Final project
Time-series data and application to stock markets

Introduction

Social and economic environment are constantly changing over time, data analysts must be able to assess and predict the effects of these changes, in order to suggest the most appropriate actions to take.

Time-series data is made up by dynamic data collected over time. Thus, it requires to have appropriate forecasting techniques to support business, operations, technology, research, etc.

Objective

The project aims at getting students familiar with time-series data and its applications by analyzing and deriving practical solutions using predictive analytics for stock markets.

This is a free-style project, meaning you have the freedom to decide how to programmatically formulate and solve problems. You'll independently determine the best approach for different tasks, justifying your choices with sound reasoning or experimental results. The evaluation will focus on the accuracy and justification of your solutions.

Project requirements

This project comprises six main tasks, each further broken down into subtasks. The following table provides a detailed overview of the project requirements.

# Task	Name	Requirements
Task 1 (15%)	Nasdaq stock price prediction (Nasdaq dataset)	Nasdaq stock price prediction: • **Task 1.1** (5%, Nasdaq multi-feature extension) Modify the demo code so that the Nasdaq stock price prediction model utilizes multiple features, such as Low, High, Open, Close, Adjusted Close prices, and Volume, rather than relying solely on one feature (Open price). • **Task 1.2** (5%, Nasdaq k^{th} day forecast): Update the demo code to enable the Nasdaq stock price prediction model to forecast the price on the k^{th} day in the future, such as the 3rd day or the 7th day ahead, instead of just the next day. • **Task 1.3** (5%, Nasdaq k days forecast): Extend the demo code so that the Nasdaq stock price prediction model is capable of predicting k *consecutive* days ahead. For example:

		○ If k=3, the model should predict the stock prices for the next three days (i.e., the 1st day, 2nd day, and 3rd day ahead). ○ If k=7, the model should predict the stock prices for the next seven days (i.e., the 1st day, 2nd day, 3rd day, 4th day, 5th day, 6th day, and 7th day ahead). You have to figure out the following points: • Training / Validation / Test split conforming to time-series data. • Cross-validation conforming to time-series data. • Time window, e.g., one-month training and one-week testing. • Company filtering, e.g., those with at least 120 historical data points, companies in certain stock exchanges, companies in certain industries, etc.
Task 2 (15%)	Vietnam stock price prediction (Vietnam dataset)	Vietnam stock price prediction: • **Task 2.1** (5%, Vietnam multi-feature extension): Modify the demo code so that the Vietnam stock price prediction model utilizes multiple features, such as Low, High, Open, Close prices, and Volume, rather than relying solely on one feature (Open price). • **Task 2.2** (5%, Vietnam k^{th} day forecast): Update the demo code to enable the Vietnam stock price prediction model to forecast the price on the k^{th} day in the future, such as the 3rd day or the 7th day ahead, instead of just the next day. • **Task 2.3** (5%, Vietnam k days forecast): Extend the demo code so that the Vietnam stock price prediction model is capable of predicting k *consecutive* days ahead. For example: ○ If k=3, the model should predict the stock prices for the next three days (i.e., the 1st day, 2nd day, and 3rd day ahead). ○ If k=7, the model should predict the stock prices for the next seven days (i.e., the 1st day, 2nd day, 3rd day, 4th day, 5th day, 6th day, and 7th day ahead). You have to figure out the following points: • Training / Validation / Test split conforming to time-series data.

		• Cross-validation conforming to time-series data. • Time window for training and testing. • Company filtering, e.g., those with at least 120 historical data points, companies in certain stock exchanges, companies in certain industries, etc. • Is it good to make use of additional Vietnam data such as dividend history, industry analysis, financial ratio.
Task 3 (20%)	Trading signal identification for Vietnam market	Vietnam trading point prediction: • **Task 3.1** (10%, Buying signal identification): Build a model to identify potential entry points for buying stocks, i.e., the model outputs an indication that can be a score or a probability suggesting it might be a good time to buy. Justify the way you build the model. • **Task 3.2** (10%, Selling signal identification): Build a model to identify potential entry points for selling stocks, i.e., the model outputs an indication that can be a score or a probability suggesting it might be a good time to sell. Justify the way you build the model. You have to figure out the following points: • Training / Validation / test split conforming to time-series data. • Cross-validation conforming to time-series data. • Time window for training and testing. • Company filtering, e.g., those with at least 120 historical data points, companies in certain stock exchanges, companies in certain industries, etc. • Is it good to do the manual feature engineering such as Simple Moving Average (SMA), Moving Average Convergence Divergence (MACD), Relative Strength Index (RSI), etc., to determine the training points?
Task 4 (30%)	Portfolio composition, risk management and portfolio optimization for Vietnam market	Vietnam portfolio and risk management: • **Task 4.1** (10%, Portfolio composition): Select a list of profitable Vietnamese companies to include in the portfolio. Evaluate the projected profit potential of each chosen company within a designated time frame. Explain the methodology used for portfolio optimization and profit estimation. • **Task 4.2** (10%, Risk management): Identify risky companies that should be excluded from the

		portfolio. Explain the risk scoring methodology in your risk scoring model. • **Task 4.3** (10%, Portfolio optimization): Outline a method to determine the optimal investment allocation within the portfolio, i.e., maximize expected return while minimizing risk. Justify your chosen approach for portfolio construction. You have to figure out the following points: • Training / Validation split conforming to time-series data. • Cross-validation conforming to time-series data. • Time window for training and testing. • Company filtering, e.g., those with at least 120 historical data points, companies in certain stock exchanges, companies in certain industries. • What should be the list of companies to hold if investors are risk-taking or prudent? Note: Building model for Vietnam market is more challenging, therefore more preferable.
Task 5 (30%, Extra credit)	Industry standard for deployment and ease of use	Industry standard for deployment and ease of use: • **Task 5.1** (10%, Model deployment): Deploy the prediction models as API services. Some keywords to research are Tensorflow Serving (TFServing), REST APIs, gRPC. • **Task 5.2** (10%, Model as SaaS): Deploy the prediction models as a web-based Software-as-a-Service (SaaS). Some keywords to research are TensorflowJS, Superset / Tableau / PowerBI. • **Task 5.3** (10%, AI automation workflow) Design an engineering flow to automate the tasks. Some keywords to research are SQL, MongoDB, Airflow, Airbyte, dbt.
Task 6 (20%)	Report	The report: • should describe the journey about your experiments, observations, findings and conclusions from Task 1 to Task 5. • must be at least **1,500 words** in length. Tips to write an effective report: • Telling story about experiment failures is as valuable as about experiment successes.

		• The report must include instructions, if necessary, on how to run your code, such as any external libraries that need to be installed.

Data folder structure

- **Nasdaq data**:
  The Nasdaq data folder is organized as follows:
  - ./csv: folder containing historical stock price data.
- **Vietnam data**:
  The Vietnam data folder is organized as follows:
  - ./stock-historical-data: folder containing historical stock price data.
  - ./dividend-history: folder containing historical dividends.
  - ./financial-ratio: folder containing financial health of companies.
  - ./industry-analysis: folder containing analysis of companies in the same industries.
  - companies.csv: list of companies.
  - ticker-overview.csv: overview of companies.

Technology stack

A priori, the project is not limited to any tools, libraries and programming languages. Here follows some examples:

- Main requirements:
  - Python (for programming language).
  - Tensorflow (for deep learning and model serving).
  - Scikit-learn (for machine learning and data analysis).
- Extra credit:
  - SQL/MongoDB (for database).
  - Airflow (for task orchestration).
  - Airbyte (as DB connector).
  - dbt (for data transformation).
  - Superset (for dashboard).

Submission

The structure of submission folder should be organized as follows:

- ./<StudentID>-project-notebook.ipynb: Jupyter notebook containing source code.
- ./<StudentID>-project-report.pdf: project report.

The submission folder is named DL4AI-<StudentID>-project (e.g., DL4AI-2012345-project) and then compressed with the same name.

Evaluation

Project evaluation will be based on successful task completion. Here's a breakdown of the grading criteria:

Task	Subtask	Description	Grade	
Task 1 - Nasdaq stock price prediction (Nasdaq dataset)	Task 1.1	Nasdaq multi-feature extension	5%	15%
	Task 1.2	Nasdaq k^{th} day forecast	5%	
	Task 1.3	Nasdaq k days forecast	5%	
Task 2 - Vietnam stock price prediction (Vietnam dataset)	Task 2.1	Vietnam multi-feature extension	5%	15%
	Task 2.2	Vietnam k^{th} day forecast	5%	
	Task 2.3	Vietnam k days forecast	5%	
Task 3 - Trading signal identification for Vietnam market	Task 3.1	Buying signal identification	10%	20%
	Task 3.2	Selling signal identification	10%	
Task 4 - Portfolio composition, risk management and portfolio optimization for Vietnam market	Task 4.1	Portfolio composition	10%	30%
	Task 4.2	Risk management	10%	
	Task 4.3	Portfolio optimization	10%	
Task 5 (Extra credit) - Industry standard for deployment and ease of use	Task 5.1	Model deployment	10%	30%
	Task 5.2	Model as SaaS	10%	
	Task 5.3	AI engineering workflow	10%	
Task 6 - Report			15%	20%
Total			**130%**	

Deadline

Please visit Canvas for details.