# 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')

A graph showing a graph of gas production

Description automatically generated

1. 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)
   1. Linear trend

A screenshot of a computer

Description automatically generated

Regression statistics

Mean Error (ME) : 1.1510

Root Mean Squared Error (RMSE) : 8.2360

Mean Absolute Error (MAE) : 6.7058

Mean Percentage Error (MPE) : 0.2482

Mean Absolute Percentage Error (MAPE) : 1.6962

A graph showing the growth of the year

Description automatically generated with medium confidence

* 1. Seasonality

A screenshot of a computer

Description automatically generated

Regression statistics

Mean Error (ME) : 19.4764

Root Mean Squared Error (RMSE) : 20.5194

Mean Absolute Error (MAE) : 19.4764

Mean Percentage Error (MPE) : 4.9112

Mean Absolute Percentage Error (MAPE) : 4.9112

A graph of blue and orange lines

Description automatically generated

* 1. Quadratic trend and seasonality (remember that quadratic trend includes a linear trend term)

A screenshot of a computer

Description automatically generated

Regression statistics

Mean Error (ME) : -38.0559

Root Mean Squared Error (RMSE) : 39.1968

Mean Absolute Error (MAE) : 38.0559

Mean Percentage Error (MPE) : -9.6359

Mean Absolute Percentage Error (MAPE) : 9.6359

A graph with blue and orange lines

Description automatically generated

1. 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)

- Based on the "Regression Statistics" the best performing model would be the linear trend, having the lowest RMSE and MAPE values signifing that it had the closest prediction accuracy compared to the othe models. It's ME and MPE are also close to 0 indicating that there is little to no bias when it does its predictions.

- The linear model has statistics of Mean Error (ME) : 1.1510, Root Mean Squared Error (RMSE) : 8.2360, Mean Absolute Error (MAE) : 6.7058, Mean Percentage Error (MPE) : 0.2482, and Mean Absolute Percentage Error (MAPE) : 1.6962. This is by far the lowest of all the statistical models thus far having little bias and closest prediction accuracy.

- The seasonality has statistics of Mean Error (ME) : 19.4764, Root Mean Squared Error (RMSE) : 20.5194, Mean Absolute Error (MAE) : 19.4764, Mean Percentage Error (MPE) : 4.9112, and Mean Absolute Percentage Error (MAPE) : 4.9112. This has the second lowest RMSE and MAE meaning it has less prediction compared to linear, but higher ME, MPE, and MAPE indicating more bias and less predictive capability.

- The Quadratic Trend and Seasonality has statistics of Mean Error (ME) : -38.0559, Root Mean Squared Error (RMSE) : 39.1968, Mean Absolute Error (MAE) : 38.0559, Mean Percentage Error (MPE) : -9.6359, Mean Absolute Percentage Error (MAPE) : 9.6359. Has the lowest ME and MPE in the negative likely because of the log in the model, but has the highest of all the other values making me less inclined to use it.

1. 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)

A screenshot of a computer screen

Description automatically generated

A graph with blue and green lines

Description automatically generated

1. 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)

Results of Dickey-Fuller Test:

Test Statistic -0.664424

p-value 0.855751

#lags used 15.000000

Number of observations Used 223.000000

Critical Value (1%) -3.460019

Critical Value (5%) -2.874590

Critical Value (10%) -2.573725

dtype: float64

A graph with blue dots and numbers

Description automatically generated

A graph with blue dots and lines

Description automatically generatedA screenshot of a computer

Description automatically generated

- The AIC and BIC values are negative, indicating a good fit with the data.

- All tested values in the function are statistically significant.

- Log Likelihood is 749.954.

- The Skew of .5 is so it is right in the center for being symmetrical.

- Heteroskedasticity (H) is 0.85, which suggests that there is no significant heteroscedasticity in the model, which is a good result in terms of the assumptions behind regression analysis.

Regression statistics

Mean Error (ME) : 9.8548

Root Mean Squared Error (RMSE) : 13.3390

Mean Absolute Error (MAE) : 10.7085

Mean Percentage Error (MPE) : 2.4658

Mean Absolute Percentage Error (MAPE) : 2.6864

1. 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)

- The ME is higher in the SARIMA model, indicating a greater average error per forecast.

- The RMSE is almost double in the SAMIRA model compared to the linear model, suggesting that the SARIMA model has larger prediction errors overall, especially when

penalizing larger errors.

- The MAE is also higher for the SARIMA model, meaning the average error magnitude is larger.

- The MPE is larger in the SARIMA model, indicating a tendency to overpredict as opposed to the relatively neutral linear model.

- The MAPE for the SARIMA model is higher than the linear model, which means the prediction errors represent a larger proportion of the actual values.

- Based on these metrics, the linear model still appears to outperform the SAMIRA model, as it has lower values across all the key metrics, suggesting it provides more accurate and less biased forecasts.

- Overall the Linear model still outperforms the SARIMA, likely because of the lack of complexity in our data being a two column without massive fluctuations and more strenuous predictors.