PROGRESS REPORT 1

CAPSTONE PROJECT DATA SCIENCE AND ANALYTICAL COHORT 3

Najibah Ali

Time -Series Analysis (Arima Model)

Predicting The Usage of Household Power Consumption for Years to Come

Week: 8

Date: 9/6/2024 (Sunday)

No.	Points	Remarks
1	Do you have data fully in hand and if not, what blockers are you facing?	**First Data:** • Data on USA historical flood *records* from 1980 – 2015 • Data has many blank rows and not complete • Dataset is not relevant with initial objective • Blockers: Data does not provide solution for problem statement • Decision to change dataset • Source: here **Second Data:** • Source out data for a household power consumption for 2007 • Data is assumed to be in complete form with manageable nulls and format • Dataset is relevant with objective • Dataset is chosen for project • Source: here
2	Have you done a full EDA on all your data?	**Non-Graphical** groupby method to display mean, mode or median Column interaction
		Graphical Outliers Correlation Matrix Relationship between features Plot bar Kde displot

3	Have you begun the modeling process? How accurate are your predictions so far?	Time Series Analysis ARIMA Model 1.Prepare Data • Read Data • Index 'DateTime' • Drop all columns, except 'Global active power' • Aggregating 'DateTime' to Lower Frequency 2.Plot the Data 3.Check for Stationary • From plot • ACF and PACF plot • ADF test 4. Transform to stationary using differencing until p<0.05 5. Define the parameter (p,d,q) using: • Manual: from ACF and PACF plot (cannot read the plot) • Auto: auto_arima package (2 way) i. ARIMA (3,1,1) ii. ARIMA (1,1,1) 6.Split Data to Train-Test-Split 7.Train the data 8.Prediction on Test set 9.For future dates prediction How accurate: MAE: 4307, 3566 RMSE: 5194, 4983 MAPE: 4.034, 5.035 Ref: https://blog.devops.dev/lets-talk-about-your-first-arima-model-cbfdcba1749e https://github.com/nachi-hebbar/ARIMA-

		Temperature_Forecasting/blob/master/Temperature_Forecast_ARIMA.ipynb
4	What blockers are you facing, including processing power, data acquisition, modeling difficulties, data cleaning, etc.? How can we help you overcome those challenges?	**Problem** Data acquisition: • Irrelevant dataset • Data quality, • Format inconsistency • Outside domain Modeling difficulties: • Stationary issue, • Model selection, • Evaluation and Validation Data cleaning: • Datetime format • Numerical column issue • Handling the nulls • Correlation issue Processing power: • Old device, take too long for ADF test
		Overcoming Data acquisition: • Change topic • Munging and cleaning • Convert the format • Research Modeling difficulties: • Research using medium and towards data science, youtube, article • Go through other's project and sharing • Follow tutorial

		• Fix the dataset shape, understanding ADF • Use auto-arima and ACF, PACF plot • Try to find combination to lower the error Data cleaning: • Combine and general format, common dtype for formatted date • Find issue and address by replace • Fill the nulls with mean • Drop column that has high correlation Processing power: • Optimize by follow how the example and sharing organize the dataset • Include only important data frame
5	Have you changed topics since your lightning talk? Since you submitted your Problem Statement? If so, do you have the necessary data in hand (and the requisite EDA completed) to continue moving forward?	**Yes, I have changed my topic since the lightning talk.** **Topic changed** **Data 1**: Predicting the risk of getting flood, - It has no continuous data to be used as prediction, - no clear time-current **Data 2**: Predicting the household power consumption - It has continuous data for column that want to be predicted - Has clear time-current - Fulfill the problem statement Data 2 is a complete dataset, and the EDA is completed on the dataset.

6	What is your timeline for the next week and a half? What do you have to get done versus what would you like to get done?	Focus on modeling the prediction model (ARIMA) and handling error and accuracy issue for the model **What have to be done:** 1) Clean the data 2) Prepare data to be appropriate for analysis 3) EDA data 4) Understanding data **What would like to do:** 1) Complete the modeling stage 2) Run the prediction on testing model 3) Handling all errors 4) Find the best combinations
7	What topics do you want to discuss during your 1:1?	**Topics:** 1. Is mean the best way to fill my null ✓ 2. What other hyperparameter I can do to lower my error (kiv) 3. Why the df and df_train produce different p-value from ADF test ✓ 4. Should run ADF on original df or training set ✓ 5. How to get p and q from ACF/PACF plot 6. Why resample data to days left blank, even there are data on that date ✓ 7. The right step to modelling ✓ 8. .rolling(30).mean().plot (kiv) 9. Why it become NaN when use in def (in nb 1) ✓

Next week goal:

- Choose better parameter
- Lower the error

Note	Remarks
1. Dataset documentation	2/6/2024 (new dataset) 1. Acquire dataset 2. Understanding the dataset

	3. Check isnull, 4. Check dtypes 5. Drop all null 6. Handle dtypes 7. Change Date and Time to pd.datetime 8. Proceed to EDA and all 3/6/2024 1. Rather than drop all nulls, fill the nulls with mean 4/6/2024 1. Combine Date and Time column, and formatted to pd.datetime 2. Fill the nulls 3. Address the dtype issue 4. Import as new csv 6/6/2024 (Start modelling) 1. Load csv 2. Datetime column as index 3. Drop all columns except the predictor column