Final Project

Due on December 12, 2025 at 11:59PM

Groups

If you decide to work on this project in a group you must create your group through Canvas. To do so, go on the left menu of the Canvas webpage of the course, select People and then Groups. From the upper right menu select "Final_Project_Group". At the moment, each group is formed by just one student. Therefore, to create a group with more than one students you can simply move your name and those of the other members under the same group.

Note: No more than three students are allowed for each group. All the group members must be added in Cc in any correspondence with the instructor and the TA.

Assistance provided by the instructor and the TA

The purpose of this assignment is to understand what **YOU** would do to address the goals of this project, how **YOU** would perform the analysis of the data provided and how **YOU** would interpret its results. Hence, there is no point in asking the instructor, the TA or others what to do as none of us will be there when your future boss/advisor/recruiter/collaborator will ask you to conduct a similar (but typically much more complicated) analysis.

In light of the above, the instructor and the TA will only address questions related to issues you may encounter with `R` when using the codes you have learned during the Labs and/or the homework assignments. You may also ask clarifications regarding the guidelines provided below. Questions not belonging to these two categories will not be addressed.

You may ask your `R`-related questions during the designated Q&A sessions (on December 3 and December 5), during office hours or, you can post your questions on the discussion board. However, no `R`-related questions will be addressed if posed after December 9, 2025. You may still inquire about the organization of the report after that date.

Submission policy

- The report of the final project has to be submitted electronically through Canvas before the due date indicated above. Late submissions will not be graded and will automatically receive a grade of zero.
- Only what readable through Canvas will be graded. (See section "Submitting the wrong files" of the syllabus for more details.)

Format

- Collect all your results in a well organized report provided in the form of a `.pdf` file. Note: for `.pdf` it is intended that the text must be searchable; submissions saved as `.pdf` images will receive a deduction of 10%.
- **The main body of the report** (i.e., excluding the references, `R`-code appendix, etc) **CANNOT exceed 10 pages** including all relevant outputs and plots. Use a 11pt font size and 3cm margin on each side of the page.

`R` code

- DO NOT include your code in the main body of your report. Specifically, the main body of the report should include only the relevant outputs and plots. All your codes should be well organized and included in the form of an Appendix at the end of the document you submit.
- A report without code in the Appendix will automatically receive a grade of zero.

1 Data

- For this project use the dataset `credit_defaultXX.txt` assigned to you on Canvas under Assignments/Final Project (`https://canvas.umn.edu/courses/516123/assignments/4707994`).
- The datasets contain information regarding credit card clients and their payment behaviors from a financial institution in Taiwan.
- The variables contained in these datasets are:
  - `LIMIT_BAL`: Amount of given credit in NT dollars (includes individual and family/supplementary credit).
  - `SEX`: Gender (1 = male, 2 = female).
  - `EDUCATION`: Education level (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).
  - `PAY_3`: Repayment status in July, 2005 (scale same as above).
  - `PAY_4`: Repayment status in June, 2005 (scale same as above).
  - `PAY_5`: Repayment status in May, 2005 (scale same as above).
  - `PAY_6`: Repayment status in April, 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).
  - `BILL_AMT3`: Amount of bill statement in July, 2005 (NT dollar).
  - `BILL_AMT4`: Amount of bill statement in June, 2005 (NT dollar).
  - `BILL_AMT5`: Amount of bill statement in May, 2005 (NT dollar).
  - `BILL_AMT6`: Amount of bill statement in April, 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).
  - `PAY_AMT3`: Amount of previous payment in July, 2005 (NT dollar).
  - `PAY_AMT4`: Amount of previous payment in June, 2005 (NT dollar).
  - `PAY_AMT5`: Amount of previous payment in May, 2005 (NT dollar).
  - `PAY_AMT6`: Amount of previous payment in April, 2005 (NT dollar).
  - `Default`: Default payment (1 = yes, 0 = no).

2 Goals

- The main goals of your analysis are (1) to understand which variables contribute the most to predicting credit card default, (2) to construct a classification model which can help predicting possible defaults. If you think it is worth/interesting to investigate further questions feel free to address those as well.

3 Preliminaries

- The response variable you want to predict/explain is provided in the variable `Default`.
- For now, if any of the quantitative variables in the dataset contains missing values impute them with the mean of that variable.
- For now, if any of the categorical variables in the dataset contains missing values exclude the observations with missingness from the analysis.

4 Main analysis to be implemented

- Construct a suitable logistic regression model to predict `Default`.
- Sensibly choose one of the procedures discussed in Handout 8 to predict `Default`.
- Implement at least one between random forest, bagging or boosting to predict `Default`.
- Using the original dataset (that is, the one you downloaded from Canvas with missing values on it) impute all the missings using iterative regression.
- Repeat all the analyses above on the newly imputed dataset and compare the resulting error rates with those obtained in the previous round of analysis (that is, those obtained on the dataset where the missings were excluded or imputed as described in Section 3).

5 Organization of the report

- The main body of the report has to include 4 sections: Introduction, Methods, Results, Discussion and **cannot exceed 10 pages**.
- In the **Introduction** you should briefly outline the goals of the analysis and mention which features were the most relevant in predicting `Default`. (This should not take more than half of a page).
- In the section **Methods** you should briefly describe the statistical methods used in the analyses implemented and why you chose them. (This should not take more than one page).
- In the section **Results** you should report all the relevant outputs and plots of your analyses. Here you should adequately interpret your results, discuss the performance of the methods

implemented, compare them in terms of predictive accuracy and using ROC curves. Justify the differences observed using different classification methods on the basis of what you have learned in class (e.g., flexibility of the model, validity of the assumptions if any, etc.). Finally, discuss the changes observed in terms of error rates when imputing missing values via iterative regression compared to when imputing the missings with the mean (for continuous variables) and/or excluding them (for categorical variables).

- In the **Discussion** you should briefly summarize your results and address clearly the questions which arose when defining the goals of the analysis. If you have any, provide some suggestion for future analyses in order to construct a model with even better predictive performance. (This should not take more than one page).
- Any source of material other than the textbooks of the course, labs and lecture notes must be cited within the text and in a dedicated section named **References**. **This implies any AI tool used must be adequately cited.**
- In addition to the above, at the end of your report, you must add an **Appendix** where you will collect all your `R` codes.
- Clearly, the Appendix and, if provided, the references section, will NOT be counted as part of the 10 pages limit.

6 Grading scheme

6.1 Preliminaries

- Data preparation: 5%.

6.2 Main analysis

- Model Selection + outputs and/or plots of logistic regression model + Error rate with no interpretation: 10%.
- Outputs and/or plots of one the procedures discussed in Handout 8 + Error rate with no interpretation: 10%.
- Outputs and/or plots of one between random forest, bagging or boosting + Error rates with no interpretation: 10%.
- Error Rates comparison before and after missing data imputation via iterative regression with no interpretation: 15%.

6.3 Organization of the report

- Introduction, Methods and Discussion sections: 10%.
- Results section:
  - Interpretation of the results, discussion of the performance of the methods implemented, comparison using ER and ROC curves: 20%.
  - Justification of the differences observed using different classification methods on the basis of what you have learned in class (e.g., flexibility of the model, validity of the assumptions if any, etc.): 5%.
  - Discussion of the changes observed on the error rates when imputing missing values with iterative regression: 10%.
- Overall form of the report: 5%.