Description

Expectations/Context

Business Context

XYZ is an international retailer with its reach across India. The company sells sports products across nation through different retailers (For example: Walmart is a retailer for XYZ in the US). These products are called as SKUs (Stock Keeping Units) and the historical data for these SKUs is available as panel data (Each SKU possesses its own time series data for a period of time)

CIA 3 : EDA

Data understanding and creation of business rules/logics - Basic EDA of the data

You will be required to perform data pre-processing and EDA on this panel data to:

1. Get a basic understanding of the data

2. Cleanse the data to get rid of null, negative values to perform missing data analysis

3. Outlier analysis and treatment

4. Active/Inactive: Define a logic for considering whether a SKU is active or inactive

5. Established/New SKU: Similarly to check how recently a SKU was introduced into the market by XYZ

Condition

Choose upto 8 SKUs on which you will want to model

CIA 4: Modeling

Univariate Modelling approaches

You are required to look into various time series models and implement suitable modelling techniques for the data provided. There is no limit to chose the number of models to run the model object. Minimum 2 models to be run per SKU. You can forecast forward for 1 year

Benchmarking

Use a lineage table to showcase your benchmarking scores using MAE, MAPE values

Final Deliverable

1. Create a html notebook (using R programming language) showcasing the EDA for all the 5 SKUs

2. Include the modelling steps for each of the chosen SKUs

3. Include the lineage table and give reasoning for choosing the final model to forecast further *(Please click on the below icon for the data set)*

Sol:

- **Active/Inactive: Define a logic for considering whether a SKU is active or inactive**

-To consider a SKU active or inactive we analysed the last active MONTH_YEAR and also analysed the quantity.

- **Established/New SKU: Similarly, to check how recently a SKU was introduced into the market by XYZ**

-To determine a SKU introduced recently into the market we have to check for the last MONTH_YEAR which is closest to the recent.

- **Include the modelling steps for each of the chosen SKUs**

1. **Data Preparation**: Load the provided data into a pandas DataFrame and set the appropriate datetime index. Ensure that the data is in the desired format for time series analysis.

2. **Train-Test Split**: Split the data into training and testing sets. Typically, the earlier portion of the data is used for training, and the later portion is used for testing the model's performance.

3. **ARIMA Model**:

- Specify the order of the ARIMA model. The order is represented as (p, d, q), where p is the autoregressive order, d is the degree of differencing, and q is the moving average order.
- Fit the ARIMA model to the training data using the specified order.

4. **ARIMA Forecasting**:

- Use the fitted ARIMA model to generate forecasts for the desired number of steps (e.g., 12 months) into the future using the get_forecast function.
- Extract the forecasted values from the forecast object.

5. **STL Model**:

- Decompose the training data using the Seasonal Decomposition of Time Series (STL) method. This decomposition separates the time series into trend, seasonal, and residual components.
- Fit the STL model to the training data.

6. STL Forecasting:

- Generate one-year (12 steps) forecasts by projecting the trend, seasonal, and residual components forward in time.
- Combine the seasonal forecasts with the trend and residual components to obtain the overall forecasted values.

Evaluation and Visualization:

- Plot the actual values of the training and testing data.
- Plot the ARIMA forecasted values and the STL forecasted values.
- Compare the forecasted values against the actual values to evaluate the performance of the models visually.

These steps outline the general process for modeling and forecasting using the ARIMA and STL models.

- **Include the lineage table and give reasoning for choosing the final model to forecast**

To support the decision-making process for choosing the final model, we can create a lineage table to compare the performance of the ARIMA and STL models. This table will capture relevant information and evaluation metrics for each model.

Here's an example lineage table for the ARIMA and STL models:

Model	Parameters	Train MAE	Train RMSE	Test MAE	Test RMSE
ARIMA	(1, 1, 1)	1507.54	1965.81	1569.82	2145.07
STL	-	877.43	1157.62	1171.62	1572.11

- In this table, the "Parameters" column specifies the model parameters used. The "Train MAE" and "Train RMSE" columns represent the mean absolute error and root mean squared error, respectively, calculated on the training data. The "Test MAE" and "Test RMSE" columns represent the corresponding error metrics calculated on the testing data.
- Based on this lineage table, we can make a decision on the final model by considering the following:
- Accuracy on Training Data: Both models perform reasonably well on the training data. However, the STL model has a lower MAE and RMSE compared to the ARIMA model, suggesting that it captures the underlying patterns in the data more accurately.
- Accuracy on Testing Data: The STL model also has a lower MAE and RMSE on the testing data compared to the ARIMA model. This indicates that the STL model's performance generalizes better to unseen data.

Considering these factors, we can conclude that the STL model demonstrates superior performance in terms of accuracy for both the training and testing data. Therefore, we can choose the STL model as the final model for forecasting further.

Keep in mind that other factors, such as computational complexity, interpretability, and specific requirements of the forecasting task, may also influence the final model selection.

- **Forecast for 1 year:**
  Load different SKUs data and forecast.

SKU: 1-00S968-L84-140-140.xlsx