Advanced Statistics Project Report

Daniele Lupico, Sofia Matarante, Simone Menghini, Michele Baldo

1) PROBLEM

Loan issuance is one of the core activities of the banking sector, but it carries significant risks, particularly with clients who have a high probability of default. This project aims to develop predictive models to estimate a client's probability of default based on historical data. The goal is to provide an effective tool to assist banks in the decision-making process and mitigate financial risk.

2) DATA

We started our analysis by examining the dataset, which consists of 45,000 units. We explored the type and distribution of variables using functions like *str()* and *summary().* This initial step revealed three key issues:

1. **Outliers in age:** The variable *person_age* displayed a maximum value of 144, which is clearly unrealistic. To improve the reliability of our analysis, we capped the maximum age at 80.
2. **Categorical variables:** Variables like *person_education* and *loan_intent* were categorical and unsuitable for direct numerical analysis. We transformed these variables into dummy variables and, in some cases, recoded them further for clarity. For instance, *person_education* was recoded into ordinal categories to better reflect its inherent ranking.
3. **Variable scaling:** The *summary()* function revealed significant variations in measures of central tendency (mean and median) across variables. This could bias our results by amplifying the impact of variables with larger coefficients. To address this, we **standardized** all variables, ensuring they were on a comparable scale.

We conducted a correlation analysis (**Graph 1, Line 53**) among the independent variables. This step helped us identify the strength and direction of relationships between variables, as well as potential multicollinearity issues that could undermine our models. Notable correlations included:

- *cb_person_cred_hist_length and person_age* (0.88)
- *person_age and person_emp_exp* (0.95)
- *person_emp_exp and cb_person_cred_hist_length* (0.84)

3) METHOD

The main method we used was logistic regression. We identified outliers and high-leverage points, and through VIF and correlation analysis, we examined multicollinearity. Additionally, we created different models to highlight the presence of confounders. After, we chose the Random Forest model for its accuracy and ability to identify key variables, making it ideal for automating loan approvals and reducing false outcomes. To compare with logistic regression, we tested Random Forest using variables from the full and adjusted regression models.

4) IMPLEMENTATION

Logistic Regression (Method 1, Line 70): After preparing the dataset, we split it into training (70%) and test (30%) sets and developed our initial regression model (*MODEL_FULL*). This model included all variables, including those with high correlations, to assess their individual impact and statistical significance.

Threshold Strategies: We adopted two decision thresholds:

1. A less conservative threshold of 0.5 for loan approval.

2. A more conservative threshold of 0.75.

This dual approach allowed us to explore the implications of different risk tolerances and provided a more adaptable framework for decision-making.

From the *MODEL_FULL* results (**Summary 1, Line 73**), several insights emerged:

- Some variables, such as the entire EDU and HOME categories, *gender_female*, and *previous_loan_defaults*, were statistically insignificant (high p-values and low z-scores).
- Despite its insignificance, *previous_loan_defaults* was retained in the model because of its practical importance in assessing borrower reliability.

To refine the model further, we assessed multicollinearity using the Variance Inflation Factor (VIF). Variables with VIF values exceeding 5 were identified as problematic. The HOME category, being both insignificant and collinear, was excluded.

After experimenting with interactions and polynomial terms to make the *previous_loan_defaults* variable significant (as we consider it important in situations like this), we ultimately settled on a simpler and more interpretable model, **model_adjusted1** (**Summary 2, Line 115**). This model excluded:

- HOME (insignificant and collinear)
- *person_emp_exp (insignificant)*
- *gender_female (insignificant)*

The adjusted model retained mostly significant variables, suggesting it was a better fit for our analysis.

Outliers: To enhance model precision, we identified and removed outliers using thresholds between -2 and 2 (**Graph 2, Line 136**). This process led to the creation of cleaned versions of our models, *model_adjusted1_cleaned* **(Summary 3, Line 145)** *and model_full_cleaned* **(Summary 4, Line 160)**.

Random Forest (Method 2, Line 211): To compare our models with Logistic Regression, we applied Random Forest using the same variables, building a *rf_adjusted_graph* model (**Graph 3, Line 237**) and *rf_full_graph* model (**Graph 4, Line 276**), both tested on 500 and 1,000 trees.

In both cases, the most important variables were *previous_loan_default*, *loan_percent_income,* and *loan_int_rate.* Unlike logistic regression, which considers *previous_loan_default* the least statistically significant despite its high coefficient, Random Forest ranks it as the most important.

5) EVALUATION

Finally, we evaluated our models on the test set using several metrics:

- **ROC curve:** It is used to visualize the classification performance of the models. In our case, the models show significant overlap, supporting our decision to select the *Model_adjusted1_cleaned* as the preferred choice. **(Graph 5, Line 295)**
- **AIC (Akaike Information Criterion):** This metric assessed model quality and facilitated comparisons.
- **Nagelkerke's Pseudo-R^2:** A logistic regression-specific metric based on Log Likelihood, which we preferred for its normalized scale (0 to 1), making it more intuitive.
- **Confusion Matrix** with metrics like: Sensitivity, Specificity, Accuracy, Misclassification Rate. **(Graph 6, Line 429)**
- **The Trivial Classifier:** with a misclassification rate of **0.222**, was used as a benchmark to evaluate the performance of our models. Compared to this benchmark, our Logistic Regression model (adjusted, **Table B**) had a much lower misclassification rate of **0.123**, and the Random

Forest model (Full, **Table D**) performed even better with a rate of **0.075**. These results show that both models are effective at identifying meaningful patterns in the data. (**Summary 5, Line 475**)

From both tables, we can observe that the metrics for the *Model_full_cleaned* are slightly better than those for the *Model_adjusted1_cleaned*. However, the difference is minimal, making it worthwhile to sacrifice a small amount of accuracy in exchange for improved interpretability and simplicity. Comparing the two tables, we consider **Table B** to be more realistic, as research suggests that the threshold used in **Table B** aligns more closely with those commonly applied by banks:

Table A:

THRESHOLD 0.5	Model_full_cleaned	Model_adjusted1_cleaned
AIC	11690	12251
Pseudo-R^2	0.753	0.739
Accuracy	0.897	0.893
Misclassification Rate	0.102	0.106
Sensitivity	0.944	0.941
Specificity	0.736	0.725

Table B:

THRESHOLD 0.75	Model_full_cleaned	Model_adjusted1_cleaned
AIC	11690.27	12250.92
Pseudo-R^2	0.753	0.739
Accuracy	0.883	0.876
Misclassification Rate	0.116	0.123
Sensitivity	0.979	0.976
Specificity	0.546	0.525

For the Random Forest analysis, we determined that using **N.Tree = 1000** and the **Random Forest Full model** is the better approach. The Random Forest algorithm independently captures interactions and non-linear relationships between variables while maintaining readability, even with the inclusion of more variables, as it does not compromise interpretability.

Table C:

N.TREE = 500	Random Forest Full	Random Forest Adjusted
Accuracy	0.924	0.915
Misclassification Rate	0.076	0.085
FPR	0.070	0.070
FNR	0.105	0.149
Sensitivity	0.975	0.963
Specificity	0.747	0.747

Table D:

N.TREE = 1000	Random Forest Full	Random Forest Adjusted
Accuracy	0.925	0.916
Misclassification Rate	0.075	0.084
FPR	0.069	0.069
FNR	0.103	0.149
Sensitivity	0.975	0.963
Specificity	0.748	0.751

6) INTERPRETATION

Thanks to the methodologies implemented, we were able to address our main question from multiple perspectives, specifically identifying the variables most crucial for a bank's decision to approve or reject a loan. The findings depend on the type of model the bank chooses to adopt. In logistic regression, the key factors influencing loan approval are the **purpose of the loan**, its **amount**, and the **credit score**. In contrast, the Random Forest model highlights variables such as **past mortgage behavior**, the **loan-to-income ratio**, and the **loan interest rate** as the most significant. Interestingly, it ranks **previous_loan_default** as the most important variable, despite its low statistical significance in logistic regression, underscoring the methodological differences between the two approaches.While logistic regression offers simplicity and interpretability, Random Forest provides greater accuracy and robustness, making it better suited for analyzing complex datasets. Additionally, the dual-threshold strategy added flexibility, enabling the exploration of different risk tolerance levels and decision-making scenarios.