Predicting Metro Interstate Traffic Volume

Project 2: White Paper

KJ MoChroi

Department of Data Science, Bellevue University

DSC680: Applied Data Science

Dr. Brett Werner

May 2nd, 2023

I will be predicting the traffic volume of an interstate highway using a multivariate time series dataset of metro interstate traffic volume in Minneapolis-St Paul, MN for westbound I-94. We are hoping to address the issue of optimal transit timing. Suppose there was a shipping company who often used this track of interstate. It would be helpful to be able to predict the traffic volume at any given time through this interstate section so that shipping container drivers could avoid the heaviest traffic and instead travel during times of low traffic volume.

The transportation of goods and people have always been an important part of human life. In the modern era, goods and people are shipped around the world at a more rapid pace than ever before. Because of the ever-increasing number of people in this world, there is also an increasing number of people and goods that need transporting. In the United States, one of the most common ways to transport people and goods is via the interstate highway system. This leads to periods of high traffic volumes on interstate highways, which leads to slow commutes for drivers. It is relevant to attempt to forecast the periods of high traffic volume, so that they can be avoided by people and companies who want an efficient commute.

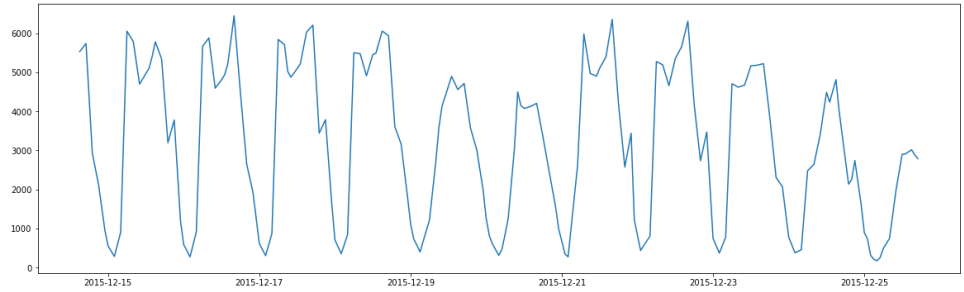
This multivariate, sequential, time-Series dataset was found on the UC Irving database website. The traffic data was provided by the Minnesota Department of Transportation and the weather data was provided by OpenWeatherMap. The dataset has 8 training features including holiday, temperature, rain, snow, clouds, weather brief description, weather longer description and one target feature which is traffic volume on westbound I-94, with 48,204 hourly readings. The data was collected from 2012 to 2018 and has integer and real number characteristics.

For data preparation, I started by removing all data instances prior to July of 2015, as there were many missing values just before this that would have interfered with modeling. I took a closer look at the traffic volume data and noted that it appears approximately sinusoidal with a

period of 24 hours as can be seen in Figure 1. It can also be seen that there is a higher traffic volume on weekdays than weekends. The next step in data preparation was to use one hot encoding to transform the categorical and descriptive variables into numeric ones so they can be used in modeling as well.

Figure 1

Distribution of Target Variable



Next, I checked the traffic volume data for stationarity using the Dickey-Fuller test available in the statsmodels library. This test indicated that the traffic volume data is stationary and that no further transformations are necessary. Then I used Granger's causality test on each of the training features with the traffic volume feature to determine which training features had an impact on traffic and which should be removed before modeling. With this method, I reduced the number of training features to about 15. Finally, I set the date time column as the index and I split the dataset into training and testing sets, with 80% of the data going into training.

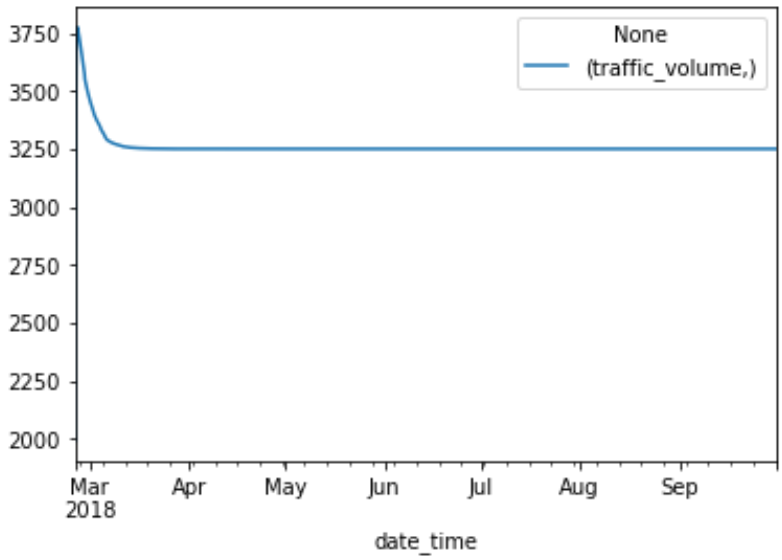
I trained three different models on this dataset, in attempts to create a model that provided value to the business problem. The first model I trained was a multivariate forecasting model called Vector AutoRegression (VAR) which took in all the features during training and made predictions on all the variables. The second model I trained was a univariate forecasting model called AutoRegressive Integrated Moving Average (ARIMA) focused only on traffic volume.

The final model I trained was not a forecasting model but rather a standard regression model called random forest regression which uses the other variables available at each timestamp to predict the traffic volume at that time.

With the VAR model, all the features are fed into the model and the model generates predictions for each of them. Even though I am only interested in predicting the traffic volume, I wanted to see if including the other features in a multivariate model could improve the predictions for traffic volume. Because I was predicting the traffic volume, I focused on the quality of the prediction on that specific feature only. I used root mean squared error (RMSE) to evaluate the model's quality and graph the predictions and the test set to compare the two visually. While I was able to generate predictions with the model, the predictions were constant, centered at the mean of the dataset as can be seen in Figure 2. While this prediction may minimize error mathematically, it did not provide value to our business problem.

Figure 2

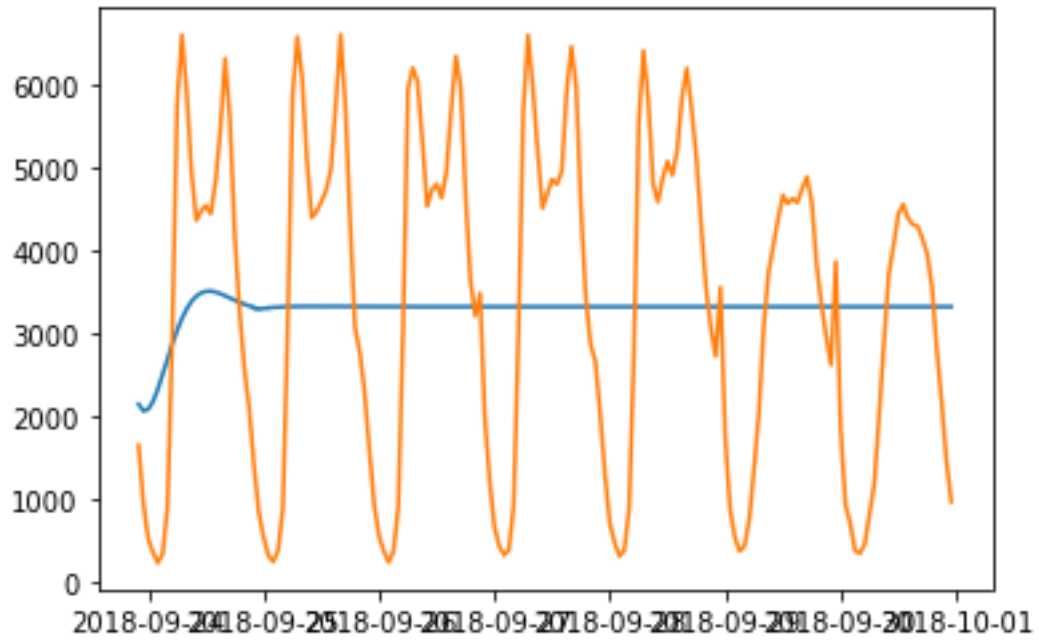
VAR Model Predictions



With the ARIMA model, I dropped all the features except for traffic volume since this is a univariate model type. I used auto-ARIMA from the pmdarima library to find the best parameters for the model, using Bayesian information criterion (BIC) as the metric by which to rank the models. The auto-ARIMA found that the best suited model was one without seasonality or difference, which makes sense given the stationarity of the dataset. Unfortunately, this model also produced a straight-line prediction centered near the mean of the dataset as can be seen in Figure 3. This model also did not provide value to the business problem.

Figure 3

ARIMA Model Predictions Overlayed Test Data

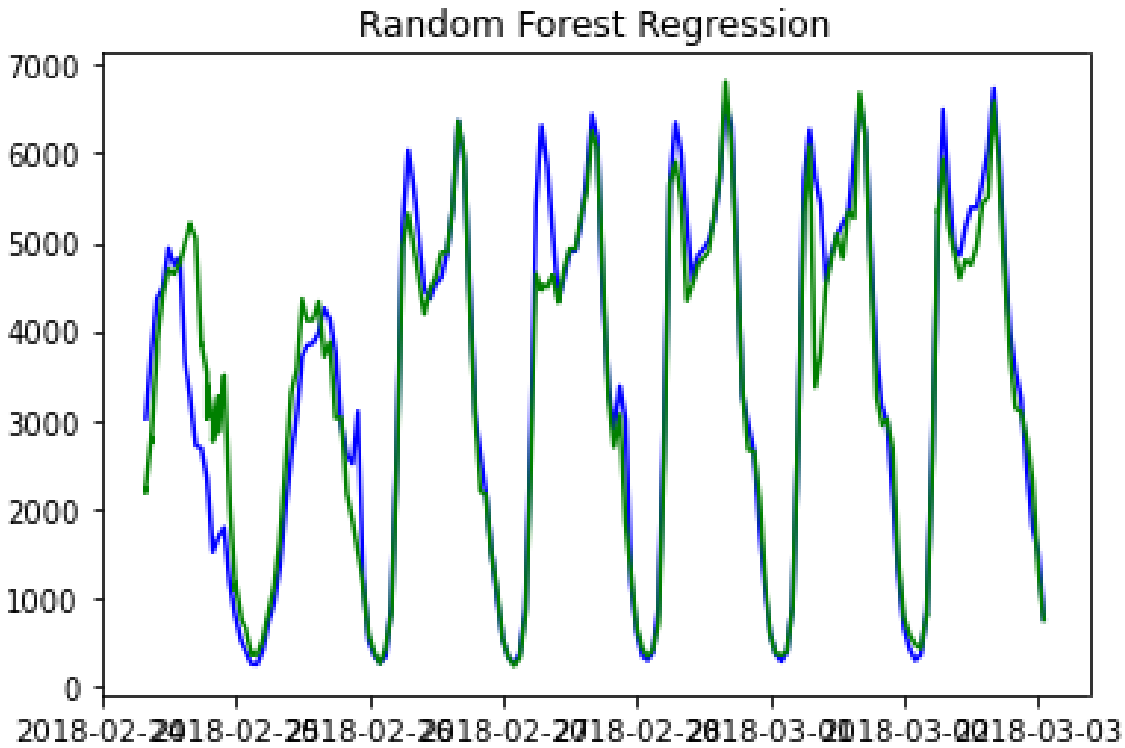


For my final model, I decided to take a different approach. I wanted to use the other features in the dataset, so I decided to pursue a standard regression model rather than a forecasting model. I still used the train-test split as before, but I fed the training features from the entire dataset into the model so that I could make predictions about the traffic volume. This

model worked phenomenally well as can be seen in Figure 4, and I was very satisfied with the results it yielded. While this is not a true forecasting model, all parameters used in making the prediction are theoretically available ahead of time, because detailed forecasts are available for weather data today. Thus, in terms of adding value to the business problem, this is the only model I created that could be used to predict traffic volume, assuming you had weather forecasts for the time you were attempting to predict.

Figure 4

RFR Model Predictions Overlayed Test Data



In conclusion, while the VAR and ARIMA models were able to be trained and make predictions, those predictions were not valuable. However, the random forest regressor was incredibly successful at modeling the sinusoidal nature of the traffic volume data. Assuming one

had access to weather predictions, this model could be used to predict the traffic volume on this interstate highway.

I made several assumptions before and during these models' training. The first assumption was that dropping nearly half of the dataset to remove missing values would not affect the integrity of the models. Furthermore, I assumed that the results of the Granger's causality test would still be relevant for the random forest regression. I also assumed that the same train-test split could be used for each of the models without issue.

The main limitation and challenge I came across came from the straight-line predictions I got from the VAR model and ARIMA model. I could not find any resources on how to improve these models to get more meaningful results. The second limitation I ran into was the time and knowledge required to train more complicated learning algorithms that could have potentially performed better than the models that I ended up using. Lastly, a limitation of this project is that it requires weather predictions to be accurate to accurately predict traffic volumes.

In the future, I think this type of modeling project could be extended to multiple highway systems to create a more systematic understanding of traffic flow throughout the United States. Having a more robust system of capturing this data could provide researchers with enough data to model the entire highway system. I would also be interested in seeing how Holt-Winters or LSTM models would perform on this data, as they handle stationary sinusoidal data particularly well.

Regarding the implementation of this model, it is important to acknowledge that this model could only predict traffic volume to the extent that weather predictions are available for that period. Meaning that this model was only able to make predictions to at most ten days out.

Implementation could be improved if the model connected to a weather API to gather the weather predictions automatically. Furthermore, if an extended forecast were required, one could pull from historical weather records to make predictions about weather based on years past and that could be used in this model as well.

Ethical considerations for this project are about the nature of transportation via highway. Transporting goods and people via the highway system is inherently unfriendly to the environment. There are other transportation methods that require less fuel and produce fewer greenhouse gas emissions, such as trains. Optimizing highway use may contribute to increased use and thus contribute to global climate change. Furthermore, requiring truck drivers to drive during less busy times of day may lead to increased safety hazards for truck drivers as well as folks on the road with them, as they may be more likely to drive at night or when they would be resting.

10 Questions from the audience:

1. What has been done here compared to the current traffic volume models out there?
2. Why did the VAR model generate a flat line prediction?
3. Why did the ARIMA model generate a flat line prediction?
4. What data could be added to the dataset to generate better results?
5. How could you change the RFR model to improve results?
6. Why an RFR model over a different type of standard regression model?
7. Why is there a dip in the top of the sine wave of the traffic volume data?
8. What were the parameters of the optimal ARIMA model?
9. What was the OOB score of the random forest regression?
10. Which features were retained after Granger's causality test?

References

Bose, E., Hravnak, M., & Sereika, S. M. (2017). Vector autoregressive models and Granger causality in time series analysis in nursing research. Nursing Research, 66(1), 12–19. https://doi.org/10.1097/nnr.0000000000000193

Hogue, J. (2019). Metro Interstate Traffic Volume Data Set. UCI Machine Learning Repository. Retrieved April 15, 2023, from https://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume

Singh, A. (2023, April 13). Multivariate time series analysis with python for forecasting and modeling (updated 2023). Analytics Vidhya. Retrieved April 15, 2023, from https://www.analyticsvidhya.com/blog/2018/09/multivariate-time-series-guide- forecasting-modeling-python-codes/