

Financial Data Science (AFIN8015)

Financial Data Analysis 2

Session-1, 2024 Total Marks 100

General Instructions

- This assignment has two parts.
- **Part-I** is on theoretical background and **Part-II** is machine learning, specifically, classification and regression models. Both parts must be documented in one document as Part-I and Part-II.
- There are two data files for this assessment: (1) 'UCI_credit_sample' for the classification task, and (2) 'data_stocks_24' for the regression task.
- For question 4: You have been assigned a company to work with (look at 'Afin8015_final_2024_company_allocation.xlsx' on iLearn) for the regression task, you must work on the company listed against your name as the company in your analysis.

You must first check if the stock prices for your allocated company are in the file 'Data_stocks_24.csv'. Inform the unit convenor in an unlikely case of the prices not available in the data file.

- The assignment requires submission of your working R code or RMarkdown file
 - If not using a RMarkdown format: The code must also be included in the appendix of the document and an R code file must be uploaded.
 - If using a RMarkdown format: The code must be visible in the document along with the output.
The RMarkdown file must be uploaded with the submission.
 - You are **encouraged** to use RMarkdown format for this assignment but it's **not** a requirement. You are free to create a word document from RMarkdown and then update the Word document.
- On your report include your name and Student ID. A cover sheet is not required.
- Your individual paper should not exceed the equivalent of around 15 A4 pages of 11pt font size with 2 spacings. This excludes any appendices, tables, and R output you may elect to incorporate in the report.
- The word count mentioned for a specific question is the maximum word count but remember it excludes any figures and/or tables.
- Marks will be awarded for depth of coverage, quality of insight, succinctness, and accuracy of answers.
- Remember descriptive, predictive, and prescriptive – running the code is part of the task – what procedure you're using and why, assumptions, what were the results and how do you explain.
- Careful of excessive and poor paraphrasing. Make sure your view is expressed.
- Marks will be deducted for poorly informed reports which lack proper formatting, referencing etc. Following minimum deduction will apply:
 - No references when required (in-text and end text) -10
 - Illegible presentation: -10
 - Lack of informed research: -10

¹Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

²Description is sourced from here <https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset>

- R code as images: -10 (you may also be asked to provide an identical machine-readable copy.
- Error in code when trying to replicate -10
- Plagiarism will be dealt with according to the university policy.
- Relevant sources must be cited in the document. Both, in-text and end text citations are required. End text references are excluded from the page limit.
Use one citation style consistently, either APA or Harvard.
- Include comments - at the start of your code include name and across code.
- Error in code when trying to replicate including issues from
 - setwd or include file path and file name as it won't work on replication.
 - make sure any files are submitted if you're referencing your code.
 - single code file either r script or RMD and single report file. No need to include data files but keep the name of the files provided unchanged
 - Don't submit as a compressed or ZIP file.
 - Don't submit as an R project.
- Comment out package install.
- Include any library used.
- Consider code readability and avoid complexity.
- Any tables in your report to be in text not pictures so that it can be read by Turnitin.
- Your code needs to be in your report and in a form that can be read by Turnitin.
- It is ok to refer to the appendix to support but need some comments in the report on what you've asked for and let the reader know if more support is provided in the Appendix.
- Quality of figures - need to be able to read.
- Only make one data load and don't create and load subsequent save files.
- Use the same numbering as the question to make it clear what task you are addressing.
- Reference authoritative sources ??Wikipedia

Before submitting - close R Studio and rerun to double-check what you're submitting runs.

¹Lichman, M. (2013). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.

²Description is sourced from here <https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset>

Assignment Questions

Scenario: You are an intern data scientist at Shootingformars Corp. and your internship period is ending soon. As part of the final deliverables, you are required to submit a report on some of the new methods in Machine Learning and Data Science you have applied on various use cases in the financial service sector during the internship. Ms Fowler, your industry mentor, has provided you with a list of questions and tasks to address in this final report. She will evaluate this report to assess your progress during the internship and your contribution to the company.

Ms Fowler has given you a sample from the default of credit card data set¹ used in [Yeh and Lien \(2009\)](#) and assigned a publicly trading stock listed in 'Afin8015_final_2024_company_allocation.xlsx' and given you a set of questions/tasks as listed in Part-I and Part-II of this document. Part-I is aimed to assess your understanding of Data Science & Machine Learning concepts and methods. Part-II of the task is to test your computing skills in conducting descriptive analysis for classification and regression exercises using some of these methods. The task requires you to create a professional standard document to be presented to your mentor. You have been given a choice of either using a traditional workflow of creating and word document and R for coding the methods separately and then bring them all together in one document or use a reproducible method with an RMarkdown file.

Part I. Data Science Concepts & Methods

- 1 ChatGPT is provided with the following prompt "Critically evaluate these three Data Science/Machine Learning methods their application and any limitations. i. Decision Trees and the main components in their structure, ii. Principal Component Analysis and iii. Text Mining.

ChatGPT Response (in pink)

Decision Trees

- **Definition:** Decision Trees are a supervised learning method used for classification and regression tasks.
- **Main Components:**
 - Root Node: represents the input data
 - Decision Nodes: represent the features or attributes of the data
 - Leaf Nodes: represent the predicted class or value
- **How it works:** Decision Trees recursively partition the data into subsets based on the values of the features, creating a tree-like model of decisions.
- **Advantages:** easy to interpret, handle missing values, and can handle both categorical and numerical data.
- **Limitations:** can be prone to overfitting, and may not perform well with high-dimensional data.

Principal Component Analysis (PCA)

¹Lichman, M. (2013). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.

²Description is sourced from here <https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset>

- **Definition:** PCA is an unsupervised dimension reduction technique used to simplify complex datasets.
- **Main Components:**
 - Eigenvectors: directions of the new coordinate system
 - Eigenvalues: amount of variance explained by each eigenvector
- **How it works:** PCA projects the data onto a new coordinate system, retaining the most important features and reducing the dimensionality.
- **Advantages:** reduces noise, improves visualization, and speeds up computation.
- **Limitations:** sensitive to scale, and may not perform well with non-linear relationships.

Text Mining

- **Definition:** Text Mining is the process of extracting useful patterns, relationships, or insights from text data.
- **Main Components:**
 - Tokenization: breaking text into individual words or phrases
 - Stopword removal: removing common words like "the", "and", etc.
 - Stemming or Lemmatization: reducing words to their base form
- **How it works:** Text Mining uses techniques like sentiment analysis, topic modeling, and named entity recognition to extract insights from text data.
- **Advantages:** extracts valuable information from unstructured data, and can handle large volumes of text.
- **Limitations:** may require domain-specific knowledge and can be sensitive to language and syntax.

1.1 Expand on key aspects of the explanation provided by ChatGPT for each of these three Data Science/Machine Learning methods and relate them to readings from this Unit.

1.2 Critique the explanations – based on your study of this subject what do you think are the strengths and weaknesses of the ChatGPT explanation. Provide support to your argument where appropriate.

1.3 Critically evaluate the application of each of these three Data Science/Machine Learning methods in Finance and their likely popularity. These should ideally be recent examples and be your critique – not mainly a paraphrase of a study.

Word limit: 500 words max each Total mark: 15 (5+5+5)

¹Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

²Description is sourced from here <https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset>

- 2 ChatGPT is prompted to “Explore and analyse the term Big Data? Outline and critically evaluate at least five (5) characteristics of Big Data. Critically analyse at least two examples of Big Data use cases in Finance.”

ChatGPT Response (in pink)

Big Data: Definition and Characteristics

Big Data refers to the large and complex sets of data that traditional data processing tools and techniques cannot manage due to their size, variety, speed, and complexity. The five key characteristics of Big Data are:

1. **Volume:** Big Data refers to enormous amounts of data, often measured in terabytes or petabytes.
2. **Velocity:** Big Data is generated at high speeds, often in real-time, from sources like sensors, social media, and applications.
3. **Variety:** Big Data comes in various formats, including structured, semi-structured, and unstructured data, such as text, images, and videos.
4. **Veracity:** Big Data's accuracy and quality can vary, requiring robust processing and analysis techniques to ensure reliability.
5. **Value:** Big Data's value lies in its potential to reveal insights and drive decision-making, but this requires effective extraction, analysis, and interpretation.

Critical Evaluation:

While Big Data offers immense potential, its characteristics also pose challenges:

- Handling large volumes and velocities requires significant computational resources and advanced analytics tools.
- Managing variety and veracity demands flexible and adaptable processing techniques.
- Extracting value from Big Data requires skilled analysts and domain expertise.

Big Data Use Cases in Finance:

1. **Risk Management:** Big Data analytics can help financial institutions identify and mitigate risks by analyzing large datasets from various sources, such as customer transactions, market trends, and social media.
2. **Fraud Detection:** Big Data analytics can detect fraudulent patterns and anomalies in real-time, enabling financial institutions to respond promptly and prevent financial losses.

Examples:

- **Credit Card Fraud Detection:** A bank uses Big Data analytics to monitor transaction data in real-time, identifying suspicious patterns and flagging potential fraud. This enables the bank to take swift action, reducing financial losses and improving customer trust.

¹Lichman, M. (2013). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.

²Description is sourced from here <https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset>

- **Portfolio Optimization:** A wealth management firm uses Big Data analytics to analyze market trends, economic indicators, and client data, optimizing investment portfolios and improving returns for clients.

Critical Analysis:

While Big Data analytics offers significant benefits in Finance, it's crucial to consider the following:

- Data privacy and security concerns must be addressed when handling sensitive financial data.
- Ethical considerations, such as avoiding biases in algorithms and ensuring transparency, are essential when applying Big Data analytics in Finance.

2.1 Critique the explanations provided by the above ChatGPT response. Based on your study of this subject, what do you think are the strengths and weaknesses of the explanations. Provide support to your argument where appropriate.

2.2 Critically evaluate the issue of data privacy mentioned in the ChatGPT response under critical analysis above. Include in your evaluation how the Australian Government has approached the issue.

2.3 Critically evaluate the ethical considerations mentioned in the ChatGPT response under critical analysis above. Include in your evaluation how the Australian Government has approached the issue.

The critique should go beyond the textbooks and in-class resources to include a synthesis that includes recent developments. All references to be cited.

Word limit: 500 words max Total mark: 15 marks (5+5+5)

¹Lichman, M. (2013). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.

²Description is sourced from here <https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset>

Part II. Classification & Regression Modelling

3 Classifying Credit Risk

You have previously (Financial Data Analysis 1), conducted a price direction prediction exercise using Logistic Regression and K-NN methods. Since that time, you have been introduced to some additional ML models to conduct classification and regression analysis. You have been asked to explore a sample credit risk dataset which can be used for credit risk modelling and forecasting. A subset of the *Default of Credit Card Clients Dataset* has been provided for you to conduct a descriptive analysis of the data.

Data Description:

This is a subset of a dataset containing information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. Content: There are 12 variables² :

- default: Default payment (1=yes, 0=no)
- LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit)
- SEX: Gender (1=male, 2=female)
- EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
- MARRIAGE: Marital status (1=married, 2=single, 3=others)
- AGE: Age in years
- PAY_0: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, ... 8=payment delay for eight months, 9=payment delay for nine months and above)
- PAY_2: Repayment status in August, 2005 (scale same as above)
- BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)
- BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)
- PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)
- PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)

Import the provided *UCI_credit_sample.csv* dataset. All processing is to be undertaken in R (e.g. do not make changes to name or content of the provided CSV file) and:

3.1 Perform an initial analysis of the data and create a box plot(s) of continuous values using ggplot2. The box plot should be separated (filled) according to the default value and labelled. Explain the purpose of any techniques used and any assumptions surrounding their use.

3.2 Create bar chart(s) for the factors with levels. The bar chart should have separate bars according to the default value. Explain the purpose of any techniques used and any assumptions surrounding their use.

¹Lichman, M. (2013). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.

²Description is sourced from here <https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset>

3.3 Present the graphs in your report along with a summary of the data. Label the graphs informatively. Provide a discussion on the variables based on the graph and summary of the data. Are there any variables which are different for the two default levels, comment.

Total marks: 15 marks
(5+5+5)

¹Lichman, M. (2013). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.

²Description is sourced from here <https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset>

4 Forecasting Stock Returns using Technical Indicators

You have been assigned a company to work with (see *Afin8015_final_2024_company_allocation.xlsx* on iLearn) for this regression task, you must work on the company listed against your name as the company in your analysis.

Conduct a comparative regression forecasting analysis to forecast stock returns for the stock you have been assigned from the NASDAQ stock market. The comparative analysis is to demonstrate regression and predictive ability of some of the ML methods in modelling and predicting the stock returns based on various technical indicators. A data file *data_stocks_24.csv* containing the stock prices for the assigned stock is provided for this task.

Data Description:

This is a dataset containing stock prices for selected NASDAQ stocks for the two years' Jan-2021 to Dec 2022. Content: There are 80 variables:

- Date: Daily dates
- Column 2-79: Stock Prices for selected stocks from NASDAQ. Columns are named according to the ticker symbols.
- Column 80 (sentiment): Daily aggregated sentiment scores generated using Alexandria (ACTA) data.

Import the provided *data_stocks_24.csv* dataset. All processing is to be undertaken in R (e.g. do not make changes to name or content of the provided CSV file) and:

4.1 Use the provided data select the closing prices for the assigned stock and create the following Technical Indicators. Explain the possible reason for using these indicators and any assumptions surrounding their use. Hint: First convert the stock prices to an XTS object using the xts package. Use the TTR and quantmod package.

Outcome Variable:

4.1.1 Daily logarithmic returns

Predictor variables:

4.1.2 One period lag of the log returns

4.1.3 One period lag of Moving Average: 5 day

4.1.4 One period lag of Exponential Moving Average: 5 day

4.1.5 One period lag of RSI with 5 day period

Total marks: 12 marks (1.5+1.5+3+3+3)

4.2 Select the sentiment indicator from the data and convert it to an xts time series object. Merge the lag of the sentiment indicator with other technical indicators and the logarithmic returns (outcome variable).

Total marks: 2 marks

¹Lichman, M. (2013). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.

²Description is sourced from here <https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset>

4.3 Create a training and testing sample from the dataset with the last 100 days of the data for the testing set and remaining as training set.

Total marks: 4 marks

4.4 Conduct a regression (forecasting) exercise using Neural Networks. The analysis should include the following:

- 4.4.1 Training on the training sample using a 'timeslice' sampling. Use 100 days for horizon (prediction window) and remaining training days in training data as window size (initialWindow) for prediction in a fixed window.**
- 4.4.2 Data pre-processing to standardise the data.**
- 4.4.3 Fit Neural networks for regression modelling using the 'nnet' method with RMSE metric and tuneLength of 5.**
- 4.4.4 Prediction on the test set and performance measures. Include the measures in the report.**
- 4.4.5 Plot of predicted and actual values. Present the plot in the report.**
- 4.4.6 Brief discussion on the accuracy of the prediction based on the performance measures.**

Total marks: 22 marks (4+3+4+4+4+3)

4.5 Conduct the forecasting exercise in item (4.4) above using Support Vector Machine (SVM) Regression with default tuning length. The analysis should include the following:

- 4.5.1 Radial kernel for the SVM Regression.**
- 4.5.2 Prediction on the test set and corresponding performance measure. Include the measures in the report.**
- 4.5.3 Plot of predicted and actual values. Present the plot in your report.**
- 4.5.4 Brief discussion on the accuracy of the prediction based on the performance measures and comparison with NNet. Which of the two models would you recommend based on the predictive performance, explain?**

Total marks: 15 marks
(4+4+2+5)

Your final report must include both Part-I and Part-II and must contain the output from the analysis conducted in R. Final code/RMarkdown file must be submitted on the relevant links on iLearn.

****End of Assignment Questions****

¹Lichman, M. (2013). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.

²Description is sourced from here <https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset>

Marking

Part 1	Data Science Concepts & Methods	
1	(5+5+5)	15
2	(5+5+5)	15
Part II	Classification & Regression Modelling	
3	(5+5+5)	15
4.1	(1.5+1.5+3+3+3)	12
4.2		2
4.3		4
4.4	(4+3+4+4+4+3)	22
4.5	(4+4+2+5)	15

References

Yeh, I.-C. & Lien, C.-h. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2), 2473–2480.

¹Lichman, M. (2013). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.

²Description is sourced from here <https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset>