Internship Project – Adewale Adeagbo

Topic: Monitoring for Time Series forecasting models

Context: Time series forecasting models are a family of Machine Learning (ML) models whose task is to use past values of a variable of interest to generate accurate forecasts of the variable's future values.

Example: Given a set of time series of historical values (ground truth) of various BT metrics up to today, we are interested in predicting the values of each of the metrics for the next 14 days. In production, we run this process every week on Sunday, and each week we generate the forecasts for the following 14 days.

Because of business processes, the ground truth data is available to us with some delay. While this is not important when making the predictions, it is important when it comes to evaluating them. For the purpose of this project, we assume that historical data has a delay of 7 days (so every Sunday when the predictions are generated, historical data is available only until the previous Sunday included).

The historical data can be considered a table with the following structure:

Series Name	Date	Historical Value

This data is fed into the model and used to generate predictions. In this project we are not concerned with the nature of the forecasting model, which we assume being a black box. Once the predictions have been generated, the output of the model can be considered a table with the following structure:

Series Name	Prediction Date	Predicted Value	Model Execution Date

Part 1: You are asked to come up with a mechanism to regularly assess the goodness of the predictions of the model against the ground truth, and to report 4 or 5 metrics of interest. You are provided with 2 CSV files which represent the tables above.

The requirements of the deliverables are the following:

- Whilst you can experiment in any language that you wish, the final product must be a SQL solution. The solution should create a new table and repeatedly append results to it so that the entire history of model performance can be audited. Feel free to use any SQL product or dialect you are familiar with.
- The solution must be reproducible (i.e. running the solution on the same set of data multiple times should lead to the exact same results).
- You should identify and monitor standard metrics that are often used in the context of Time Series forecasting.
- You should take care of handling any numerical errors that arise when calculating the metrics.
- You should take care of handling any missing or late data when calculating the metrics.

- You should write a script to automate the solution to run every Monday at 04:00 UTC time using the latest available historical data.
- In addition to the raw metrics (that are useful to Data Analysts and Data Scientists), you should think how to effectively summarise the results in a clear and concise manner for the business stakeholders.

Part 2: You are asked to come up with a mechanism to monitor the historical data to understand whether its underlying distribution is changing over time. If the distribution of the data changes over time, it might be necessary to re-train the model to ensure that the performance do not degrade.

We will discuss this once you've done some progress on the first point.