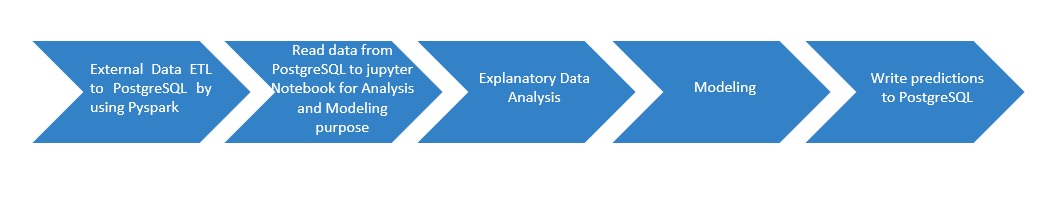
**Due:** 05.01.2025 – 23.59

**Presentation:** 10.01.2025

**Instructions:**

You should make an end to end analytics project that includes the following steps.

****

Your task is make a Daily buy sell prediction model with minimum one stock and one index. Stock and index list is above.

**Stock List:**

|  |  |  |
| --- | --- | --- |
| Stock | | |
| ID | Name | Ticker |
| 1 | Apple Inc | AAPL |
| 2 | NVIDIA Corp | NVDA |
| 3 | Microsoft Corp | MSFT |
| 4 | Broadcom Inc | AVGO |
| 5 | Meta Platforms Inc | META |
| 6 | Amazon.com Inc | AMZN |
| 7 | Tesla Inc | TSLA |
|  |  |  |
| Index | | |
| 1 | S&P 500 | ^GSPC |
| 2 | NASDAQ 100 | NQ=F |
| 3 | Russell 2000 | RTY=F |
| 4 | Dow Jones Industrial Average | ^DJI |

Stock are selected from the Nasdaq 100 companies.

Prediction must be one over one of these strategies:

1. Buy-Hold-Sell
2. Buy-Sell
3. Next day price

Based on your analysis and model evaluations, clearly state the approach and the models you chose for prediction. Provide an interpretation of the performance metrics to justify your model selection. You can use *Measurement methods* document.

You must show minimum these 4 methods *Return on Investment (ROI), Net Profit, Win Rate (Accuracy), Beta* to measure model success.

All actions will have **1/1000 fee.** You must apply that fee for calculating net profit.

**We will measure the final model performance with following method:**

Assume, you have 100.000 $ at the beginning of the prediction period.

Your model gain:

With your **stock prediction model** you will take actions; as a result you will have 150.000 $ or worth of stock shares at the end.

Things to consider when developing a **stock prediction model**:

* You should run the model every day. You will make a buy, sell or hold decision depending on one of the 3 strategies you choose above. Remember! Actions must be made based on the next day open price.
* There will be no fee on the days you do not take action. You are expected to reflect the fee for every action you take.

Stock market gain:

If you have bought the stock at the first day and never sell till the end of the period, you would have 125.000 $.

**Your success will be 150.000 / 125.000 = 1,2**

**That ratio will be our final score = 1,2**

**Train Period:** 0/0/0 -01/01/2020

**Test Period:** 01/01/2020 - 01/01/2023

**Validation Period:** 01/01/2023 - 01/01/2024

**Prediction Period:** 01/01/2024 - 31/12/2024

\* The data for the period to be predicted should not be included in the training data.

**Tools to be used:**

1. PySpark: For data processing and ETL tasks.

2. PostgreSql: To store the data and save prediction results.

3. Python & Jupyter Notebook: For data analysis, modeling, and visualization.

4. Yahoo Finance API/LIB [1]: For fetching stock data.

**Step 1:** **Fetching the Data (Using Yahoo Finance API)**

You are required to fetch historical stock data for the 10 Nasdaq companies using the Yahoo Finance API. The API can be used to download relevant stock data, such as opening price, closing price, trading volume, etc. for the last ten years. Add. Info: Analyst recommendation, Market Cap, Finansal Data (Income Statement, Balance Sheet, Cash Flow) and other related ticker data like gold, UKOIL, VIX, DXY, US2Y, US10Y .

Once the data is fetched, you will use **PySpark** to process the data and store it in **PostgreSQL.** The goal here is to create an automated ETL pipeline that extracts stock data from **Yahoo Finance** and loads it into **PostgreSQL**.

**Step 2: Data Analysis and Feature Engineering**

After storing the data in PostgreSQL, you should retrieve it into a **Jupyter Notebook** for further analysis. The following tasks must be performed:

**1. Exploratory Data Analysis (EDA):**

Analyze the data to identify trends, patterns, and key features. Provide descriptive statistics and summary metrics. Feel free to be creative and think outside the box in this step.

**2. Feature Engineering:**

Generate meaningful features from the available stock data. These features should be engineered in a way that helps improve the predictive models.

**3. Data Visualization:**

Use appropriate visualizations to illustrate key findings from your data analysis. Ensure that your visualizations provide insights into stock behavior over time.

**Step 3: Modeling**

Using the features generated from your analysis, you should build predictive models to forecast stock decisions for 10 selected Nasdaq stocks. The focus should be on creating a robust and accurate model using the following techniques: (You should train the model using at least one technique.)

**1. Time Series Modeling:**

Consider using time series models such as ARIMA or Prophet to predict future stock prices based on historical data.

**2. Machine Learning Models / Deep Learning Models:**

Additionally, explore regression-based models such as Random Forest, XGBoost, or any other suitable techniques for predicting stock decision (buy, sell, hold).

During the modeling process, you should experiment with different approaches and evaluate the models using the following metrics:

- Mean Absolute Error (MAE)

- Root Mean Squared Error (RMSE)

- R-squared (R²)

Select the best-performing model based on these metrics and save it for further use.

**Step 4: Writing Predictions to PostgreSQL**

Once you have selected your best model, you are required to create a prediction pipeline in **PyCharm** that:

1. Reads stock data from **PostgreSQL**.

2. Applies the necessary data preparation steps.

3. Uses the trained model to make predictions on a test dataset.

4. Writes the predictions back into **PostgreSQL**.

This step ensures that your entire project workflow from data extraction to predictions is automated and reproducible.

**Deliverables Summary:**

* **ETL Process (PySpark with PostgreSQL):**

Document the process of fetching stock data and loading it into **PostgreSQL**. You should describe how you managed the ETL process, including any data cleaning steps.

* **Exploratory Data Analysis (EDA) & Visualization:**

Provide detailed analysis of the stock data, including visualizations and feature engineering steps. This should include any statistical analysis performed and insights drawn from the data.

* **Modeling:**

Submit a report detailing the models you tested, their performance metrics, and the rationale for selecting your final model.

* **Predictions in PostgreSQL:**

Ensure that the predicted stock decisions are written back into **PostgreSQL**. Provide clear documentation of this step in your final report.

**Notes will be given to jupyter notebook, presentation of results and final scripts written in Pycharm.**

**Grading Criteria:**

|  |  |
| --- | --- |
| **Task** | **Points** |
| ETL to PostgreSQL | 25 |
| EDA - Visualization | 10 |
| Feature Extraction | 20 |
| Modeling | 15 |
| Final scoring | 15 |
| Presentation | 15 |

**References:**

[1] <https://pypi.org/project/yfinance/>

[2] <https://finance.yahoo.com/>

[3] <https://www.postgresql.org/download/>

[4] <https://dbeaver.io/download/>

**In case you have questions:**

[arcacakmakoglu@gmail.com](mailto:arcacakmakoglu@gmail.com)

batuhankose@gmail.com