Assignment

Objectives

The objective of this assignment is to simulate a real-life data-science situation that can be approached using the process described in class: i) finding a source of data, ii) acquiring and storing it, iii) cleaning and preprocessing it, iv) extracting meaningful visualisations, v) building a model for inference. You are also free to use any additional methods you find are well suited for the problem.
The outcome of this assignment consists of two deliverables: a) a written report in the form of an academic paper and, b) the code-base to support it. You are allowed to discuss ideas with peers, but your code, experiments and report must be done solely based on your own work.

Overview

Assume you are a junior Data Scientist at Money, a UK investment company and your project manager, Al, provides you with the following list of public companies:

- Apple Inc. (AAPL),
- Microsoft Corp. (MSFT),
- American Airlines Group Inc (AAL),
- Zoom Video Communication Inc (ZM)

Your duty is to select **one** of these companies and study their market trends to ultimately be able to advice on when and whether Money should buy, hold, or sell this stock.

Al asked you to follow the company guidelines, which advise this process:

1. Select a company and acquire stock data in the time window specified in Section 2.
2. Collect any other data of events (e.g., climate changes, a pandemic, season, etc.) that might have an impact on the company's stocks.
3. Choose the storing strategy that most efficiently supports the upcoming data analysis.
4. Check for any missing/noisy/outlier data, and clean it, only if necessary.
5. Process the data, extracting features that you believe are meaningful to forecast the trend of the stock.
6. Provide useful visualisations of the data, exploiting patterns you might find.
7. Train a model to predict the closing stock price.

Details for each task are described in *Section 2*. Details of how each task is marked are in *Section 3*.
The code for each task should allow anyone to reproduce your work. Details on the code setup are in section 3.2.

2. Tasks details

Task 1: Data Acquisition

You will first have to acquire the necessary data for conducting your study. One essential type of data that you will need, are the stock prices for the company you have chosen from **April 2017 to April 2022** as described in Section 1. Since these companies are public, the data is made available online. The first task is for you to search and collect this data, finding the best way to access and download it. A good place to look is on platforms that provide free data relating to the stock market such as *Google Finance* or *Yahoo! Finance*.

There are many valuable sources of information for analysing the stock market. In addition to time series depicting the evolution of stock prices, acquire auxiliary data that is likely to be useful for the forecast, such as:

a) **Social Media, e.g., Twitter:** This can be used to uncover the public's sentimental response to the stock market
b) **Financial reports:** This can help explain what kind of factors are likely to affect the stock market the most
c) **News:** This can be used to draw links between current affairs and the stock market
d) **Climate data**: Sometimes weather data is directly correlated to some companies' stock prices and should therefore be taken into account in financial analysis
e) **Others**: anything that can justifiably support your analysis.

Remember, you are looking for *historical* data, not live data.

Task 2: Data Storage

Once you have found a way to acquire the relevant data, you need to decide on how to store it. You should choose a format that allows an efficient read access enabling the training of a parametric model. Also, the data corpus should be such that it can be easily inspected. Data can be stored locally, on your computer.

Task 3: Data Preprocessing

Now that you have the data stored, you can start preprocessing it. Think about what features to keep, which ones to transform, combine or discard. Make sure your data is clean and consistent (e.g., are there many outliers? Any missing values?). You are expected to:

1. Clean the data from missing values and outliers, if any.
2. Provide useful visualisation of the data. Plots should be saved on disk, and not printed on the juptyer notebook.

3. Transform your data (e.g., using normalization, dimensionality reduction, etc.) to improve the forecasting performance.

Remember that data preparation and pre-processing is a very important task in AI projects. Please motivate carefully in the report your steps.

Task 4: Data Exploration

After ensuring that the data is well preprocessed, it is time to start exploring the data to carry out hypotheses and intuition about possible patterns that might be inferred. Depending on the data, different EDA (exploratory data analysis) techniques can be applied, and a large amount of information can be extracted.
For example, you could do the following analysis:

- Time series data is normally a combination of several components:
  - *Trend* represents the overall tendency of the data to increase or decrease over time.
  - *Seasonality* is related to the presence of recurrent patterns that appear after regular intervals (like seasons).
  - *Random noise* is often hard to explain and represents all those changes in the data that seem unexpected. Sometimes sudden changes are related to fixed or predictable events (i.e., public holidays).
- Features correlation provides additional insight into the data structure. Scatter plots and boxplots are useful tools to spot relevant information.
- Explain unusual behaviour.
- Explore the correlation between stock price data and other external data that you can collect (as listed in Sec 2.1)
- Use hypothesis testing to better understand the composition of your dataset and its representativeness.

一般由滑动平均获得趋势

相变 分形结构

At the end of this step, provide key insights on the data. This data exploration procedure should inform the subsequent data analysis/inference procedure, allowing one to establish a predictive relationship between variables.

Task 5: Inference

Train a model to predict the closing stock price on each day for the data you have already collected, stored, preprocessed and explored from previous steps. The data must be spanning from April 2017 to April 2022.
You should develop two separate models:

1. A model for predicting the closing stock price on each day for a 1-month time window (until end of May 2022), using only time series of stock prices.

2. A model for predicting the closing stock price on each day for a 1-month time window (until end of May 2022), using the time series of stock prices *and* the auxiliary data you collected.

[**NOTE**] During training, make sure that all data is historical (don't use stock data or auxiliary data from the month you are predicting).

Which model is performing better? How do you measure performance and why? How could you further improve the performance? Are the models capable of predicting the closing stock prices far into the future?

[IMPORTANT NOTE] For these tasks, you are not expected to compare model architectures, but examine and analyse the differences when training the same model with multiple data attributes and information from sources. Therefore, you should decide a single model suitable for time series data to solve the tasks described above. ~~Please see the lecture slides for tips on model selection and feel free to experiment before selecting one~~.

The following would help you evaluate your approach and highlight potential weaknesses in your process:

1. Evaluate the performance of your model using different metrics, e.g. mean squared error, mean absolute error or R-squared.
2. Use *ARIMA* and *Facebook Prophet* to explore the uncertainty on your model's predicted values by employing confidence bands.
3. Result visualization: create joint plots showing marginal distributions to understand the correlation between actual and predicted values.
4. Finding the mean, median and skewness of the residual distribution might provide additional insight into the predictive capability of the model.

3. Deliverables

3.1. Report

The report should be written in the form of an academic paper using using the following template (both latex and MS word templates) in ***DAPS_assignment_kit*** (DAPS_assignment_kit.zip). The criteria for each part are detailed in the template. For beginners in latex, we recommend overleaf.com, which is a free online latex editor.

Once you finish your report, please export it into a PDF document and name it with the following format (Using your SN number):

Report_DAPS_22-23 _SN12345678.pdf

The paper should be at most **8 pages** long excluding references, with an additional maximum of

2 pages for references.

The paper must include the following sections:

- **Abstract**. This section should be a short paragraph (4-5 sentences) that provides a brief overview of the methodology and results presented in the report.
- **Introduction**. This section describes the problem with an emphasis on the motivations and the end goal. Please make this introduction meaningful, be short but impactful with clear description of the context but also your project understanding.
- **Data description**. This section details the data that was used for this study. For each data set, should clearly describe the content, size and format of the data. The reason for selecting each data set should also be provided in this section.
- **Data acquisition.** This section presents the data acquisition process, explaining how each data set was acquired, and **why** did you choose the specific data acquisition method.
- **Data storage**. This section explains and **justifies** your data storage strategies.
- **Data preprocessing.** This section should describe in detail all the preprocessing steps that were applied to the data. A **justification** for each step should also be provided. **In case no or very little preprocessing was done, this section should clearly justify why.** It is really important for you to clearly motivate and explain your reasoning.
- **Data Exploration**. This section should summarize any data exploration task you have resorted to in order to find particular patterns within the data. Strong emphasis will be given the to justification and the reasoning that you applied in this phase.
- **Data inference**. This section should first describe the inference problem, then **explain** and **justify** the methodology used to approach the problem and finally present the results.
- **Conclusion**. This last section summarises the findings, highlights any challenges or limitations that were encountered during the study and provides directions for potential improvements. Please do not forget the main goal of this project. What should you learn from the inference? What is the actual conclusion of your study?

Please make sure you complement your discussion in each section with relevant equations, diagrams, or figures as you see fit. Note that your work will be evaluated based on the report. Information not appearing in the report cannot be deduce, so please provide reasoning and motivation behind each step.

Note: figures need to be readable to be considered.

3.2. Code

In addition to the report, you should also provide all the code that was used for your study.
The assignment will be released using GitHub classroom, as we have shown in the formative assessments throughout the course.
In the repository, you will find this ***README.md***, and a jupyter notebook named ***assignment.ipynb***

The code you submit must:

- **Reproducible.** We will run the automated procedure described below. The code will be considered reproducible if a) the automated procedure succeeds and b) the code returns the same results included in the report. Notice that this requires that the code does not request for manual inputs, and does not stop pause its execution for, e.g., plot graphs (do not use plt.imshow(), but rather save the image to disk). Furthermore, an environment.yml file containing a conda recipe is required to ensure a reproducible environment. More details on the procedure will be released in due course.
- **Documented.** We will evaluate the quality of the documentation against the section 3.8 of the Google Python Style Guide.
- **Of good quality.** We will evaluate the quality with respect to the PEP8 style guide. Please, if time complexity is not an issue, aim for readability and clarity when writing your code.
- **Well organised.** Please, organise your code following the guidelines we provided during the course. Aim at not repeating yourself, and at grouping together methods and classes that are shared across tasks. You are free to choose the criteria to organise your code, as long as the choices are justified.
- **Version controlled.** Please, use GitHub to submit your code following the best practices suggested in the course. Prefer modular commits to one big final commit, and aim at pairing each commit with a specific purpose, e.g. “add module to acquire data”. We will provide more details on how each of these assessment criteria will be deployed in due course.

4. Marking Scheme (+ Submission Instructions)

4.1 Marking Scheme

The mark will be decided based on both the **report (70% of final mark)** and **corresponding code (30% of final mark)**. In particular, we will mark based on following scheme:

REPORT		70%
Abstract		**5%**
Introduction		**5%**
Data Description		**15%**
Data Acquisition	Stock prices over time	**5%**
	External data*	**10%**
Data Storage	Data locally stored on your computer in a tidy format of your choice (*csv, pkl, numpy files*)	**6%**
	A cloud-based database	**4%**
Data Preprocessing	Data Cleaning	**5%**
	Data Visualization	**5%**
	Data Transformation	**5%**
Data Exploration	EDA	**10%**
	Hypothesis Testing	**10%**
Data Inference	Development of model using stocks	**3%**
	Development of model using stocks and other data sources	**4%**
	Evaluation metrics implementation	**3%**
Conclusion		**5%**

* You are expected to find at least one extra data source to acquire, store, preprocess, explore and use in the second model of inference step.

CODE		30%
Reproducibility		25%
Documentation		25%
Code Quality		25%
Code Organization (Architecture)		25%
Github “Usage” [extra]		10% [extra]

4.2. Submission

When accepting an assignment, you will not only receive a personal repository, but GitHub Classroom will also open a Pull Request (PR) for you. The PR is a place where we can exchange Questions and Answers about your solution to the exercise. To submit your assignment, please comment on the PR with @epignatelli I submit! This way we will receive the notification you submitted and know that the assignment is completed. **Make sure you push your work to remote.** Commits sent after the deadline will not be considered.

Report Submission

Submit **only** reports in PDF format. Upload the file on Moodle via the “*ELEC0136 Assignment (Report) Submission*” page.