# **INFS 5096 Customer Analytics in Large Organisations**

# **Assignment 2 part 1 – Credit Scoring**

This is the first part of the Assignment 2. It is worth 20 points of your final grade. Your task is to prepare the data, run an analysis, answer research questions below and write a brief report with your findings.

# Dataset

You are provided with the data set having information on credit card customers. Please check data description at the end of this document for variables description.

# Research questions

1. Load and prepare the data, make decisions what variables to include or exclude from the model. Most probably you will need some data transformation/feature engineering before you start building your model. Report and explain your decisions.
2. Prepare Credit Scoring model and Credit Scorecard. Report decisions made during model development stage, final scorecard, statistical summary of results, discuss quality of the model.
3. Use any other tool / predictive model to predict “bad” customers, discuss quality of the model.
4. Provide a discussion comparing credit score and predictive models. Compare quality of models. Compare potential application of these models – how good or bad each of them would serve the business. Tell something about the population.

# Submission

You must submit a formal report in MS Word or PDF format. Your report will include:

1. Introduction (7 marks).
2. Dataset description and data preparation discussion (10 marks).
3. Credit Scoring model with related discussion (40 marks).
4. Predictive model with related discussion (30 marks).
5. Discussion comparing two models (8 marks).
6. Conclusion (5 marks).
7. Appendix with some extra information, if required.

Your report should demonstrate completeness in covering all research questions and **brevity as no one loves reading long reports**. “A picture is worth a thousand words” – use data visualisations to illustrate and support your research findings.

**There is a page limit: no more than 8 pages from the beginning of Introduction to the end of Conclusion. Cover page, content list, appendix, references are not included.**

Students should not try to cheat. Too small fonts, too small pictures will be ignored, so that would be missing information or mistake. It should be very easy for the reader to read and understand what you do and why you do that.

Appendix will not be considered for marking – all important stuff should be in the main report.

Don’t use generative AIs (ChatGPT, etc.) – they waste too much space and don’t deliver any information value to your reports.

# Software

It is expected that students will use SAS Enterprise Miner for this assessment and follow in-class presentation as a guideline. If you don’t have access to SAS Enterprise Miner, then you can use other tools, for example “scorecard” package in R or “scorecardpy” package in Python. However, support will be limited for alternative tools, and SAS-based presentation remains the guideline for important steps, decisions and results.

If you have any questions – feel free to ask on the forum or by email. You can discuss this exercise with me and other students. You are encouraged to share ideas but not solutions. Remember about academic integrity.

# Data description

The data set represents the information about credit cards from the Republic of China (Taiwan) and uses New Taiwan (NT) dollar as a currency. Current change rate is roughly 20 NT dollars per 1 AUD, so some accounts in the data are large, however not too large to be unrealistic. This is real data.

There are 25 variables. First 6 variables are IDs and demographic information. Remaining variables are history of repayments, and it is a bit confusing – see extra comments after the list of variables.

**ID**: ID of each client

**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 (-2=there is nothing to pay, -1=pay duly, 0=partial repayment, 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.payment**: Customer defaulted payment on the next month (1=yes, 0=no)

Comments for the history:

1. History in the data goes from right to left.
2. Payments should be done on the next month, e.g. April bill should be paid in May (BILL\_AMT6 should be paid in PAY\_AMT5).
3. Bill Amount includes past debt plus new expenses. So, if there were no payments in April, bill amount for May goes up.
4. Negative bill amount (BILL\_AMT) is not a mistake, it means that the previous payment was larger that the actual bill and the account has extra cash in it (no debt at all). Negative payment (PAY\_AMT) would be a mistake in the data but I have not seen any.
5. Variable payment status (PAY\_) shows the customer’s status with respect to the same month bill. For example, PAY\_6 shows of BILL\_AMT6 was paid, however respected payment would be recorded in PAY\_AMT5. In general, I found payment status (PAY\_) inconsistent and illogical.

|  |  |
| --- | --- |
| Variable | Description |
| ID | ID of each client |
| LIMIT\_BAL | Amount of given credit in NT dollars (includes individual and family/supplementary credit) |
| SEX | Gender (1=male, 2=female) |
| EDUCATION | 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 | PAY\_0: Repayment status in September 2005 (-1=paid in full, 1=1 month delay, 2=2 months delay, ..., 9=9+ months delay) |
| PAY\_2 - PAY\_6 | Repayment status from April to August 2005 (scale same as PAY\_0) |
| BILL\_AMT1 - BILL\_AMT6 | Amount of bill statements from April to September 2005 (NT dollars) |
| PAY\_AMT1 - PAY\_AMT6 | Amount of previous payments from April to September 2005 (NT dollars) |
| default.payment | Customer defaulted payment on the next month (1=yes, 0=no) |