Final project
Time-series data and application to stock market

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 so that it is up to you to decide how to programmatically (and reasonably) formulate problems and come up with solutions. You will have to decide on your own (with justification or experimental result) what problems are feasible, i.e., acceptable model accuracy. The evaluation is based on how accurate and reasonable your solution and justification are.

Project requirements

The project has five main tasks and one extra credit section. Here follows the table summarizing the project requirements:

Task	Task title	Description
1 (main)	Nasdaq stock price prediction (Nasdaq dataset)	Nasdaq stock price prediction. Basic questions to answer: • Nasdaq stock price prediction, i.e., how much the price increases/decreases in a certain time window. • How does the price change in different time windows, e.g., one day, one week/seven days in a row, one month/thirty days in a row, etc.? You have to figure out the following things: • 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. • ...
2 (main)	Vietnam stock price prediction (Vietnam dataset)	Vietnam stock price prediction. Basic questions to answer: • Vietnam stock price prediction, i.e., how much the price increases/decreases in a certain time window. • How does the price change in different time windows, e.g., one day, one week/seven days in a row, one month/thirty days in a row, etc.? You have to figure out the following things: • 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.
3 (main)	Vietnam/Nasdaq trading point identification (Vietnam or Nasdaq dataset)	Vietnam/Nasdaq trading point prediction. Basic questions to answer: • What is a good signal for buying stock of certain company? • What is a good signal for selling stock of certain company? You have to figure out the following things: • 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? Note: Building model for Vietnam market is more challenging, therefore more preferable.
4 (main)	Vietnam/Nasdaq portfolio/risk management (Vietnam or Nasdaq dataset)	Vietnam/Nasdaq portfolio/risk management. Basic questions to answer: • What is the list of companies to hold? What is the profit within a certain period? • What is the list of companies to get rid of? Why? • How to combine potential scores and risk scores into a portfolio to optimize investment strategy? You have to figure out the following things: • 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.
5 (main)	Report	Report:

		• Describing the journey about your experiments, observations, findings and conclusions. • Minimum six-page report. • Using AI-powered report writing tools is strictly prohibited. Note: • Experiment failures are as valuable as experiment successes. • The report must include instruction, if necessary, of how to run your code such as external libraires to be installed, etc.
6 (extra)	Industry standard for deployment and ease of use	Industry standard for deployment and ease of use: • Deploy the prediction models as API services. • Deploy the prediction models as a web-based Software-as-a-Service (SaaS). • Design an engineering flow to automate the tasks. Some keywords to research: • API services: Tensorflow Serving (TFServing), REST APIs, gRPC. • Web-based SaaS: TensorflowJS, Superset / Tableau / PowerBI. • Engineering and automation flow: SQL, MongoDB, Airflow, Airbyte, dbt.

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.

    - crawl-vn-data.py: script for crawling Vietnam stock data.

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

The project evaluation is based on the completion of these tasks:

- (20%) Nasdaq stock price prediction.
- (20%) Vietnam stock price prediction.
- (20%) Vietnam/Nasdaq trading point identification.
- (20%) Vietnam/Nasdaq portfolio/risk management.
- (20%) Report.

Extra credit (25%): industry standard for deployment and ease of use.
Total: up to 125% of the ordinary project grading schema.

Disclaimer: this is a free-style project so that it is up to you to decide how to programmatically (and reasonably) formulate problems and come up with solutions. You will have to decide on your own (with justification or experimental result) what problems are feasible, i.e., acceptable model accuracy. The evaluation is based on how accurate and reasonable your solution and justification are.

Deadline

Please visit Canvas for details.