Time series Analysis and Modeling

DATS 6313

Final Term Project (FTP)

The objective of FTP is to apply the course learning objectives to a real dataset for modeling & prediction. The required software for the FTP is python and you are allowed to use packages of interest in python to accomplish the FTP tasks. The first step in FTP is the data selection. The dataset must satisfy the following criteria:

- It must be a time series dataset.
- It must be equal sampled dataset.
- It must be a multivariate dataset.
- It must at least 5000 samples. The dataset with less than 5000 samples requires pre-authorization.
- It must be a non-classified dataset.
- If the dataset is not cleaned, you can not simply delete the missing observation due to time dependency. Please come and talk to me if you don’t know how to clean your dataset.
- It is preferable to be seasonal and trended.
- Update the provided google excel sheet with the selected dataset as soon as possible. The deadline to select term project dataset is by 3/15/2023.
- https://docs.google.com/spreadsheets/d/1kK6Nrlvx48vc8guyT4tc8jiqqLMs84nW3rIro-tF7-o/edit#gid=0
- If two students select the same dataset, the dataset will be assigned to the first student and the second student needs to pick another dataset. First come first serve.

There are several resources available to acquire dataset i.e.

- https://www.kaggle.com/
- https://archive.ics.uci.edu/ml/index.php

A formal report and a presentation of your term project is required by the deadline. For the citation format you can use APA.

SPECIFIES

The final formal report must be typed and should contain the following sections:

1- **Cover page.**
2- **Table of content.**
3- **Table of figures and tables.**
4- **Abstract.**
5- **Introduction**. An overview of the time series analysis and modeling process and an outline of the report.
6- Description of the dataset. Describe the independent variable(s) and dependent variable:

a. Pre-processing dataset: Dataset cleaning for missing observation. You must follow the data cleaning techniques for time series dataset.
b. Plot of the dependent variable versus time. Write down your observations.
c. ACF/PACF of the dependent variable. Write down your observations.
d. Correlation Matrix with seaborn heatmap with the Pearson's correlation coefficient. Write down your observations.
e. Split the dataset into train set (80%) and test set (20%).

7- **Stationarity**: Check for a need to make the dependent variable stationary. If the dependent variable is not stationary, you need to use the techniques discussed in class to make it stationary. Perform ACF/PACF analysis for stationarity. You need to perform ADF-test & kpss-test and plot the rolling mean and variance for the raw data and the transformed data. Write down your observations.

8- **Time series Decomposition**: Approximate the trend and the seasonality and plot the detrended and the seasonally adjusted data set using STL method. Find the out the strength of the trend and seasonality. Refer to the lecture notes for different type of time series decomposition techniques.

9- **Holt-Winters method:** Using the Holt-Winters method try to find the best fit using the train dataset and make a prediction using the test set.

10- **Feature selection/elimination:** You need to have a section in your report that explains how the feature selection was performed and whether the collinearity exits not. Backward stepwise regression along with SVD and condition number is needed. You must explain that which feature(s) need to be eliminated and why. You are welcome to use other methods like VIF, PCA or random forest for feature elimination.

11- **Base-models**: average, naïve, drift, simple and exponential smoothing. You need to perform an h-step prediction based on the base models and compare the SARIMA model performance with the base model predication.

12- Develop the **multiple linear regression** model that represent the dataset. Check the accuracy of the developed model.

a. You need to include the complete regression analysis into your report. Perform one-step ahead prediction and compare the performance versus the test set.
b. Hypothesis tests analysis: F-test, t-test.
c. AIC, BIC, RMSE, R-squared and Adjusted R-squared
d. ACF of residuals.
e. Q-value
f. Variance and mean of the residuals.

13- **ARMA and ARIMA and SARIMA model** order determination: Develop an ARMA, ARIMA and SARIMA model that represent the dataset.

a. Preliminary model development procedures and results. (ARMA model order determination). Pick at least two orders using GPAC table.
b. Should include discussion of the autocorrelation function and the GPAC. Include a plot of the autocorrelation function and the GPAC table within this section).
c. Include the GPAC table in your report and highlight the estimated order.

12, 0
12, 12

14- Estimate ARMA model parameters using the **Levenberg Marquardt algorithm**. Display the parameter estimates, the standard deviation of the parameter estimates and confidence intervals.

15- **Diagnostic Analysis**: Make sure to include the followings:

  a. Diagnostic tests (confidence intervals, zero/pole cancellation, chi-square test).

  b. Display the estimated variance of the error and the estimated covariance of the estimated parameters.

  c. Is the derived model biased or this is an unbiased estimator?

  d. Check the variance of the residual errors versus the variance of the forecast errors.

  e. If you find out that the ARIMA or SARIMA model may better represents the dataset, then you can find the model accordingly. You are not constraint only to use of ARMA model. Finding an ARMA model is a minimum requirement and making the model better is always welcomed.

16- **Deep Learning Model:** Fit the dataset into multivariate LSTM model. You also need to perform h-step prediction using LSTM model. You can use tensorflow package in python for this section.

17- **Final Model selection**: There should be a complete description of why your final model was picked over base-models ARMA, ARIMA, SARIMA and LSTM. You need to compare the performance of various models developed for your dataset and come up with the best model that represent the dataset the best.

18- **Forecast function**: Once the final mode is picked (SARIMA), the forecast function needs to be developed and included in your report.

19- **h-step ahead Predictions**: You need to make a multiple step ahead prediction for the duration of the test data set. Then plot the predicted values versus the true value (test set) and write down your observations.

20- **(Bonus points +5)**: Design a dashboard (web-based app) for your FTP using python Dash & Flash. You need to show the developed web-based app is working. Please talk to me if you have question about the bonus points. The created dashboard must be deployed through GCP and be available for worldwide use.

21- **Summary and conclusion:** You should state any limitation of the final model and suggestions for other types of models that might improve performance.

22- A **separate appendix** should contain documented python codes that you developed for this project.

23- References

24- The **soft copy of your python programs** needs to be submitted to verify the results in the report. Make sure to include the dataset in your submission. Make sure to run your code before submission. If the python code generates an error message, 50% of the term project points will be forfeited.

25- Include a **readme.txt** file that explains how to run your python code. All the results in your report must be regenerated to grant the final grade.

26- The FTP is defined to be individual unless an approval is granted for collaboration. All the coding must be done individually, and it must be genuine. Copying a code from internet without proper citation will be considered as a **plagiarism** and FTP grade will be disregarded. Make sure to write your own code to avoid future complications.

27- All figures in your report must include a proper x-label, y-label, title, and a legend [if applicable]. Pick an appropriate theme or style for the plotted graphs. If you have a table inside your report, then make sure to include a proper title. Including grid is optional.

28- The final presentation is due by **Wednesday April 26th**. You will be given 10 minutes to present your term project and 5 minutes for Q/A. The presentation weighs 20% of the term project grade. You need to create a power point for your presentation.

29- The final formal report submission is due by **Wednesday**, **May 3rd** .

Upload the **final report (as a single pdf**) plus **the .py file(s)** through BB by the due date.