Harvard Extension Data Science

Dynamic Modeling and Forecasting in Big Data

Instructor: William Yu

Assignment 3

- Assignment submission:
  (1) A R mardown/script (or Jupyter notebook if you prefer).
  (2) A word file to summarize the results, discussions, and visualizations if any.

Part A: A Practice to Enhance Prophet's Prediction Accuracy

- In the class with H03c_prophet.R script, we learn Facebook's Prophet package, which I think is a great tool to detect and forecast various seasonality in the univariate time series data in particular for data with higher frequency.
- However, I think its detection of trend (and cycle) component could be problematic. Therefore, I suggest when you use it, you can use its seasonality forecast, such as weekly, yearly, etc but be cautious or drop its trend forecast and use your own trend/cycle forecast.
- Here let's make some effort to see if we can make Prophet's trend forecast a bit more reasonable.
- Let's revisit the data of W09e_tsa.xlsx, in which the train (in-sample) data is up to 6/30/2024 and the test (out-of-sample) data is from 7/1/2024 to 9/19/2024. In the H03c script, we have trained the m11 model and made out-of-sample forecast – forecast11a. But it seems the forecast is not ideal. The reason is its trend forecast, which is forecast to increase over time.
- One possible reason is that it confuses the low air travel in 2020/2021 and the recovery of air travel in 2022 as an upward sloping trend. One possible remedy is to add the Covid period as a special "holiday" or dummy variable into the model. Another possible remedy is to use some variable as "add_regressor" into the model, such as daily new Covid infection/death, etc.
- *Hint: for instance, you can try the following (or try different duration):*
  - holiday = 'covid',
  - ds = seq(as.Date('2020-3-15'), to=as.Date('2022-3-1'), by='days'), …
- Calculate the root mean squared error (RMSE) for m11 and your new model in the test set (7/1/2024-9/19/2024). Hint: rmse11 = sqrt(mean((forecast11a$yhat-tsa_test$travel_test)^2))
- You can also try adding US holidays to see if it improves testset errors.
- *Take-Aways:* (1) Prophet is not good at identifying a correct trend/cycle component. (2) The sample (2019 to 2024) is not long enough for us to figure out a correct trend.

Part B: Model/Features Selection

Part B is optional for students pursuing undergraduate credit.

- In Assignment 1 Part B, we learn to use a linear model to predict COVID-19 cumulative death rates (deathp) on January 23, 2021 and on December 31, 2022, respectively. Let's call it the baseline model.
- However, we have not used all the variables in the collected dataset (New2.csv). Now let's pretend we don't have much prior knowledge of these variables.
- New Model 1: You need to use some model/feature selection methods to decide what variables to be included in the linear regression model, for example, stepwise methods (See answer key of Assignment 1, AR01.R).
- *Note:* before you work on the variable selection, you would like to do some work on the dataset. (a) You might need to remove one of the two highly correlated variables (say correlation >0.9) by checking the simple correlated table. (b) You can use VIF to remove variables which could cause multicollinearity problem.
- New Model 2: If you know how to run a random forest model (or a xgboost model), also give it a try.
- Use deathp on 1/23/2021 as the dependent variable (y) in the trend set to train the models and use deathp on 12/31/2022 as the dependent variable (y) in the test set and calculate the root mean squared error (RMSE) in the test set. Hint: rmse = sqrt(mean((y_prediction_test - y_test)^2)).
- Compare the test RMSE among the baseline model, new model 1 and new model 2 and briefly discuss them.
- Note: I don't have a right/correct model or answer to this question. This question is meant to be an open-ended exercise. Feel free to try anything.