NASHVILLE

PREDICTING PUBLIC TRANSPORTATION SAFETY RISK

SPRINGBOARD
Capstone Project 2

Image Creator: jmsilva | Credit: Getty Images

2023 I MARCH

TAMARA HORNE

Image credit: www.newschannel5.com

Table of Contents

Introduction

Nashville, TN built a 2.1 million square foot convention center in 2013 to host both local and international events, but limited public transportation options are available between the convention center and the airport. The Nashville Department of Transportation and Multimodal Infrastructure is considering adding additional public transportation between the two to help reduce the pressure on the congested roadways. They would like to understand the safety implications of doing so as part of their decision-making process.

Using tree based models on three different aggregations, I was able to achieve R Squared values of 0.99, 0.83, and 0.94 respectively. These models and aggregations provide insight into the most important features influencing the safety risk for the nation as a whole, those influencing the safety risk of individual locations, and those influencing the safety risk of each of the modes.

The resulting recommendation is that in terms of safety risk, Nashville would be better served by increasing service to one of its existing modes instead of adding an additional mode of transportation like light rail. Implementation details can be found in my GitHub Repository Folder (https://github.com/tamarahorne/Springboard/tree/main/Capstone%20Project%202).

Data Acquisition & Wrangling

I utilized the National Transit Database's monthly modal time series data from the U.S. Department of Transportation's website (link provided in resources section) which contained 133,196 rows of data gathered from 2014 to present day. Each row contains data for one month for each reporting agency, consisting of features ranging from population and size information for each area; the types of modes, organizations, and service; and the number of safety events (fatalities, injuries, derailments, security events, etc.) which occurred for that location during the given month.

To prepare the data for modeling, I filled NaNs in over 4000 rows using appropriate data found elsewhere in the dataframe and an additional 27 using information available online through the FTA. Forty eight rows had no identifying location information and so were removed, resulting in a dataframe of 133,148 rows.

Storytelling & Inferential Statistics

The dataframe had 65 columns, so it was necessary to gain a better understanding of their contents to determine their usefulness to the project. After exploring the data, I determined that the values in many columns were accounted for in the Total Fatalities and Total Injuries columns, so the totals columns were retained and the others were dropped. I was also able to drop one of the two ID columns and the Primary UZA Code column which was an alternate representation of Primary UZA Population. I further reduced the dataframe by removing rows containing data for modes not relevant to the project. Nashville would not be eligible to develop an Alaska railroad, for example. The reduced dataframe contained 129,752 rows and 26 columns, and the Nashville subset contained 750 rows and 26 columns of data.

I then took a first glance at the relationships between some of the features and was interested to see the strong correlation between 'Total Collisions' and 'Total Events', both in the national data and the Nashville data (f*igure 1).* After further exploration, I found the values in 'Total Collisions' and other columns were included in the value of 'Total Events', so the contributing columns were removed.

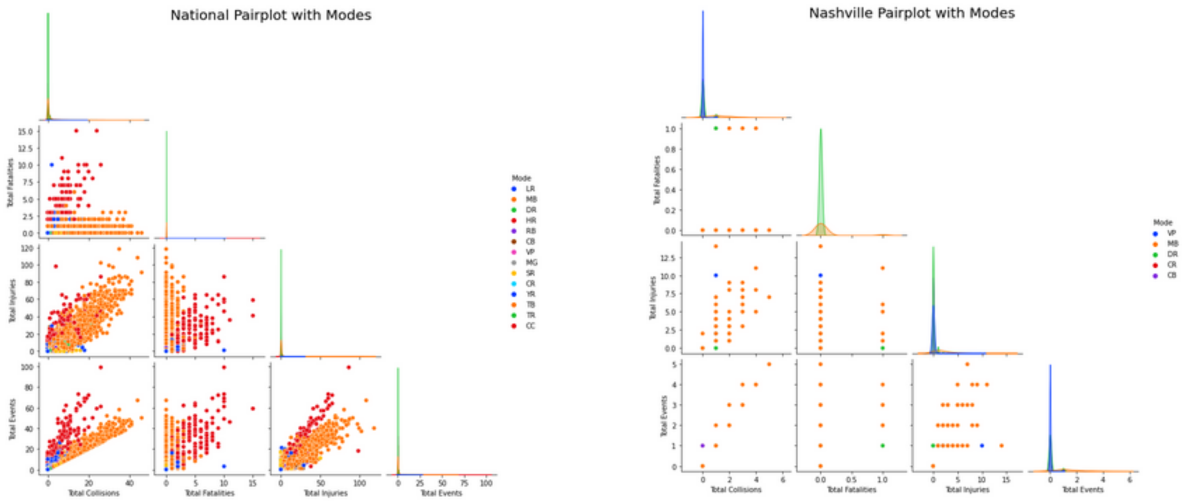
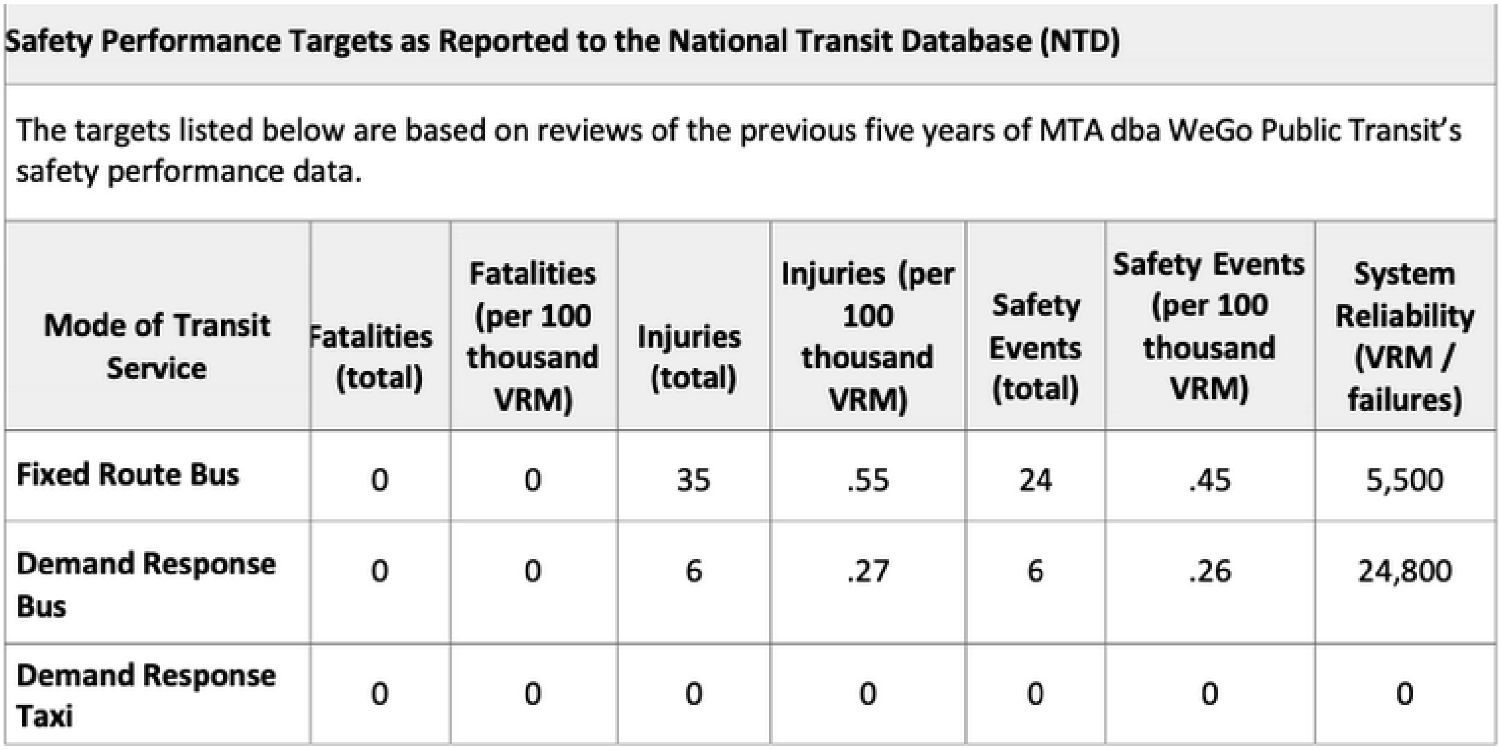


figure 1

Coloring the data points by mode revealed that the majority of reported safety risk incidences were associated with bus service (MB) in both the national data and Nashville, and heavy rail (HR) in the national data.

Storytelling & Inferential Statistics (continued)

I defined the target using the safety performance measures used by the Metropolitan Transit Authority (*figure 2*), the largest agency serving Nashville, which matches the example given by the Federal Transit Authority (FTA) in their Safety Performance Targets Guide (link available in resources section).



Safety Performance Targets as Reported to the National Transit Database (NTD)

The targets listed below are based on reviews of the previous five years of MTA dba WeGo Public Transit's safety performance data.

Mode of Transit Service	Fatalities (total)	Fatalities (per 100 thousand VRM)	Injuries (total)	Injuries (per 100 thousand VRM)	Safety Events (total)	Safety Events (per 100 thousand VRM)	System Reliability (VRM / failures)
Fixed Route Bus	0	0	35	.55	24	.45	5,500
Demand Response Bus	0	0	6	.27	6	.26	24,800
Demand Response Taxi	0	0	0	0	0	0	0

figure 2

System Reliability was outside the scope of this project, so I defined the target as

(Total Fatalities + Total Injuries + Total Events) / Vehicle Revenue Miles

Exploratory Aggregation

I performed some exploratory aggregations in order to see relationships between features over time. I looked at a subset of locations which had modes up to and including the ones that Nashville has, and then a second subset of locations which had those modes plus light rail. The nine locations in the group which had light rail showed significant higher averages for injuries and events (*figure 3*).

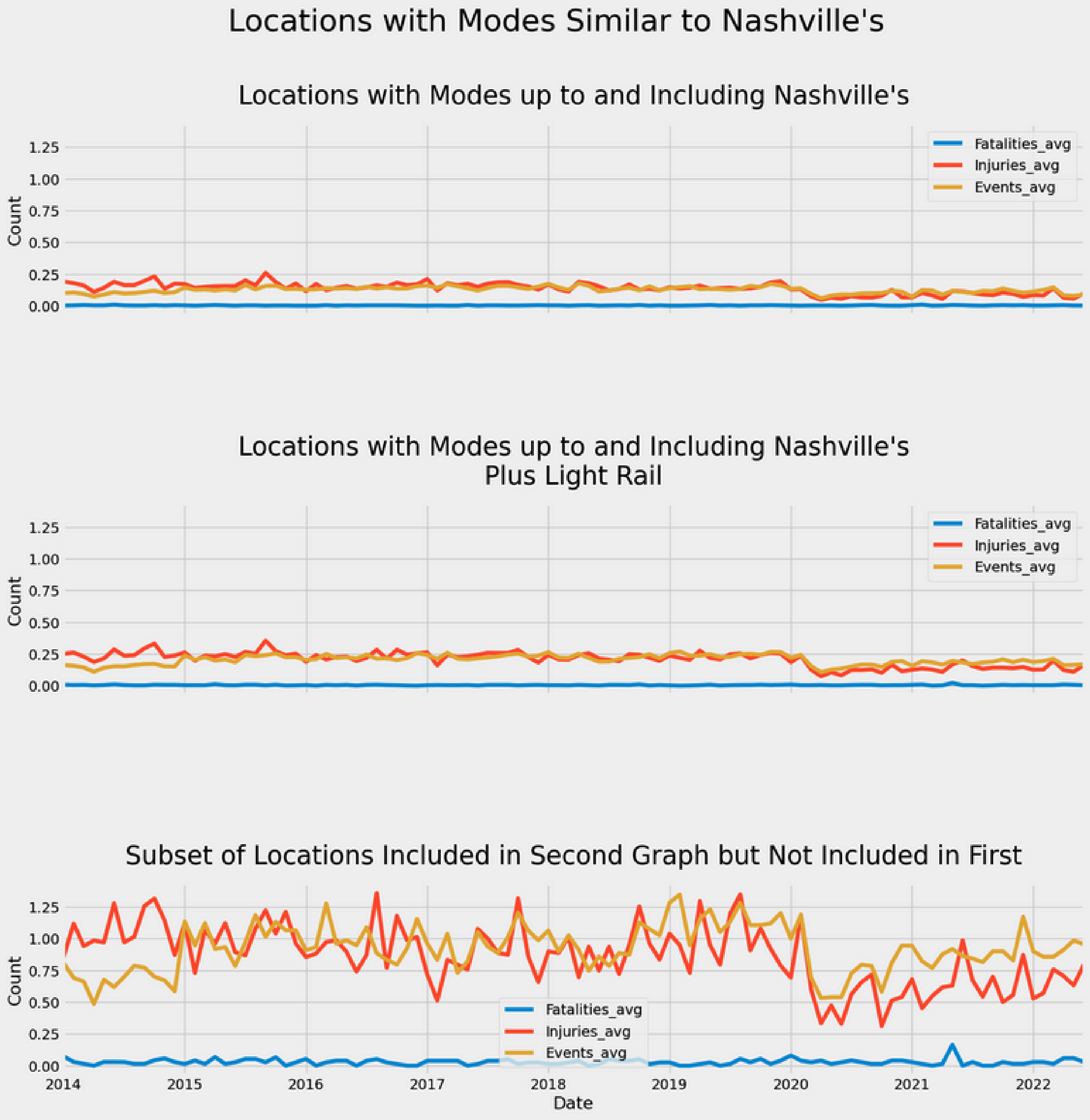


figure 3

Baseline Modeling

modeling_df

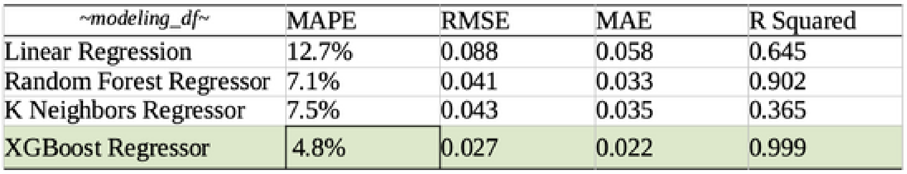
To begin modeling, I created a dataframe (modeling_df) using all of the national data aggregated to date. This dataframe contains 102 rows – one for each month of data. The total fatalities, injuries, and events were summed for each month; the population and size data were first averaged within each location (Primary UZA Name) and then summed to created a national level figure. I developed a baseline model using Linear Regression, but with a MAPE at 12.7%, I suspected another type of model might perform better.

Extended Modeling

Next, I tried K Neighbors Regressor and Random Forest Regressor which both had MAPEs around 7% and RMSEs around 0.04. Encouraged by the successfulness of Random Forest, I decided to try one more tree-based ensemble machine learning algorithm, XGBoost.

Findings

XGBoost Regressor ended up being the winner with a MAPE of 4.8% and an RMSE of 0.027. The full results of the four models are displayed in the table below (*figure 4*).



~modeling_df~	MAPE	RMSE	MAE	R Squared
Linear Regression	12.7%	0.088	0.058	0.645
Random Forest Regressor	7.1%	0.041	0.033	0.902
K Neighbors Regressor	7.5%	0.043	0.035	0.365
XGBoost Regressor	4.8%	0.027	0.022	0.999

figure 4

Modeling with the data aggregated only to the date is handy for seeing the big picture: the features influencing the safety risk ratio for the nation as a whole. I preserved the time order by sorting the data by date and setting shuffle to 'false' when splitting the data into train/test sets, so the regression models would be attempting to predict the more recent data by training on the earlier data.

With XGBoost run on the whole modeling_df dataframe, we can see the feature relationships displayed below in the SHAP summary plot (*figure 5*). Ridership is overwhelmingly the feature of greatest importance, and higher ridership results in higher SHAP values. In other words, having more riders increases the safety risk.

Findings (continued)

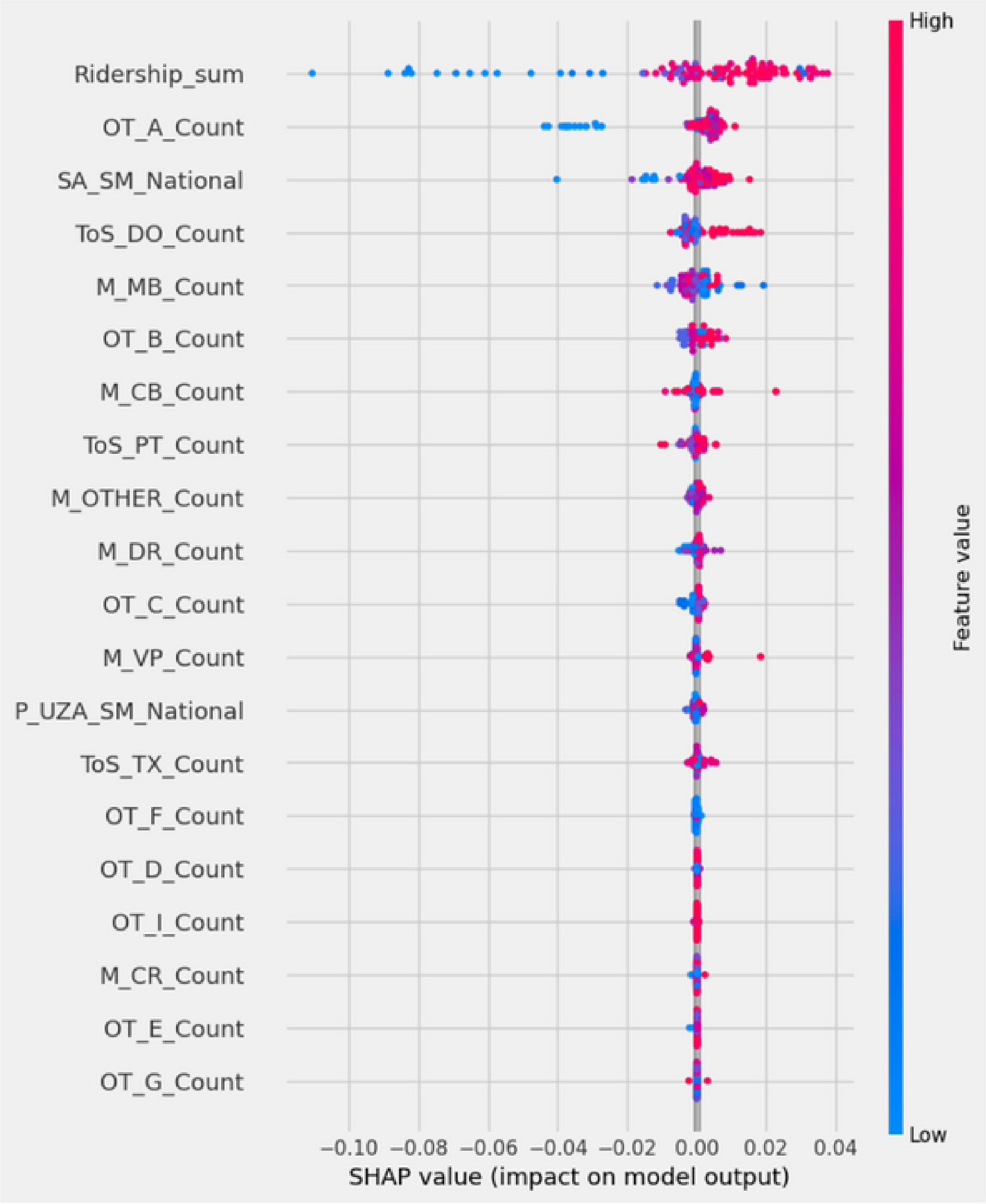


figure 5

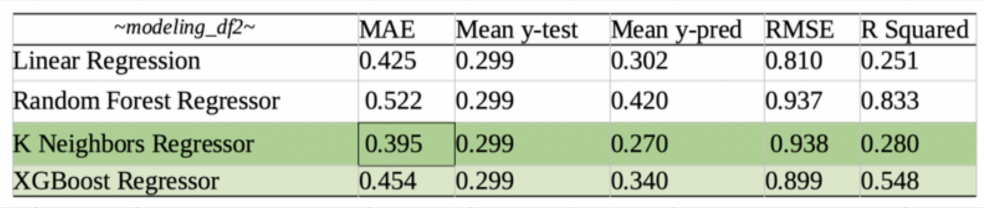
The second feature, OT_A_Count, refers to the 'Independent Public Agency or Authority of Transit Service' organization type. This is the most frequently occurring organization type in the national data. The third feature SA_SM_National, is the average service area square miles for the nation during each given month. Lower SHAP values of both of these features correlate with lower feature importance values, suggesting lower safety risk. One caution here that correlation does not imply causation.

Second Aggregation & Modeling

modeling_df2

The first aggregation was a fine starting point for understanding safety risk. However, I wanted to see if it would be possible to predict at more granular levels. Could I predict safety risk when the data was aggregated to date and location? Could I predict it if the data was aggregated to date and mode? I performed the same four regression models on these two aggregations to find out.

With the data aggregated to date and location (modeling_df2), K Neighbors had the best MAE. (It was not possible to get the MAPE because zero values in the target caused division by zero.) XGBoost had a better MAE than Random Forest but was not quite as good as K Neighbors (*figure 6*). I also found that my residuals were shifted to the left of center for all models (*figure 7*) and regrettably, I did not have time to explore the reasons for this before submission.



~modeling_df2~	MAE	Mean y-test	Mean y-pred	RMSE	R Squared
Linear Regression	0.425	0.299	0.302	0.810	0.251
Random Forest Regressor	0.522	0.299	0.420	0.937	0.833
K Neighbors Regressor	0.395	0.299	0.270	0.938	0.280
XGBoost Regressor	0.454	0.299	0.340	0.899	0.548

figure 6

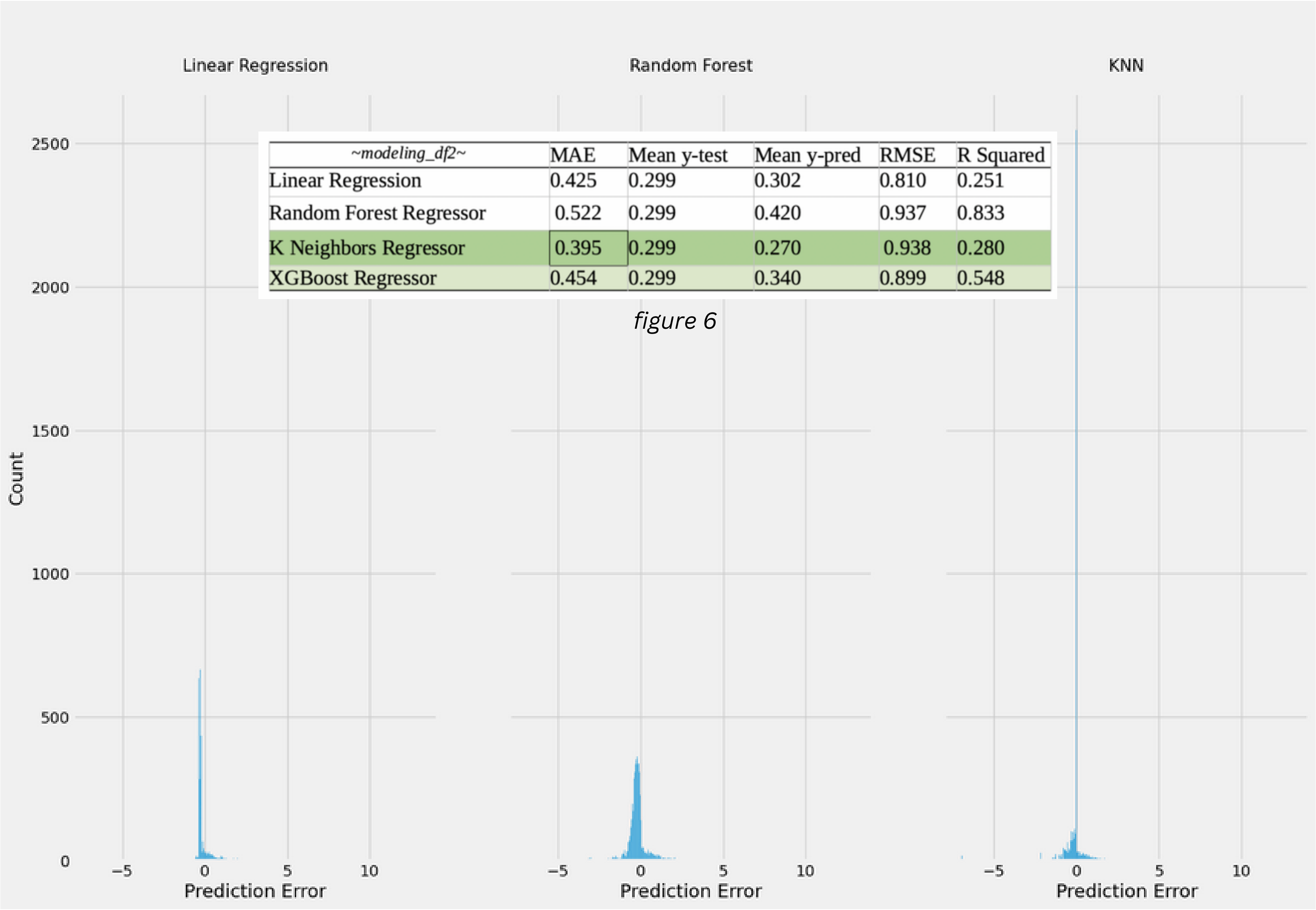


figure 7

Second Aggregation & Modeling (continued)

modeling_df2

The XGBoost Regressor run on 1000 samples from the whole modeling_df2 dataframe suggests the size and population of the Primary UZA have the biggest influence on safety risk (*figure 8*), followed by ridership, and then the size and population of the service areas. A Primary UZA is equivalent to a city or metropolitan area whereas a service area refers to an area serviced by a given agency. A service area could be the same as it's corresponding UZA or a subset thereof. Service areas for different agencies can also overlap within a UZA.

National

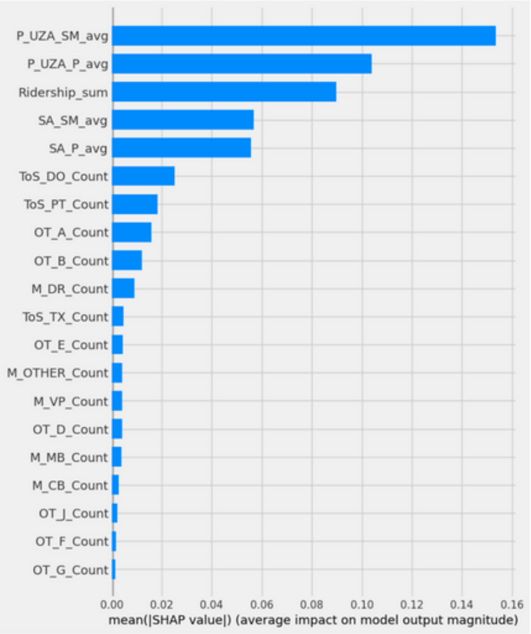


figure 8

Nashville

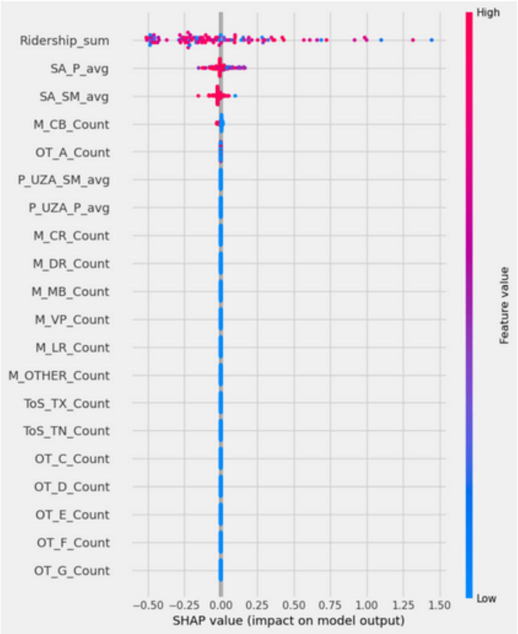


figure 9

When we look just at Nashville's data using the same aggregation, we can see the order of feature importance is slightly different (figure 9). UZA population and size are no longer a factor because they are a single value for Nashville. Nashville does have three service areas, though, so those values continue to influence safety risk for this location, although ridership's importance takes a clear lead.

We can see a scattering of both high and low SHAP values mixed with both high and low feature importance values for ridership. The same is true for service area population (SA_P_avg), which is the average population for that service area in each given month. The distribution of SHAP values for SA_P_avg, though, is far more narrow than those for ridership. We also see two curiously large SHAP values for low ridership, which would be great to look into if time allowed.

Second Aggregation & Modeling (continued)

modeling_df2

Even though the modeling for this dataframe was not the most trustworthy because of the shifted data, the aggregation itself was informative. The graphic below (*figure 10*) is a portion of the whole plot showing the average safety risk ratio per location, where the dashed vertical blue line shows the national average. Nashville's safety risk ratio is higher than the national average but it is not among the highest in the nation.

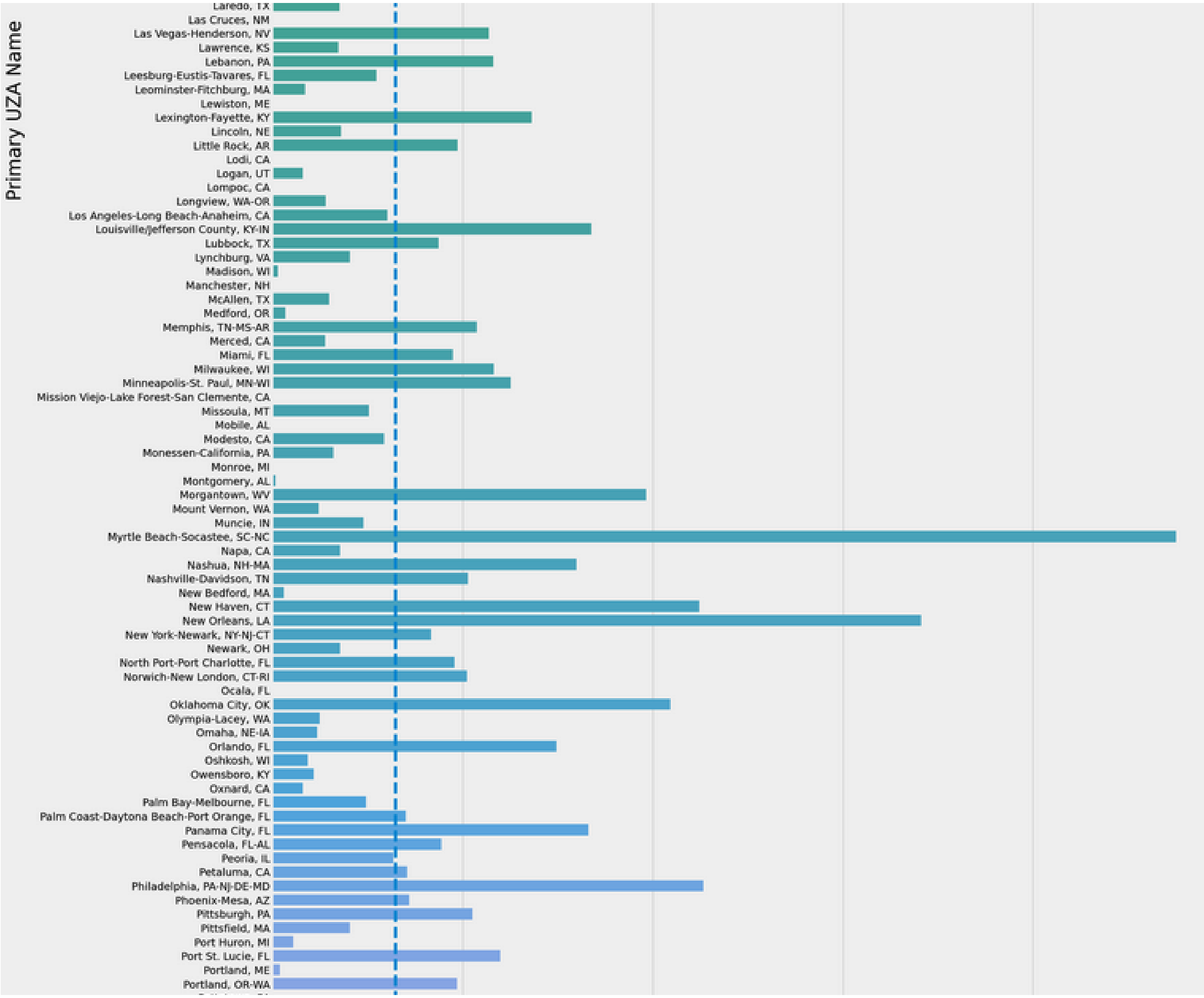
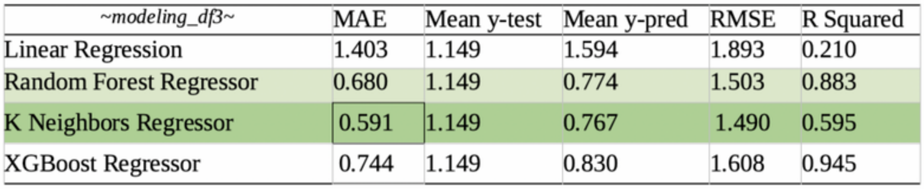


figure 10

Third Aggregation & Modeling

modeling_df3

With the data aggregated to date and mode (modeling_df3), K Neighbors again had the lowest MAE. (MAPE was not possible again due to zeros in the target.) Random Forest Regressor had the second lowest MAE and was almost equivalent to K Neighbors in terms of their RMSEs. XGBoost did not perform as well as Random Forest for this aggregation, so Random Forest was used for interpretation.



~modeling_df3~	MAE	Mean y-test	Mean y-pred	RMSE	R Squared
Linear Regression	1.403	1.149	1.594	1.893	0.210
Random Forest Regressor	0.680	1.149	0.774	1.503	0.883
K Neighbors Regressor	0.591	1.149	0.767	1.490	0.595
XGBoost Regressor	0.744	1.149	0.830	1.608	0.945

figure 11

With Random Forest run on the whole modeling_df3 dataframe, we see the order of feature importance has changed from what they were for modeling_df (*figure 12*). Here, the service area square miles (SA_SM_National) is the feature of greatest importance with ridership (Ridership_sum) taking second place.

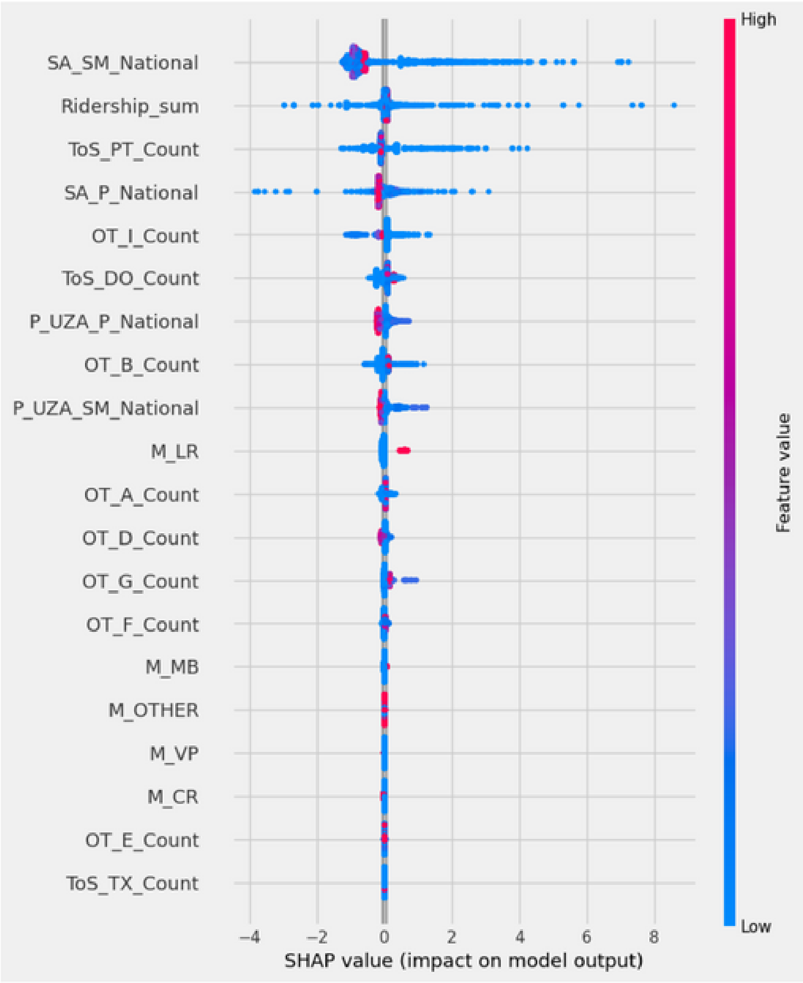


figure 12

Third Aggregation & Modeling (continued)

modeling_df3

Taking a closer look, we can see from the SHAP dependence plot below (*figure 13*) that the size of the service area (SA_SM_National) is most closely associated with Ridership (Ridership_sum). Higher SHAP values are associated with the smaller service area sizes - under 25000 square miles. Negative SHAP values and low Ridership values follow every level of service are size and intermixed with these are a few a few high Ridership values which do not follow a pattern. There are a small number for very small service areas and several for service areas around 125,000 - 150,000 square miles.

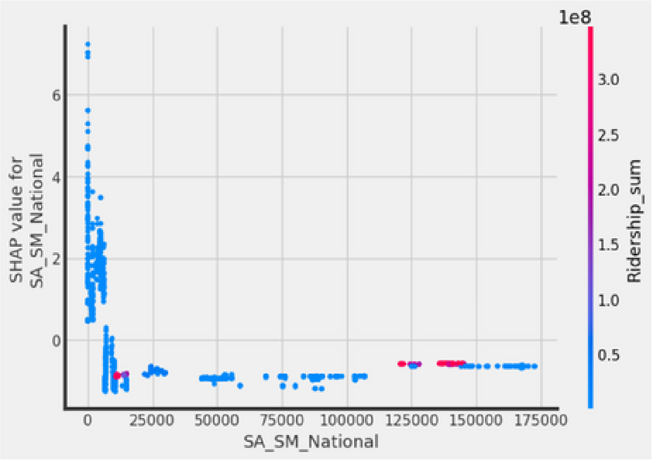


figure 13

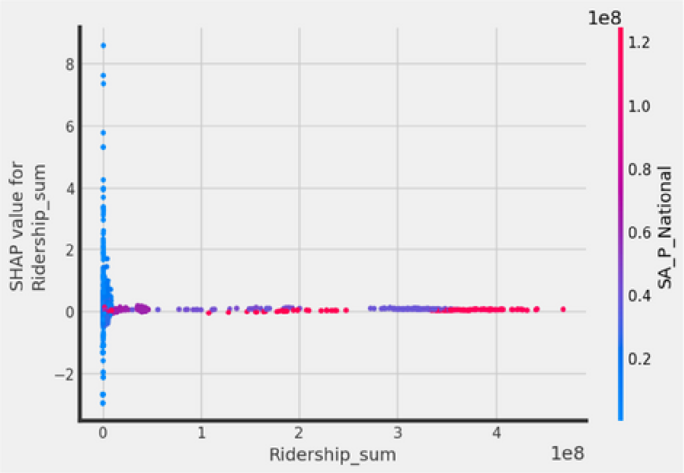


figure 14

The dependence plot for ridership (*figure 14*) shows it is most closely associated with the population size of the service area (SA_P_National). Higher ridership (over about 300 million) is positively associated with higher service area population, but the relationship is not perfectly linear below that level of ridership. We can see a mix of both medium and high population sizes when ridership is between 1 and 2 million for example.

All of the lowest population sizes are associate with the lowest ridership, though, and there are a wide range of SHAP values for that cluster. The SHAP values for all other values for ridership fall just north of zero, meaning they have only a slightly increasing affect on safety risk.

Third Aggregation & Modeling (continued)

modeling_df3

Additionally, with the data aggregated this way, I was able to collect the average safety risk ratios for each mode and compare them to the overall average, which is represented by the vertical dashed blue line (*figure 15*).

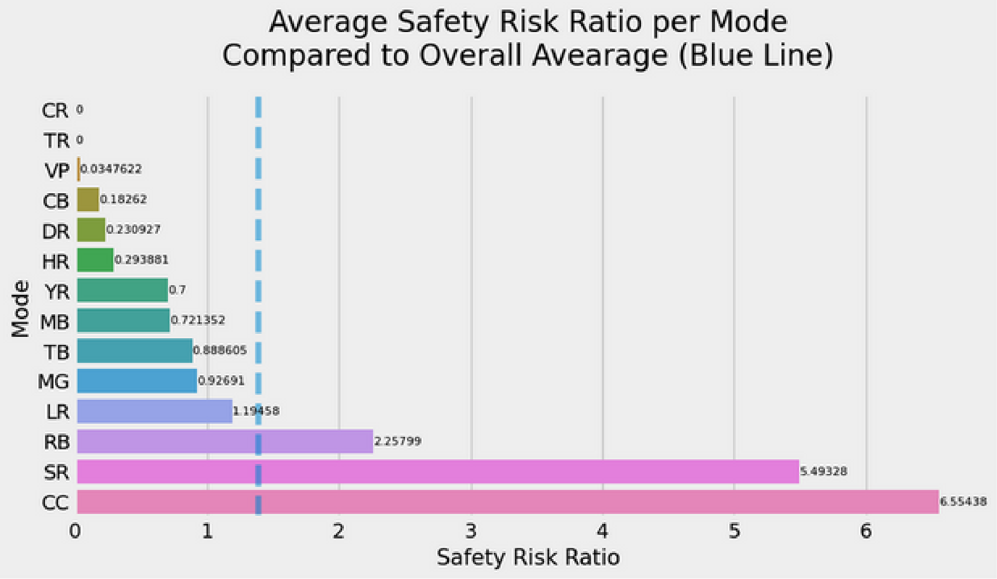


figure 15

Nashville currently has modes CB, CR, DR, MB, and VP. Of these, MB has the highest safety risk ratio in the national data, confirming what I saw in Figure 1. Light rail is a mode Nashville might consider adding from the airport to the convention center, but we can see that the safety risk ratio for LR is higher than any of the modes Nashville current has. So, modeling on this aggregation is particularly valuable in helping Nashville understand the implications of adding light rail (LR) between the airport and the convention center versus adding a new commuter rail line (CR) or potentially just increase the bus service (MB). It is clear that adding light rail would likely result in greater safety risk than increase any of it's current modes of transportation.

Conclusions:

KNeighbors, Random Forest, and XGBoost were the most successful models overall depending on the aggregation. KNeighbors regularly had the lowest MAPE or MAE. The first aggregation predicts the safety risk for the nation as a whole. The second allows as the opportunity to see how the features impacting Nashville's safety risk compare to those affecting the nation. The third aggregation allows us examine the safety risk for the individual modes of transportation so Nashville can better assess the safety risk of adding a new mode of transportation and compare that to the risk associated with increasing service for modes of transportation which are already in present in the area.

Future Work:

- Discover the reason for the shifted data in modeling_df2.
- Explore the two low ridership values which has the highest SHAP values for modeling_df2.
- Use K Neighbors to further explore modeling_df2 and modeling_df3.
- Import the data about failures (needed for System Reliability) which is stored in separate annual Vehicle Maintenance xlsx files available through the NTD. Then incorporate the new data into the definition of the target to better align with the MTA's safety performance measures.
- Perform time series analysis on each aggregation to better understand the relationship of each over time. Included in this might be gaining an understanding of how the pandemic affected public transportation and whether or not the post-pandemic rebound is happening in an expected way. Are we returning to the same balance of feature importance post pandemic or are we on a new path? Is the small subset of post pandemic data a better predictor of future safety risk or is that predicting capability equivalent to the predicting capability of the whole dataset?

Recommendations for the Clients:

- Adding light rail would increase the overall safety risk for Nashville. I would not recommend recommend adding this mode of transportation.
- Adding service to an existing mode of transportation would be advisable.
  - Caution: The added vehicle revenue miles would be similar depending on the mode of transportation chosen, but not exactly the same. For example, rail would take a different path than a bus. The added vehicle revenue miles will alter the safety risk since it is part of the definition of the target.
  - The existing data shows that the safety risk for Nashville's current modes in ascending order is as follows: CR, VP, CB, DR, MB. So, Nashville should consider adding one of the modes earlier on that list to minimize safety risk.

Consulted Resources:

- National Transit Database Monthly Modal Time Series Data: https://datahub.transportation.gov/Public-Transit/Monthly-Modal-Time-Series/5ti2-5uiv/data
- National Transit Database Monthly Modal Time Series Information Page: https://datahub.transportation.gov/Public-Transit/Monthly-Modal-Time-Series/5ti2-5uiv
- Federal Transit Administration Glossary: https://www.transit.dot.gov/ntd/national-transit-database-ntd-glossary
- National Transit Database Mode Definitions: https://www.ftis.org/iNTD-Urban/modes.pdf
- Guidance for Type of Service "TN": https://www.transit.dot.gov/sites/fta.dot.gov/files/docs/NTD%202108%20FRN%20Webinar%20Presentation.pdf
- Metropolitan Transit Authority Safety Performance Measures: https://www.wegotransit.com/file.aspx?DocumentId=102
- Federal Transit Administration Safety Performance Target Guide: https://www.transit.dot.gov/sites/fta.dot.gov/files/2021-06/SPTs-Guide-v2-20210629.pdf

Packages:

- pandas
- matplotlib
- seaborn
- numpy
- sklearn
- scipy
- xgboost
- SHAP

Appreciation:

Special thanks to my mentor, AJ Sanchez, for all of his time, guidance, and encouragement. We do not walk alone.