Time series Analysis & Modeling

DATS 6450

Term Project

Obtain some real experimental (not simulated) time series dataset from a public data base. The dataset must have at least 5000 samples with multiple features. Develop a linear model representation of the data using Python. You can use either the codes your developed in class or you can use any python packages. A formal report and a presentation of your term project is required by the deadline.

SPECIFIES

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

1- **Cover page.**
2- **Table of content.**
3- **Abstract.**
4- **Introduction**. An overview of the time series analysis and modeling process and an outline of the report.
5- Description of the dataset. Describe the independent variable(s) and dependent variable:
    a. Plot of the dependent variable versus time.
    b. ACF of the dependent variable.
    c. Correlation Matrix with seaborn heatmap and Pearson's correlation coefficient.
    d. Preprocessing procedures (if applicable): Clean the dataset (no missing data or NAN)
    e. Split the dataset into train set (80%) and test set (20%).
6- **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. You need to make sure that ADF-test is not passed with 95% confidence.
7- **Time series Decomposition**: Approximate the trend and the seasonality and plot the detrended the seasonally adjusted data set. Find the out the strength of the trend and seasonality. Refer to the lecture notes for different type of time series decomposition techniques.
8- **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.
9- **Feature selection:** You need to have a section in your report that explains how the feature selection was performed. Forward and backward stepwise regression is needed. You must explain that which feature(s) need to be eliminated and why.
10- 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: 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.

11- **ARMA (ARIMA or SARIMA) model** order determination: Develop an ARMA (ARIMA or 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- Estimate ARMA model parameters using the **Levenberg Marquardt algorithm**. Display the parameter estimates, the standard deviation of the parameter estimates and confidence intervals.
13- **Diagnostic Analysis**: You need to derive at least 2 ARMA (ARIMA or SARIMA) model and pick the best one. 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.
14- **Base-models**: average, naïve, drift, simple and exponential smoothing.
15- **Final Model selection**: There should be a complete description of why your final model was picked over base-models, holt-winters, regression, ARMA, ARIMA, SARIMA. 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.
16- **Forecast function**: One the final mode is picked; the forecast function needs to be developed and included in your report.
17- **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.
18- **Summary and conclusion:** You should state any limitation of the final model and suggestions for other types of models that might improve performance.
19- A **separate appendix** should contain documented python codes that you developed for this project.
20- References
21- 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.
22- Include a **readme.txt** file (if applicable) that explains how to run your python code.
23- The final presentation is due by **Wednesday December 9th**. 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.
24- The final formal report submission is due by **Wednesday, December 16th**.

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