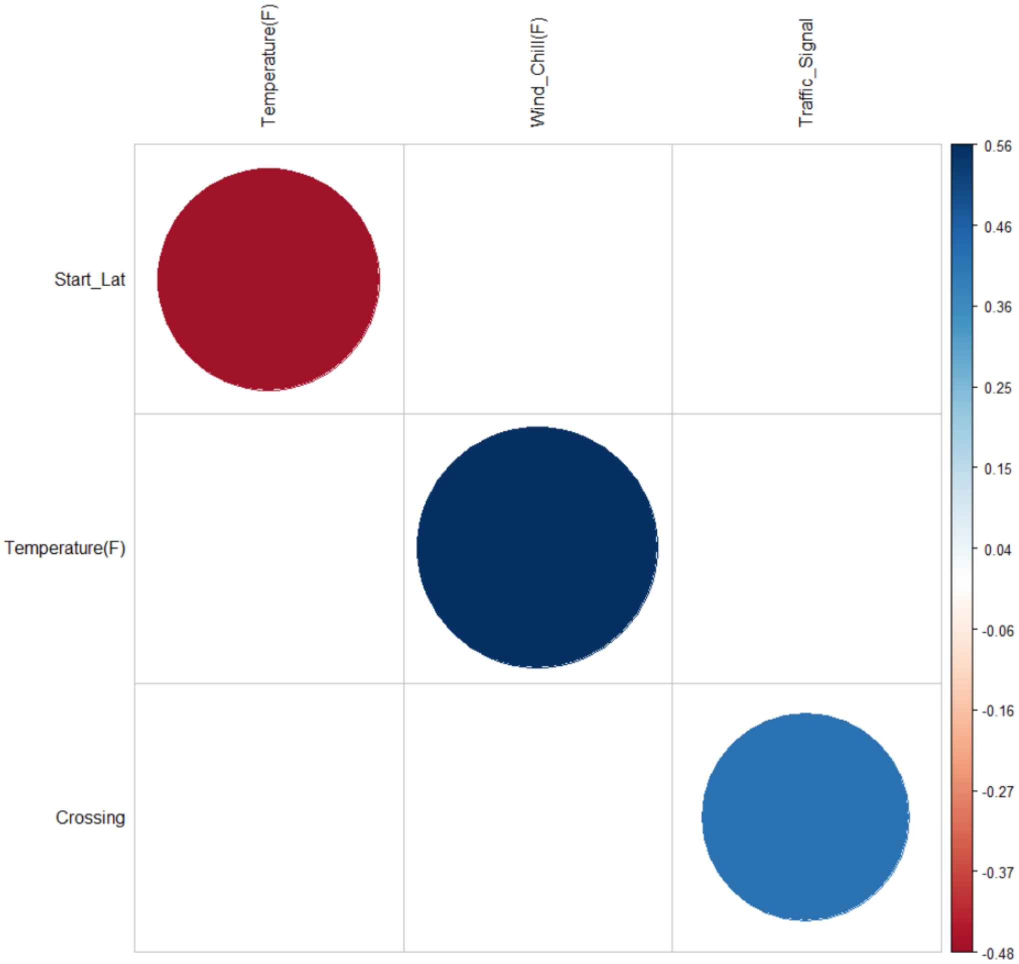


Exhibit 7: CART Model for All of USA and All Years

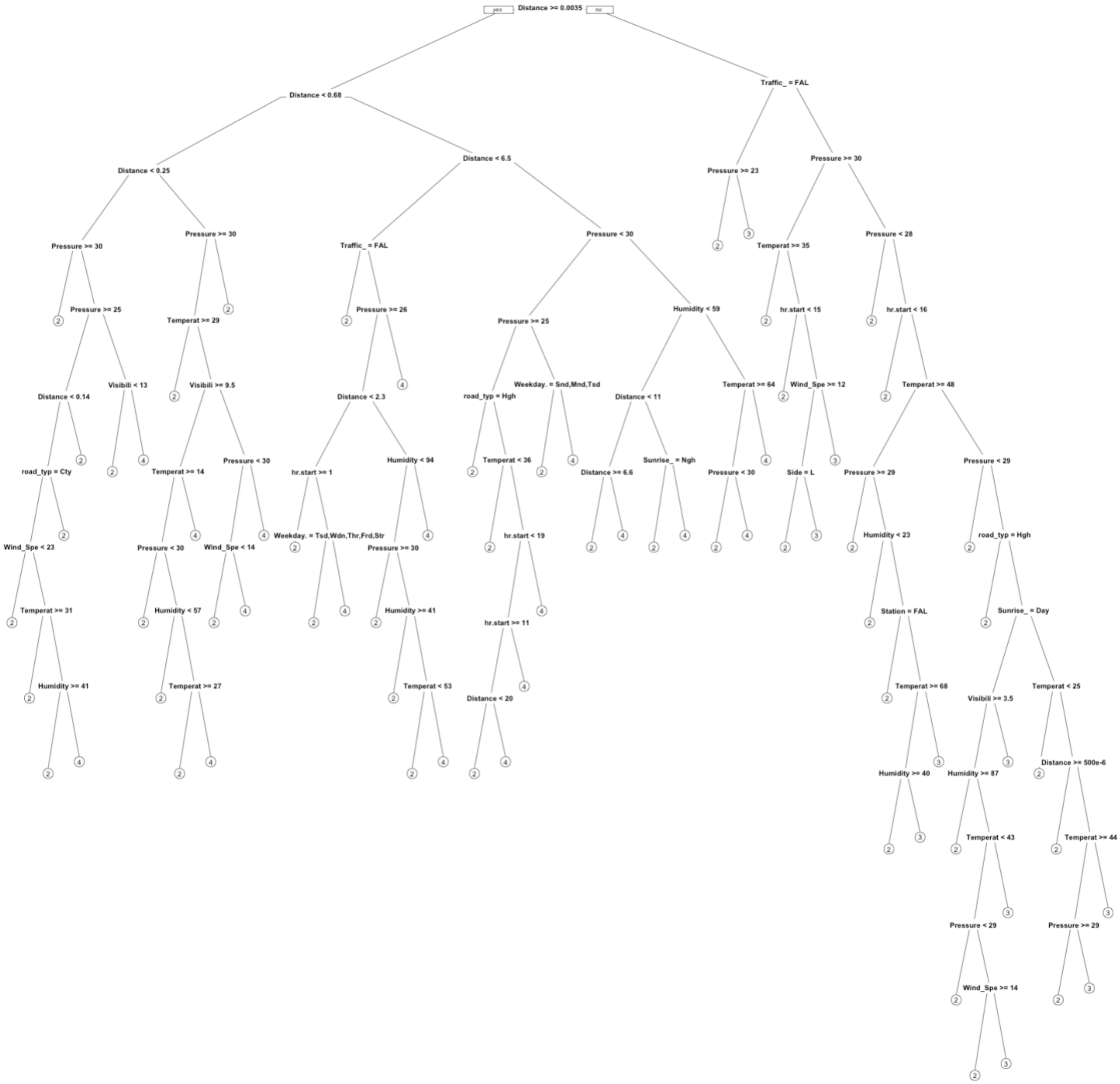
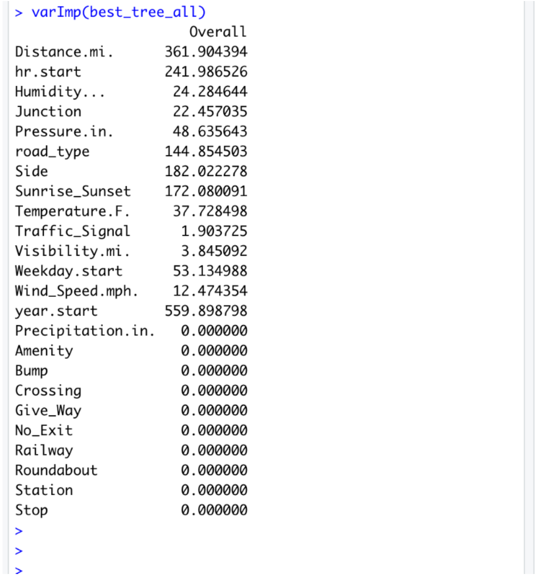


Exhibit 8: CART Model for All of USA and All Years and Year as IV



```
> varImp(best_tree_all)
                    Overall
Distance.mi.     361.904394
hr.start         241.986526
Humidity...       24.284644
Junction          22.457035
Pressure.in.      48.635643
road_type        144.854503
Side             182.022278
Sunrise_Sunset   172.080091
Temperature.F.    37.728498
Traffic_Signal     1.903725
Visibility.mi.     3.845092
Weekday.start     53.134988
Wind_Speed.mph.   12.474354
year.start       559.898798
Precipitation.in.  0.000000
Amenity            0.000000
Bump               0.000000
Crossing           0.000000
Give_Way           0.000000
No_Exit            0.000000
Railway            0.000000
Roundabout         0.000000
Station            0.000000
Stop               0.000000
>
>
>
```

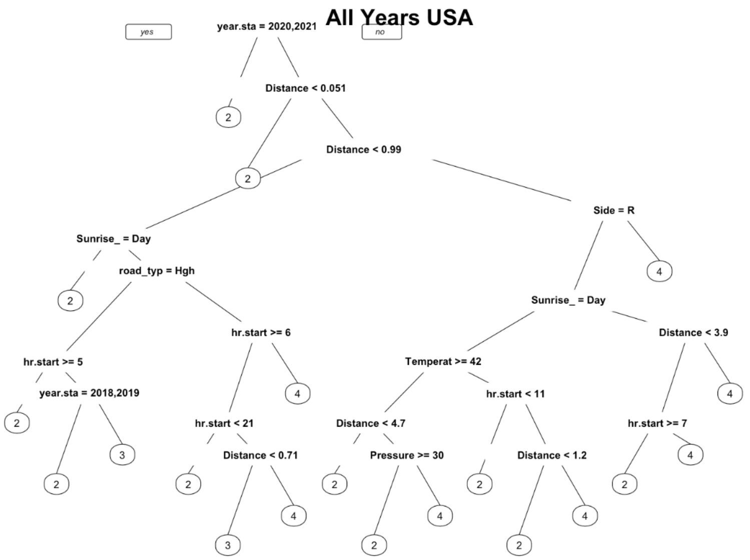

All Years USA

Exhibit 9: Per state XGBoost accuracies for 3 considered time spans

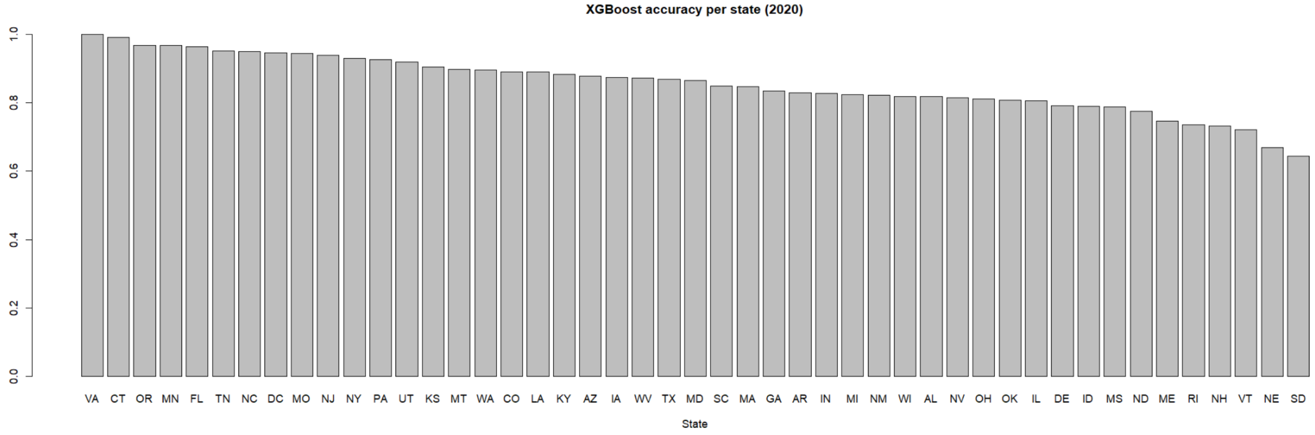
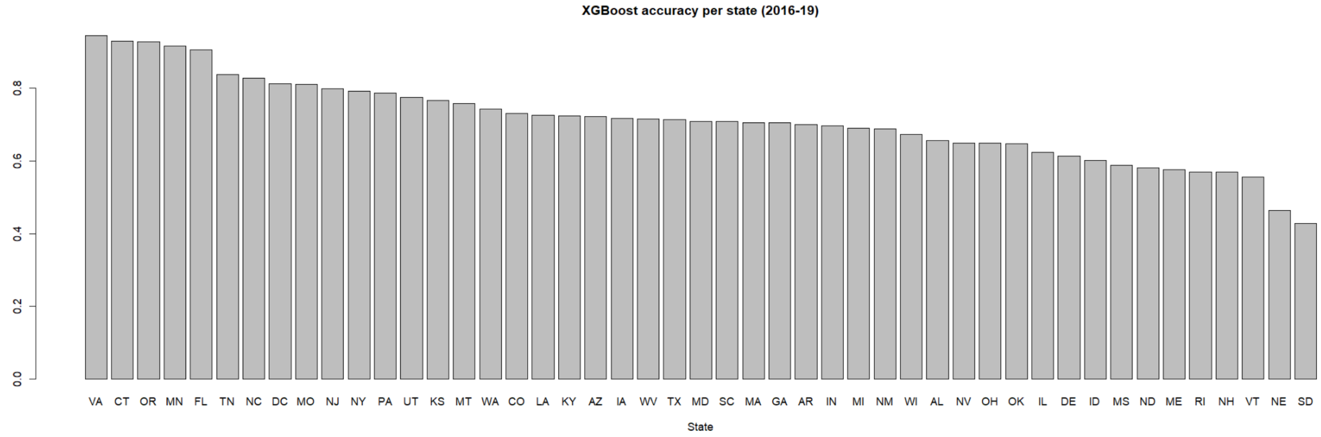

XGBoost accuracy per state (2020)

Exhibit 9 (con):



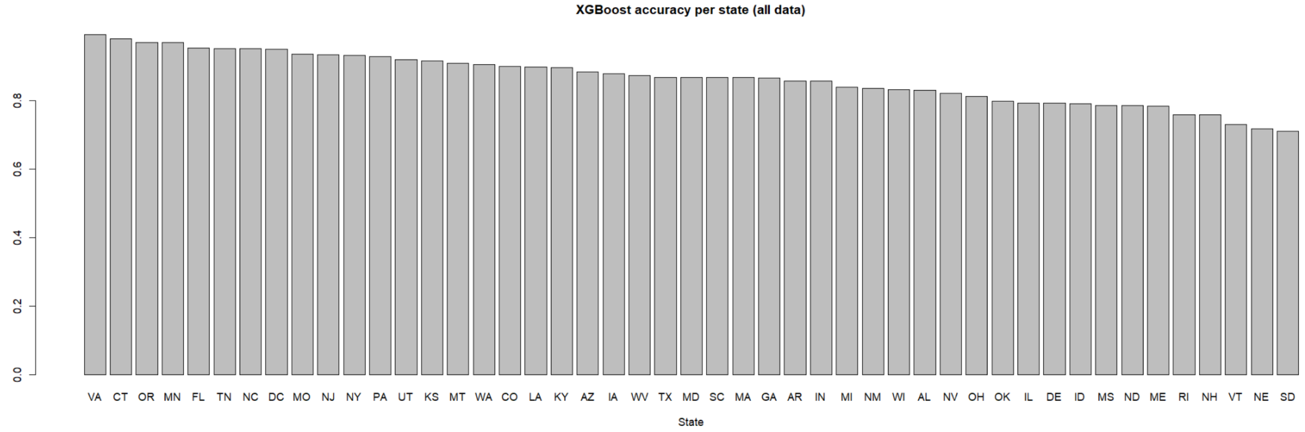
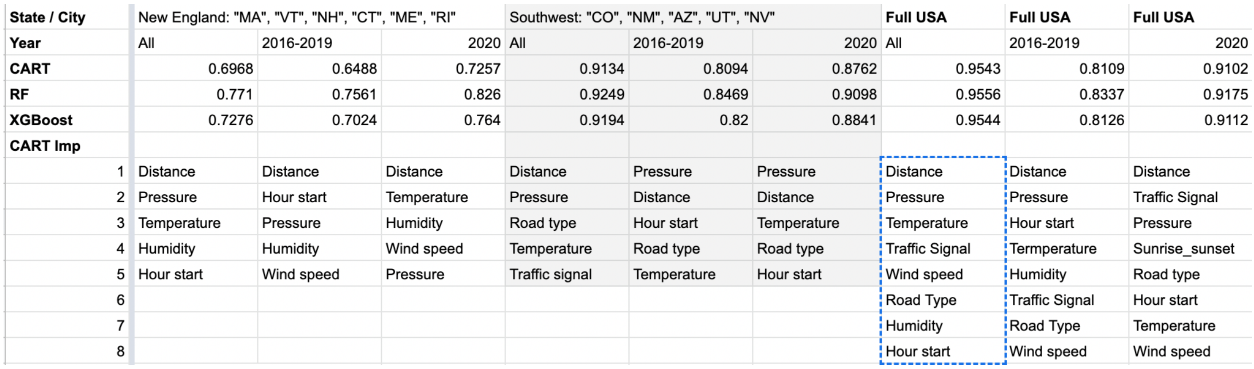


Exhibit 10: Model Accuracy for State- and City-Level Analysis

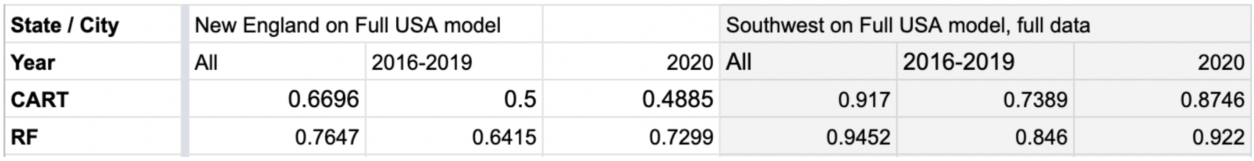
https://docs.google.com/spreadsheets/d/1cfnMTL-EVjcBTM6cTh3tyG8i4w2-G7tBQApk_LTUWnE/edit?usp=sharing

Exhibit 11: Comparison of Regional Models to Full USA model accuracies



State / City	New England: "MA", "VT", "NH", "CT", "ME", "RI"			Southwest: "CO", "NM", "AZ", "UT", "NV"			Full USA	Full USA	Full USA
Year	All	2016-2019	2020	All	2016-2019	2020	All	2016-2019	2020
CART	0.6968	0.6488	0.7257	0.9134	0.8094	0.8762	0.9543	0.8109	0.9102
RF	0.771	0.7561	0.826	0.9249	0.8469	0.9098	0.9556	0.8337	0.9175
XGBoost	0.7276	0.7024	0.764	0.9194	0.82	0.8841	0.9544	0.8126	0.9112
CART Imp									
1	Distance	Distance	Distance	Distance	Pressure	Pressure	Distance	Distance	Distance
2	Pressure	Hour start	Temperature	Pressure	Distance	Distance	Pressure	Pressure	Traffic Signal
3	Temperature	Pressure	Humidity	Road type	Hour start	Temperature	Temperature	Hour start	Pressure
4	Humidity	Humidity	Wind speed	Temperature	Road type	Road type	Traffic Signal	Termperature	Sunrise_sunset
5	Hour start	Wind speed	Pressure	Traffic signal	Temperature	Hour start	Wind speed	Humidity	Road type
6							Road Type	Traffic Signal	Hour start
7							Humidity	Road Type	Temperature
8							Hour start	Wind speed	Wind speed

Exhibit 12: Accuracies of regional data run on Full USA model



State / City	New England on Full USA model			Southwest on Full USA model, full data		
Year	All	2016-2019	2020	All	2016-2019	2020
CART	0.6696	0.5	0.4885	0.917	0.7389	0.8746
RF	0.7647	0.6415	0.7299	0.9452	0.846	0.922

Exhibit 13 - Link to code

https://drive.google.com/drive/folders/1_8qz5b-Sm3TuuDmzoVlm40aLJhr6BFEY?usp=share_link