Time Series Gaussian Process

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

Gaussian Processes are a generalization of the Gaussian probability distribution and can be used as the basis for sophisticated non-parametric machine learning algorithms for classification and regression. Gaussian probability distribution functions summarize the distribution of random variables, whereas Gaussian processes summarize the properties of the functions, e.g., the parameters of the functions. Gaussian processes can be used as a machine learning algorithm for classification predictive modelling

We have already covered the concepts of Autoregression modelling, Moving Average Smoothing techniques, ARIMA model and Multiple linear Regression.
In this project, we will be implementing the Gaussian model on the given dataset.

Data Description

The dataset is “Call-centres” data. This data is at month level wherein the calls are segregated at domain level as the call centre 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.
The total number of rows are 132 and number of columns are 8:

- Month, healthcare, telecom, banking, technology, insurance, no of phonelines and no of channels.

Aim

This project aims to build a Gaussian model on the given dataset.

Tech stack

- Language - Python
- Libraries - pandas, numpy, matplotlib, seaborn, sklearn, scipy

Approach

1. Import the required libraries and read the dataset
2. Perform descriptive analysis
3. Data pre-processing
    - Converting date to numeric
    - Setting date as index
4. Exploratory Data Analysis (EDA) -
    - Data Visualization
5. Check for normality
    - Density plots
    - QQ-plots
6. Gaussian process
    - Initiate kernels
    - Perform train-test split
    - Create a Gaussian process regressor model
    - Fit the model
    - Generate predictions
    - Plot the results
7. Difference
    - Create a residual column (difference)
    - Check for normality
    - Train test split
    - Initiate kernel
    - Create a gaussian model
    - Fit the model
    - Generate predictions on test data
    - Plot the results

Modular code overview

```
input
  |_CallCenterData.xlsx


src
  |_Engine.py
  |_ML_Pipeline
            |_Gaussian_Stationary.py
            |_Gaussian_Trend.py


lib
  |_Gaussian Process.ipynb


output
  |_Visualization plots(.png)
  |_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 10 different plots. The best model is fitted in a pickle format.

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. How to perform data preprocessing for time series data?
5. Exploratory Data Analysis (EDA)
6. What is Normal distribution?
7. What is the Gaussian process?
8. What is meant by Gaussian Kernel?
9. What is the Gaussian process for time series?
10. How to build a Gaussian model?
11. Fit and train a Gaussian model on the training data
12. How to perform predictions on the test data?
13. Hypermeters in Gaussian process
14. How to plot the results?
15. Evaluate the results with R-squared and MSE scores