Master Guide Documentation

Andres Camilo Viloria Garcia
aviloriagarcia00@mylangara.ca
andresvdata@gmail.com
Data Analytics
Langara College
Vancouver, BC

Diana Ortiz
dortizmontes00@mylangara.ca
dianasayuriortizm2021@gmail.com
Data Analytics
Langara College
Vancouver, BC

Introduction
1. Overview
2. Data Sources
3. Data Quality
4. MPE Metric Deep Dive
5. XGBoost Hyperparameter Tuning
6. Predicted vs Actual csv's and Graphs
7. Appendices

Introduction

The Master Guide Documentation serves as a comprehensive record of the entire project's Phase 5 for the Capstone project “Stock Forecasting”, encompassing insights data understanding, feature engineering, data quality, XGBoost hyperparameter tuning, and comparison of performance between models in visual representations. To access this documentation, datasets, and supplementary materials please click here. Inside this folder you will find seven different folders, each serving a specific purpose or containing specific types of files.

1. Overview

Phase 5 evaluated data quality limitations, explored hyperparameter tuning for XGBoost, and analyzed evaluation metrics. This phase aimed to enhance the precision and effectiveness of machine learning models by analyzing and gathering a deeper understanding of the data and model functioning. The final deliverables will serve as valuable contributions to the research team and provide insights into potential improvements moving forward.

1.1 Scope Statement : The scope document contains the objectives, the expected deliverables and a list of milestones with the respective dates. This document will provide a guideline of the work achieved by this project phase.
1.2 Phase 4 Model Comparison Initial Results : This was the initial model comparison file where all the metrics for individual stocks were saved into a table. The most relevant sheets are : 1 Day Phase 4 and 5 Days Phase 4. It important to note that in this document there are no recorded metrics for portfolio.
1.3 Phase 5 Model Comparison: This file presents the updates implemented in Phase 5, where all metrics for all models, the training data size (both 60 and 240 days), and metrics and stocks information (including individual and portfolio data) were completed for the corresponding comparison.

2. Data Sources

In previous project phases, the procedures used to transform the data were not well-organized or properly documented with comments. To address this, in phase 5, we have gathered all the procedures following a logic process into a single folder. This will make it easier for future students to understand and follow the data transformation process.
This folder contains the main dataframe, the demographics and stock prices for each stock in a csv, as well as the csv for commodities.Additionally there is a jupyter notebook containing the codes used for data extraction and hot-encoding process required to obtain the final dataframe used in the models.

This folder contains 1 folder and three files:

- **Folder 2.0** csv's : Folter that contains all the csv needed to get the final dataframe used in the models.
  - **2.0.1 Stocks:** stock prices for each stock in a csv
  - **2.0.2 commodities :** stock commodities for each stock in a csv
  - **2.0.3 df.csv**: After performing the feature engineering process in the Data Extraction Merging.html file, we have obtained the final table df.csv. This table contains the raw data that will be used as input for the models.
- **Files 2.1, 2.2 and 2.3 - Data Extraction Merging:** The feature engineering process is being conducted in a Jupyter notebook (2.3). To enhance readability, the code has been converted into two additional formats: PDF (2.2) and HTML (2.1). This allows for easier access and comprehension of the feature engineering steps.

  To obtain the main data frame Yahoo Finance was used as a primary data source, extracting both demographic values and stock prices. Additionally a group of csv's were provided by Albert that include the prices of commodities and Bonds, and lastly a hot-one encoding processes was done to help identify relevant dates on the data such as holidays, post holidays , monday mornings and friday afternoons.

  Portfolio of 29 stocks:

  - AAPL: Apple Inc.
  - XOM: Exxon Mobil Corporation
  - IBM: International Business Machines Corporation
  - KO: The Coca-Cola Company
  - CVX: Chevron Corporation
  - BA: The Boeing Company
  - PFE: Pfizer Inc.
  - MSFT: Microsoft Corporation
  - T: AT&T Inc.
  - WMT: Walmart Inc.
  - F: Ford Motor Company
  - NFLX: Netflix Inc.
  - JPM: JPMorgan Chase & Co.
  - MCD: McDonald's Corporation
  - GE: General Electric Company
  - NVDA: NVIDIA Corporation
  - JNJ: Johnson & Johnson
  - BAC: Bank of America Corporation
  - C: Citigroup Inc.
  - AMZN: Amazon.com Inc.
  - INTC: Intel Corporation
  - CSCO: Cisco Systems Inc.
  - TSLA: Tesla Inc.

- GOOGL: Alphabet Inc. - Class A
- AMD: Advanced Micro Devices Inc.
- BABA: Alibaba Group Holding Limited
- VZ: Verizon Communications Inc.
- DIS: The Walt Disney Company
- META: Meta Platforms, Inc.

3. Data Quality

Within this folder, you will find a report and a Jupyter notebook dedicated to data quality analysis. This analysis delves deeper into the exploration of the data utilized in the models, with a particular focus on identifying and addressing null values and outliers.

You will find 1 report (3.1) and two files with the same information but different format: HTML (3.2) and Jupyter notebook (3.3) as shown as follows:

- **3.1** Data Quality Report: This report summarizes the key insights from an Exploratory Data Analysis (EDA) (files 3.2 or 3.3) and focuses on data quality assessment. It identifies missing values, non-historical data, and negative values in the dataset. Recommendations include further investigation, imputation techniques for missing values, removal of non-historical data, and appropriate handling of negative values. Detailed analysis and Python files are provided for future improvements.

- **3.2 and 3.3** - EDA: This notebook explores the dataset and aims to understand in a more detailed way the data resulted in the feature engineering process. This file includes the following insights:
    - **Variables Details:** This provides a definition of the columns or variables that are considered in the dataset.
    - **EDA and validation** : A comprehensive guide on the meaning of demographic variables, as well as an analysis on missing values and outliers can be found along with cross examination to validate the null values with other sources such as YChart.

4. MPE Metric Deep Dive

In this folder a report of a discrepancy between the MPE metric calculations in the published research paper and the way the MPE metric had been calculated in previous phases.

This folder contains two files:

- **4.1 MPE Metric Deep Dive**: This file is the report regarding the discrepancy.

- **4.2 MPE:** This file shows the difference between the 2 different ways to compute the MPE based on what was mentioned in 4.1 file..

5. XGBoost Hyperparameter Tuning

This folder contains all the necessary information to understand this stage of the process. We proceed with the tuning of two hyperparameters, namely "max_dept" and "n_estimators," for XGBoost. For more detailed information and insights, please refer to the file "5.1 Hyperparameter Tuning.pdf".

6. Predicted vs Actual csv's and Graphs

This folder contains the deliverables related to the objective of comparing the prediction behavior among different models. The comparison was conducted using both the entire portfolio data and individual stock data. It includes the relevant outputs and documentation highlighting the performance and insights gained from this comparative analysis.

This folder includes:

- 1 folder : This folder contains all the data needed to generate the graphs (6.1, 6.2, 6.3)
- 3 files:
  - A report that summarizes all the graphs obtained : 6.1 Report of Comparison Graphs for Stock Predictions.pdf
  - the code to generate the graph in two different formats:
    - 6.2 Development of Comparison Graphs for Stock Predictions.html
    - 6.3 Development of Comparison Graphs for Stock Predictions.ipynb

7. Appendices

This folder includes two folders:

7.1 Models Jupyter Notebooks: This folder contains the codes needed to run the prediction of all models including Random forest, MLP and XGBoost.

7.2 Minutes: During Phase 5 of the project, a total of 13 meetings were held. Each meeting was documented to record the discussions that took place, the attendance of participants, and the plans outlined for the subsequent meeting. These records provide a comprehensive overview of the progress and decision-making throughout this phase of the project.

Conclusions

- In previous phases, it was difficult to quickly find the necessary information as the project progressed. This phase focused on gathering and explaining relevant information to make it easier for future students to access and understand. The “data source” folder served this purpose.

- We successfully documented some insights in the current data that might be improved in the future phases, such as issues with data quality and null value presence. We suggested improvement actions in the document for more accurate results in the future. One such suggestion was to review the consistency of demographic values with the corresponding dates to enhance the overall accuracy and reliability of the results obtained from the data.

- In the hyperparameter tuning process, it was noticed that changing the "n_estimator" parameter didn't have a big impact on the model's performance. However, adjusting the "max_depth" parameter made a noticeable difference. Different values for "max_depth" affected how accurate the predictions were, as shown by the MAPE and MPE values. This means that from the ranges of values selected for the hyperparameter tuning, choosing the right "max_depth" value is more important for improving the model's performance, while changing "n_estimator" had less of an effect.

- XGBoost showed better performance in predicting both individual stocks and portfolios. There were no significant differences in performance based on visual observations and metric values. This means that training the model with the entire portfolio allows us to estimate the values of both individual stocks and the overall stock simultaneously, without the need for separate processes.

Suggestions

- The performance graphs in folder 6 revealed some sudden peaks in the predictions. To better understand these occurrences, it's recommended to conduct further investigation.
- Applying statistical methods will help determine if there are significant differences in predicting individual stocks and the overall portfolio. This will provide more reliable insights.
- In the process of XGBoost hyperparameter tuning, it is recommended to consider evaluating other variables that have been researched and may impact the perfomance. Some of these parameters include: eta, subsample, sampling_method, Colsample_bytree,colsample_bylevel,colsample_bynode, Lambda, Alpha and eval_metric.
- For the XGBoost model, it's important to analyze the stability of prediction values specifically for the stocks F and WMT. By conducting a detailed analysis, we can uncover the reasons behind any inconsistencies or fluctuations in the predictions. It's recommended to further investigate these cases to gain a better understanding.