

DSBA/MBAD 6211: Homework Assignment 4 (50 pts)

Instructions: This is an individual assignment. The submitted solution and answers should be your own. The data file for this homework is **gas_prod.csv**, which is to be downloaded from Canvas. The dataset has the monthly production of natural gas in millions of barrels. You are asked to use Python to perform Forecasting tasks and answer the given questions. Create a new Word document and save it as Forecasting_xxxx (where xxxx is your ninernet login name). Write your full name on the first page of the Word document. Where required, write your answers or paste screenshots in this Word document. You need to submit both the Word document and Python Code file. **Your Python code should run correctly for your assignment to be graded. Code that generates error will result in loss of points (up to a maximum of 20%)**

Variables and models naming requirements:

- Include your **name initials** to the data frame names as well as model names in your Python coding. This is required for your work to be graded.
- For instance, my initials are **CS**, and in my coding, I would name the data frames as **dfCS**, **dfCS.train**, and **dfCS.test**. I would also name the models as **ts_lmCS**, **arima_CS**, etc.

Questions

1. Generate a plot of the monthly gas production with proper labels. Paste a screenshot in the Word document. (5) Once you have read the csv data, before moving to any other step, use the following code to convert the "Month" column to a proper date column of equal monthly intervals.

```
# Assume the sequence starts in January 2004 and ends in December 2023 (12 months x 20 years)
start_date = '2004-01'
```

```
# Generate a date range starting from the start_date, with a frequency of one month,
# and length equal to the dataset
df['Date'] = pd.date_range(start=start_date, periods=len(df), freq='M')
```

2. Split the data into training and validation sets. Use the last three years of data for validation set and the remaining as the training set. Run the following forecasting models. (5)
 - (a) Linear trend
 - (b) Seasonality
 - (c) Quadratic trend and seasonality (remember that quadratic trend includes a linear trend term)

3. In the Word document, write a short paragraph for each model about the model performance using any three metrics. Which model performs best? Justify your answer. (15)
4. Take the gas_prod for any 12 consecutive months from the dataset. In the Word document, show in a table the original gas_prod values, a simple moving average forecast with 6-month rolling period, and an exponential smoothing forecast with $\alpha=0.2$. (5)
5. Build and run a forecasting model using SARIMA. Explain the process of choosing the values for p, d, and q for the SARIMA model (Use ACF and PACF plots; Do not use the automated search algorithm to pick the best SARIMA parameters). Write a paragraph commenting on the SARIMA results, including the model performance and the AR and MA parameters. (15)
6. Compare the best model chosen from step 3 with the SARIMA model. Which model is better for forecasting between the two? Justify your answer. (5)