Time Series Multiple linear regression

Business Objective

A time series is simply a series of data points ordered in time. In a time series, time is often the independent variable, and the goal is usually to make a forecast for the future.

Time series data can be helpful for many applications in day-to-day activities like:

- Tracking daily, hourly, or weekly weather data
- Monitoring changes in application performance
- Medical devices to visualize vitals in real-time

Linear regression is widely used in practice and adapts naturally to even complex forecasting tasks. In this project, we will deal with the multiple linear regression model. The aim of the multiple linear regression is to model a dependent variable (output) by independent variables (inputs).

In this series of time series projects, we have already covered three major topics, Time Series Project to Build an Autoregressive Model in Python Build a Moving Average Time Series Forecasting Model in Python and Time Series Forecasting Project-Building ARIMA Model in Python.

In this project, we will be implementing the Multiple linear regression model on the given dataset.

Data Description

We will be using “Call_centres” data. This data is at the month level wherein the calls are segregated at the domain level as the call center operates for various domains. There are also external regressors like no of channels and no of phone lines which essentially indicate the traffic prediction of the inhouse analyst and the resources available.
There are about 130 rows and 8 columns in the dataset

- Month, healthcare, telecom, banking, technology, insurance, no of phonelines and no of channels.
- The multiple linear regression model will be built using three variables, banking (dependent variable), no of phonelines, and no of channels (independent variables)

Aim

This project aims to build a Multiple linear regression model on the given dataset

Tech stack

- Language - Python
- Libraries - pandas, numpy, matplotlib, scipy, scikit learn, gplearn

Approach

1. Import the required libraries and read the dataset
2. Data pre-processing
    - Setting date as the index
    - Setting frequency as month
3. Exploratory Data Analysis (EDA) -
    - Data Visualization
4. Check for normality
    - Density plots
    - Q-Q plots
5. Multiple linear regression model
    - Train test split
    - Train the model
    - Fit the model
    - Make predictions
    - Plot the results
6. Residual analysis
    - Remove autocorrelation with varying lag values
    - Check for the normality of the variables
    - Train and fit the model
    - Make predictions and plot the results
7. Symbolic regression model
    - Create a model
    - Train the model
    - Fit the model
    - Make predictions and plot the results

Modular code overview

```
input
  |_CallCenterData.xlsx


src
  |_Engine.py
  |_ML_Pipeline
            |_MLR.py
            |_PreprocessPlots.py
            |_SymbolicRegression.py


lib
  |_MultipleLR.ipynb


output
  |_Visualization plots(.png)
  |_symbolic_regression_model.pkl
```

Once you unzip the modular_code.zip file, you can find the following folders within it.

1. input
2. src
3. output
4. lib

1. Input folder **-** It contains all the data that we have for analysis. The following csv is used.
    - CallCenterData.xlsx

2. Src folder - This is the most important folder of the project. This folder contains all the modularized code for all the above steps in a modularized manner. This folder consists of:
    - Engine.py
    - ML_Pipeline
      The ML_pipeline is a folder that contains all the functions put into different python files which are appropriately named. These python functions are then called inside the engine.py file.

3. Output folder **-** The output folder contains all the visualization graphs. There are around 20 different plots. The symbolic regression model is saved in a pickle file. Similarly, you can save other models that can be used later.

4. Lib folder **-** This is a reference folder. It contains the original ipython notebook that we saw in the videos. The ppt used during the videos is also present here.

Project Takeaways

1. Introduction to Time series
2. Understand the basics of time series
3. Importing the dataset and required libraries
4. Pre-processing the data
5. Exploratory Data Analysis (EDA)
6. Check for the normality in data using the density plots and Q-Q plots
7. Understanding the difference between correlation and autocorrelation
8. What is a multiple linear regression model?
9. How to build a multiple regression model?
10. Training a multiple linear regression model
11. Remove autocorrelations to better train the model
12. Understand the difference between regression and autoregression
13. Understand symbolic regression model
14. Build a Symbolic regression model
15. Train a Symbolic regression model
16. Evaluate the results with performance metrics like RMSE