**Q1. AdEase Time Series**

Using hints except Complete Solution is Penalty free now

**Use Hint**

Ad Ease is an ads and marketing-based company helping businesses elicit maximum clicks @ minimum cost. AdEase is an ad infrastructure to help businesses promote themselves easily, effectively, and economically. The interplay of 3 AI modules - Design, Dispense, and Decipher, come together to make it this an end-to-end 3 step process digital advertising solution for all.

You are working in the Data Science team of Ad ease trying to understand the per page view report for different wikipedia pages for 550 days and forecasting the number of views so that you can predict and optimize the ad placement for your clients. You are provided with the data of 145k wikipedia pages and daily view count for each of them. Your clients belong to different regions and need data on how their ads will perform on pages in different languages.

**Dataset:**

https://drive.google.com/drive/folders/1mdgQscjqnCtdg7LGItomyK0abN6lcHBb

**Data Dictionary:**

There are two csv files given

1. **train_1.csv**: In the csv file, each row corresponds to a particular article and each column corresponds to a particular date. The values are the number of visits on that date.

The page name contains data in this format:

SPECIFIC NAME _ LANGUAGE.wikipedia.org _ ACCESS TYPE _ ACCESS ORIGIN

having information about the page name, the main domain, the device type used to access the page, and also the request origin (spider or browser agent)

2. **Exog_Campaign_eng**: This file contains data for the dates which had a campaign or significant event that could affect the views for that day. The data is just for pages in English.

There's 1 for dates with campaigns and 0 for remaining dates. It is to be treated as an exogenous variable for models when training and forecasting data for pages in English

**Concepts Tested:**

- Exploratory data analysis
- Time Series forecasting- ARIMA, SARIMAX, and Prophet

**What does "good" look like?**

- Importing the dataset and doing usual exploratory analysis steps like checking the structure & characteristics of the dataset
- Checking null values and understanding their reason.
- Understanding the page name format and splitting it to get different information.
- Separating different values from it like title, language, access type, and access origin.
- Visualizing the data and getting inferences from them
- Converting the data to a format that can be fed to the Arima model (Pivoting etc)
- Checking if the data is stationary
    - Dickey-Fuller test
- Trying different methods for stationarity.
    - Decomposition of series.
    - Differencing the series.
- Plotting the ACF and PACF plots
    - Give insights about the characteristics of the time series.
- Modeling
    - Creating and training the Arima model
    - Getting the exogenous variable and using it to train a sarimax model
    - Use facebook prophet for forecasting
- Finding a way(grid search / etc) to find the best params for at least 1 modeling approach.
- Defining functions for all of the tasks.
- Comparing results for all languages and creating inferences and recommendations from them
- The MAPE for previous batches has been in the range of 4-8%

**Evaluation Criteria (100 points)**

- Importing the dataset and doing usual exploratory analysis steps like checking the structure & characteristics of the dataset **(10 points)**
- Exploratory Data Analysis **(20 points)**
    - Separating the data
    - Analyzing and visualizing the data
    - Getting inferences
- Checking stationarity **(20 points)**
    - Formatting the data for the model
    - Dickey fuller test
    - Decomposition
    - Differencing
- Creating model training and forecasting with ARIMA, SARIMAX **(20 points)**
    - ACF and PACF plot.
    - Training the model.
    - Forecasting for different languages/regions.
    - Plotting the final results
- Forecasting with **(20 points)**
    - Facebook prophet
- Creating a pipeline for working with multiple series **(10 points)**

**Questionnaire:**

1. Defining the problem statements and where can this and modifications of this be used?
2. Write 3 inferences you made from the data visualizations
3. What does the decomposition of series do?
4. What level of differencing gave you a stationary series?
5. Difference between arima, sarima & sarimax.
6. Compare the number of views in different languages
7. What other methods other than grid search would be suitable to get the model for all languages?

Discussion forum link: https://www.scaler.com/academy/mentee-dashboard/discussion-forum/p/ask-me-anything-business-case-adease/21147