STAT5003
Group Project

Semester 2, 2024

THE UNIVERSITY OF
SYDNEY

Project overview

- **Goal of the project**
  - To solve a classification problem
  - Apply course techniques to solve the problem

- **Split into two stages**
  - Planning and Exploratory Data Analysis
    - Take your dataset and perform an exploratory data analysis or initial data analysis to describe the dataset.
  - Project deadline and presentation submission
    - Submit a detailed report that comprehensively introduces the dataset, describes the problem being addressed, the statistical methods being applied, and the conclusions of the analysis for the dataset.
    - Submit a live/video presentation that summarises key findings of the project in a clear and succinct manner.

Deliverables

- **Project is worth 35% of your total mark**

- **First deliverable (10%) - Week 7**
  - Plan and IDA/EDA: Sunday evening

- **Deliverable Two (25%) - Week 12**
  - Final presentation: Lecture and Tutorial/Lab session (10%)
  - Final report: Sunday evening (15%)

Group sizes

- Group sizes are to be 5 people per group
- Groups to be formed by the end of Week 5
  - If your group size is less than 5 people by the end of Week 5, I may randomly assign some students to your group

- Groups to be formed by students in different tutorial rooms
- Data can be chosen by students/group
  - Recommend listing a short label of dataset next to group
  - Encourages like-minded students to join your group

- Each member of the group will be required to assess the group via
  - assessing the peer contribution (which affects the mark)
  - articulating their own contribution

Types of learning problems

- **Unsupervised**: classes **unknown**, want to discover them from the data. Cluster analysis is a type of unsupervised learning problem.

- **Supervised**: classes are **predefined**, want to use a set of labelled objects to form a classifier for classification of future observations.
  - If labels are discrete, it is a classification problem
  - If labels are continuous, it is a regression problem

- Your project must solve a classification problem
  - Since we generally have 5 people in a group, each group needs to try **at least 4 classification methods**

Example classification problems

THE UNIVERSITY OF SYDNEY

- **Two class classification**
  - Tumour/No tumour classification
  - Predict winning party in two party in an election
  - Identify credit card fraud

- **Multi-class prediction (more difficult)**
  - Identify species based on images
  - Diagnosis of a patient based on symptoms

Selecting a dataset

THE UNIVERSITY OF SYDNEY

- Can select your own dataset
- Can be found in a repository
  - UCI Machine learning repository
  - Kaggle datasets
  - Your own workplace
  - Others, can be creative.

- Datasets you cannot use
  - Anything that is already in R, or in the R packages covered throughout the lectures.
  - For example you cannot use
    - `iris` data
    - `Ionosphere` & `Sonar` (`mlbench`)
    - `Boston Housing`
    - `Breast Cancer`
    - `Titanic` dataset
    - `Pima Indian Diabetes`
    - `Cleveland heart data`
    - Wine Reviews data, `MNIST` etc ...

Something to note about datasets

- Not recommended to use data that requires lots of feature engineering and pre-processing
  - Image data
  - Time series data
- The dataset should be challenging in some way.
- Some examples of challenges to solve:
  - Big, e.g., more than 10,000 samples
  - Messy, e.g., lots of missingness
  - Complex, e.g., more than 100 features, mix of numeric and categorical features
  - Involve combining different data sources
  - Multi-class classification
  - Your dataset should satisfy >= 3 criteria, but we'll NOT mark according to how difficult/challenging the data analysis is
- Please indicate why your dataset is challenging at the end of IDA
- Please ensure your features/predictors are **interpretable** (penalty given if not)
- If you are worried about your choice of dataset and problem, please contact a staff member. See more information after marking rubric

Deliverable One: Project plan & Initial Data Analysis (IDA)

THE UNIVERSITY OF SYDNEY

- **Overview of the problem**
  - Describe how this is a classification problem
  - Provide context about why this problem is interesting
- **Dataset description**
  - Describe the data (how many samples are there, are the columns of the data numeric, categorical)
  - What challenges do you envision for your dataset (lots of missing values, high-dimensional data, etc.)
- **Are there features that are highly correlated**
- **Evaluation metrics that are planned to be used**
  - Describe how you plan to evaluate your classification model

Visualization & IDA requirements

- Use visualization to show some properties of the data. Things you may want to explore:
  - Are there outliers
  - What percentage of values are missing
  - Perform and describe any appropriate data cleaning
  - For high dimensional data, can you plot the data in a lower dimension for visualization purpose (e.g., PCA plots)
  - Do histograms/density estimations show anything interesting

- Overall, for the Project plan & IDA:
  - Describe the techniques you plan to implement and when you plan to create the report and presentation.
  - Report/Plan length: less than 6 pages in pdf format; **marks deducted if significantly longer than this limit**
    - Please don't make the report **unnecessarily** long
    - Make the EDA succinct: listing uninformative figures/tables are very bad
  - Compile to **PDF** to check length, and then submit in **HTML**

Example: Good and Bad EDAs

THE UNIVERSITY OF SYDNEY

- A 10/10 report will have
  - An interesting, well defined problem
  - Uses a complex dataset
  - Well written report
  - Well thought out, achievable project plan
  - Appropriate choice of evaluation metric(s)
  - Description of any data cleaning and data wrangling
  - Visualisation of all the relevant features in the dataset
  - Used the appropriate type of plot for the data that is being explored.
  - Report includes excellent explanations of what the plots mean

- A 2/10 report will have
  - An ill defined problem
  - Uses a very simple dataset (e.g. <200 samples, <5 features)
  - Report full of mistakes, incomplete
  - Project plan lacks detail, does not seem achievable
  - No evaluation metric stated
  - Minimal effort in visualization
  - No explanation of the plots
  - Generated plots show little insight and do not help with data exploration

Project dataset guidance

- To gain a good mark, the dataset needs to have some challenges/complexity
- You cannot pick a dataset that is too small or too simple
- If in doubt:
  - Describe your chosen data and goals **qualitatively** to a member of the teaching team between weeks 5 to 7.
  - The teaching team can give feedback about its appropriateness and feasibility.
    - We will not download and inspect the data for you.
- Remember the plan and IDA are the task of the **student group**

Deliverable Two: Final presentation and report

- Final presentation and report will be due at the end of week 12
- Each team will have around 10 mins. This includes:
  - The goal of this is to assess your ability to summaries your key finds succinctly and your ability to communicate your ideas effectively.
    - Akin to a companywide proposal if you were to communicate to a broad audience that doesn't specialise in machine learning/statistics.
  - This presentation should be close to an elevator pitch describing the dataset and your KEY findings.
  - Required to use slides
    - can either be live or pre-recorded.
    - encouraged to give a live talk and upload a pre-recording to complement.
  - For example, you can record via zoom, phones, or any other device. The pre-recording is a single file.
    - The quality of the video recording won't be assessed.

Pecha Kucha slides: Duration and Q&A

THE UNIVERSITY OF SYDNEY

- Presentation should have
  - 20 slides, each slide is shown for 20 seconds
  - Total presentation time = 20 x 20 = 400 seconds = 6 mins 40 seconds.
- Required to create slides that transition every 20 seconds (autoplay)
  - Can be done in Microsoft PowerPoint that supports autoplay
  - Strongly recommend using R markdown instead (e.g., use an output format that supports “autoplay” option). Units in “autoplay” measured in milliseconds. See the example YAML below

```
---
output:
  xaringan::moon_reader:
    nature:
      autoplay: 20000
---
```

- Following the presentation, time for Q&A, there will be around 10 mins total time per group to include this Q&A and transition time between groups.

Final report

- Final report should contain:
  - Overview of the problem
  - Dataset description
  - Initial data analysis/visualization of the data
  - Feature engineering
  - Classification algorithms used
  - Classification performance evaluation
  - Conclusion
- Report needs to be in **R markdown**, capable of producing to pdf or html
- Length: around 8 to 12 pages in pdf format; **marks deducted if significantly longer than this limit**
- Submit in **HTML** (please contact the teaching team if the HTML file is significantly longer than the pdf file)

Example: Good and Bad Final Reports

THE UNIVERSITY OF SYDNEY

- Good final report
  - Used all the data in the prediction or argued a very good case on **why** some data should be removed
  - Performed appropriate feature engineering and/or dimensional reduction
  - Tried **at least 4** classification algorithms
  - For classification algorithms requiring parameters, performed parameter tuning
  - Report clearly described methods used and the results obtained
  - Correctly evaluated different classification models and consider **more than one** performance metrics
  - Report clearly described the model comparison
  - Presentation, and any references to the report, were clear and appropriate

Example: Good and Bad Final Reports

THE UNIVERSITY OF SYDNEY

- Bad final report
  - Only used a small subset of the data for the prediction
  - Dataset not cleaned properly
  - No feature engineering performed when it is necessary
  - Only tried one classification method
  - No parameter tuning
  - Report is full of mistakes, does not contain enough detail, and/or does not describe the methods used appropriately
  - Parts of the report are internally inconsistent
  - Presentation was poorly delivered, and/or project references during the presentation were unclear and/or inappropriate

Notes and suggestions

- This group project guide is intended to give a guide about what the project entails

- It is not intended to be prescriptive about all aspects of the dataset chosen and its complexity

- The goals of the project are
  i. to have an interesting modelling problem that requires a group effort to analyse the data and use the classification techniques in the course to **gain insights into the data**
  ii. to test you know the course concepts, **not just mindlessly plugging stuff in**
  iii. to test you about how to **optimise the hyperparameters** during training. If you use cross-validation, **pay attention to the information leakage**

Notes and suggestions

- There is **no need** to give a detailed breakdown of exactly how each classifier works in either the EDA/Presentation or report (but a *brief* description of the methods you use in the final report is good).

- In the EDA/Plan, all that is needed is a plan of intent on which classifiers intend to be used, perhaps with justification (see below). It is just a plan, and **you can change to other classification methods in your final report**.

  - "Because I have a lot of class labels in this multi-class problem with many observations/features, it is not feasible to fit all pairwise classifiers due to hardware/time constraints."

- *For EDA, please follow the rubric, only show relevant information about why your dataset is interesting and worthy of a group effort to solve. Also, create a rough plan of how you intend to solve it.*

- For all figures in the IDA, presentation slides, and final report, please remember to label the axis in a clear way.

Notes and suggestions

- Teamwork is important!
  - Schedule a regular meeting
    - **Working together** (e.g., in person) before the due dates for the EDA, presentation, and final report is strongly suggested
  - If you cannot agree with your teammate(s)
    - Example: My teammate wants me to do xxx, but I think it is technically incorrect to do such things
    - Please **post on Ed** about your queries about this technical issue
      - *Make your query very clear, otherwise we cannot say anything*
    - The teaching team will try to answer this technical issue for you
  - Get results as early as possible!
    - **Don't wait to Week 12 to have all results ready!**
    - You won't have time to revise anything if you found some fundamental issues in your dataset and data analysis

Frequently Asked Questions

THE UNIVERSITY OF SYDNEY

- *Is my dataset complex enough? It has xxx features and looks hard to me.*

  - The number of features isn't required at a crisp/hard cutoff value. The description in Project guide is only a guide, but **you need to find a dataset satisfying three requirements there as possible as you can**. There is **no hard boundary**. The main takeaway is that the dataset shouldn't be too simple. It should require a collective effort (group effort) over many weeks to solve and present a report/presentation and analysis.

  - In fact, the size of the data isn't necessarily the core issue. There can be smaller datasets that require the correct approach and careful analysis to produce a good outcome.

  - The guide (as the name suggests) merely gives some guidance on what kind of difficulties or challenges need to be dealt with in using a classification modelling technique.

Frequently Asked Questions

THE UNIVERSITY OF SYDNEY

- *Can I change datasets?*
  - No. Historically it hasn't happened.
  - The only common issue is students generally get worried when they realise it is hard to get good classification rates (performance is low). However, **the performance of the models isn't being assessed**.
  - The marking is more on the methodology and reasoning, e.g., a group picking a difficult dataset but analysing it correctly with lower performance should get higher marks than a group picking a dataset that made poor modelling decisions but had better raw performance metric values (e.g., 90% accuracy on a bad analysis isn't better than 60% accuracy on a difficult dataset.

Frequently Asked Questions

THE UNIVERSITY OF SYDNEY

- *What should I do in the EDA?*

  - We have given some guidance in the project guide and the rubric. Please read them if you haven't already.

  - There is no formulaic step-by-step way of doing an EDA. The structure of the data usually informs further investigation.

  - We recommend using your own judgement to describe your dataset and highlight interesting aspects from both an intuitive and modelling aspect. Be sure to justify any positions or statements made.

Frequently Asked Questions

- *Can I use programming language x instead of R to do something.*
  - Short answer, no.
  - Longer answer, this course is intended to be entirely in R for computational statistical methods. Using Rmarkdown (or quarto if you want to use the latest implementation) of reproducible document creation.

Frequently Asked Questions

- *I can't choose between dataset X and dataset Y. What should I do?*

  - This depends on details of the complexity of each dataset and a blanket statement can't be made without more information.

  - If the two datasets are deemed equally difficult or challenging, then students should pick the one that looks the most interesting to the group, i.e., the one that will keep their interest over the course of the project.

  - The group can look at the different challenges between datasets and if they identify some challenges as more interesting, which can help decide on which dataset to use.

Frequently Asked Questions

THE UNIVERSITY OF SYDNEY

- *The rubric/guide says the EDA/report needs to be x pages long. What do I do about code? Do we make an appendix and is it included in the page count.*

  - **It is strongly recommended to use "code_folding" and set it to default to hide.** This means that the code is hidden by default but can be unfolded and viewed at the choice of the reader. More details at the following website: https://bookdown.org/yihui/rmarkdown-cookbook/fold-show.html.

  - **You need to include all source codes to ensure a reproducible report.**

  - The precise length is not intended to be checked. As the overall goal of the project guide is a guideline and not a requirement, the length isn't enforced. Hence, the EDA/report won't be graded against the length directly. However, **if it looks too short or too long, then that will impact on the grade**.

Frequently Asked Questions

THE UNIVERSITY OF SYDNEY

- *I have 200 features. I think I should produce at least 100 visualizations to show how complex my dataset is.*
  - No, please don't do this. The goal is to highlight key features and interesting challenges in the data. This could possibly be done with only a handful of features.

- *How much do I need to do for the EDA/plan?*
  - The goal is to demonstrate how your dataset is interesting, ideally with a motivating useful goal to solve (will doing it have an impact on you or some business/life problem).
  - Then show a plan of how you intend to solve it as a group. **Nothing more**. Please see the rubric/guide and responses above if unclear.

Rubrics

- **Marking rubrics will be available on Canvas**
  - Read carefully to check if you follow the rubrics
- **If some teammates don't make contributions**
  - Approach to the teaching team
  - Reflect in the peer review
- **Good luck with your group project!**