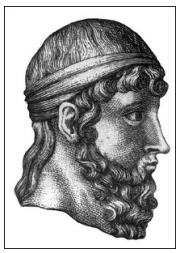
Plato's *Academy* Assignment 1



Credit Risk Modelling with Approved-Loan Data: Statistical Inference, Prediction, and Selection Bias

---

1. Introduction and Context

Credit risk modelling is a core activity in banking, quantitative finance, and financial regulation. Institutions use borrower characteristics and loan attributes to estimate the probability of default and to price risk accordingly.

The LendingClub dataset provides detailed historical information on consumer loans that were **approved and issued** by the platform. Importantly, the dataset **does not include rejected loan applications**. As a result, all observed default outcomes are conditional on LendingClub's internal approval process.

This assignment requires students to build predictive models of default risk **while explicitly recognising and addressing the inferential limitations** arising from selection on approval. The focus is not only predictive performance, but also **statistical reasoning, uncertainty quantification, and methodological awareness**.

---

2. Learning Objectives

Upon successful completion of this assignment, students should be able to:

1. Distinguish between prediction and inference in applied modelling contexts
2. Apply classical statistical inference techniques to real financial data

3. Build and evaluate machine learning models for binary outcomes
4. Identify and discuss selection bias in observational datasets
5. Communicate results with appropriate caveats, assumptions, and limitations

---

3. Data Description and Scope

3.1 Primary Dataset

You will use the publicly available LendingClub historical loan dataset.

- Each observation represents **one approved and issued loan**
- Variables include:
    - Borrower characteristics (e.g. income, employment length, credit score)
    - Loan terms (e.g. interest rate, loan amount, grade, term)
    - Loan outcomes (e.g. fully paid, charged off, default indicators)

3.2 Scope Limitation

You must explicitly recognise that:

- The dataset excludes rejected loan applications
- All analyses are conditional on approval
- Estimated relationships may not generalise to the full applicant population

Failure to acknowledge this limitation will be penalised.

---

4. Research Questions

Your analysis must be structured around the following questions:

1. Which borrower and loan characteristics are statistically associated with default among approved loans?
2. To what extent can default be predicted within the approved-loan population?
3. How stable are these relationships across time and borrower subgroups?
4. What are the inferential consequences of modelling only approved loans?

---

5. Data Preparation and Cleaning (Required)

This section must be fully documented in your notebook.

5.1 Initial Inspection

You must:

- Load the dataset and report:
    - Number of observations
    - Number of variables
- Identify variable types (continuous, categorical, binary, date, identifier)

5.2 Handling Missing Values

You must:

- Quantify the proportion of missing values for each variable
- Classify missingness as:
    - Structurally missing
    - Potentially informative

    - Likely noise

For each variable with missing values, you must **explicitly justify** one of the following actions:

- Drop the variable
- Drop affected observations
- Impute (and explain the method)
- Retain missingness as a separate category

5.3 Leakage and Availability Filtering

You must:

- Identify variables that would **not be available at loan origination**
- Remove all such variables from predictive modelling
- Provide a short written justification for why these variables constitute leakage

---

6. Exploratory Data Analysis (EDA)

This section must precede all modelling.

6.1 Descriptive Statistics

You must compute and report:

- Means, medians, and standard deviations for key continuous variables
- Frequency tables for categorical variables
- Overall default rate in the dataset

6.2 Group Comparisons

You must compare distributions between:

- Defaulted vs non-defaulted loans
- At least two borrower risk strata (e.g. credit score bands or loan grades)

Use plots and tables to support your discussion.

---

7. Statistical Inference (Core Requirement)

This section is **compulsory** and carries substantial weight.

7.1 Formal Hypothesis Testing

You must conduct **at least three distinct hypothesis tests**, including:

Test 1: Mean Comparison

- Use a two-sample t-test to compare a continuous variable (e.g. income) between defaulted and non-defaulted loans
- State:
    - Null hypothesis
    - Alternative hypothesis
    - Test statistic
    - Degrees of freedom
    - p-value
- Interpret the result in substantive (non-technical) terms

Test 2: Proportion or Association Test

- Use a chi-square test or proportion test to examine default rates across categories (e.g. loan grades)
- Clearly explain what dependence or independence means in this context

Test 3: Robustness Test

- Use either:
    - A non-parametric alternative, or
    - A test applied to a different variable

Explain why this test is informative.

7.2 Confidence Intervals

You must construct **confidence intervals** for at least three quantities, such as:

- Default rate within a specified borrower group
- Difference in mean interest rates between defaulted and non-defaulted loans
- Difference in default probabilities between two categories

Each interval must be:

- Numerically computed
- Interpreted correctly (frequentist interpretation)

7.3 Likelihood and Estimation

Using logistic regression:

- Explain why coefficients are estimated via **maximum likelihood**
- Interpret at least two coefficients in terms of:
    - Log-odds
    - Odds ratios
- Discuss assumptions underlying MLE and their plausibility in this dataset

---

7.4 Selection Bias Discussion (Written)

You must include a dedicated subsection explaining:

- Why estimated default probabilities are conditional on approval
- How LendingClub's approval filter may distort observed relationships
- Why coefficients cannot be interpreted as population-level causal effects

This section is conceptual and does not require code.

---

8. Predictive Modelling

8.1 Model Specification

You must build **at least two models**, including:

- Logistic regression
- One non-linear model (e.g. random forest or gradient boosting)

8.2 Training and Validation

You must:

- Split data into training and test sets using a time-aware or stratified approach
- Justify your split strategy

8.3 Model Evaluation

You must evaluate models using:

- ROC-AUC
- Precision, recall, and F1 score
- Calibration analysis

Explain what each metric measures and why it is relevant.

---

9. Generalisation and Stability Analysis

You must perform **at least one** of the following:

- Train on early-period loans and test on later-period loans
- Compare model performance across borrower subgroups
- Validate against a second consumer credit dataset (if available)

Discuss whether results appear stable or regime-dependent.

---

10. Interpretation and Economic Reasoning

You must:

- Identify the most important predictors
- Provide economic intuition for their effects
- Discuss how approval-only data likely influences model behaviour

---

11. Deliverables

1. **Jupyter Notebook (`.ipynb`)**
    - Fully executable
    - Clearly structured
    - Includes code, outputs, and explanations
2. **Written Report**
    - Formal academic tone
    - Clear section headings
    - Explicit limitations and assumptions

---

12. Assessment Criteria

- Statistical correctness
- Inferential depth
- Handling of selection bias
- Predictive modelling rigor
- Clarity of exposition