Credit Risk Analysis Report

Isao Pastelin Kohagura-Eduardo Botello Casey-Emilio Fernández González

November 12, 2025

Business Problem

For this study, the `credit_risk_data.csv` database was analyzed in order to build and evaluate a model capable of predicting the probability that a new applicant will fall into the default category (does not pay their loan) or non-default category (does pay their loan).

The objective of the analysis using Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) is not to directly reduce the percentage of rejected applicants, but to improve the accuracy of the loan approval process, which can indirectly reduce the number of unjustified rejections and maintain a controlled level of risk.

Thanks to this analysis, LendSmart will be able to strengthen its credit approval policy by reducing classification errors:

- Fewer people who would have paid their loan will be rejected (false negatives).
- Fewer people with a high probability of default will be approved (false positives).

Likewise, a better risk estimate will allow for a more informed adjustment of the approval threshold, optimizing the balance between profitability and delinquency.

The following sections present the main findings of the analysis, the comparative performance of both models, and the justification for the final selection of the most appropriate model, along with the recommendations derived from the study.

Key Findings and Insights

Our analysis revealed a clear separation between the variables for the default and non-default groups. This preliminary result suggests that the model will perform well, as there is a significant differentiation between the groups, which is favorable for applying discriminant methods such as LDA and QDA.

The analysis LDA identified three main variables that influence the classification of applicants: `payment_history_score`, `job_stability_score`, and `credit_utilization`.

The first two have negative coefficients—payment_history_score (-15.67) and job_stability_score (-12.80)—indicating that as these scores increase, so does the probability of belonging to the non-default group. In contrast, the variable credit_utilization (11.31) shows a positive coefficient, suggesting that higher values are associated with a greater probability of being in the default group.

In addition to these three key variables, two others were observed to have a smaller but still relevant influence: debt_to_income_ratio (4.55) and credit_score (-4.24).

Taken together, these results allow us to profile a high-risk individual as someone with a poor payment history and job stability, combined with a high debt-to-income ratio and a low credit score.

Below is a table showing the weights of each variable.

ID	Feature	Coefficient
6	payment_history_score	-15.672012
3	job_stability_score	-12.800572
5	credit_utilization	11.317957
8	debt_to_income_ratio	4.554815
4	credit_score	-4.238531
2	employment_years	-2.873156
9	savings_ratio	-2.681301
12	residential_stability	-1.751812
1	annual_income	-1.659910
7	open_credit_lines	-1.462896
18	education_level_High School	1.137332
10	asset_value	-1.070954
0	loan_amount	-0.745283
11	age	-0.495603
14	marital_status_Single	0.480114
15	marital_status_Widowed	0.354983
16	education_level_Bachelors	-0.157395
19	education_level_Masters	-0.071111
13	marital_status_Married	-0.037813
17	education_level_Doctorate	0.017360

Unlike the LDA model, QDA does not provide directly interpretable coefficients, since the separation between classes is based on quadratic and non-linear boundaries. This means that the contribution of each variable to the final result depends on non-linear interactions, which makes the individual interpretation of the coefficients meaningless.

Model Performance and Selection

As discussed in the Key Findings and Insights section, the results obtained were outstanding for both models: LDA and QDA achieved perfect classification, as evidenced by both their confusion matrices and performance metrics, where no false positives or false negatives were observed.

Figure 1: Confusion Matrix - LDA

387	0
0	136

Figure 2: Confusion Matrix - QDA

387	0
0	136

These results confirm that the discriminant methods applied were highly effective in distinguishing between default and non-default individuals within the analyzed dataset.

However, despite identical performance in terms of accuracy, interpretability was a decisive factor in choosing the LDA model. Its linear structure allows for a better understanding of the contribution of each variable and more accurate recommendations for the company, while the QDA model, although equally effective, lacks easily interpretable coefficients due to its quadratic nature.

Final Recommendations

After evaluating both models, we recommend that LendSmart deploy the LDA (Linear Discriminant Analysis) model as the preferred solution since it offers greater interpretability and operational simplicity, allowing the company to clearly understand how each factor contributes to the decision process. However, it is important to note that this model was trained on historical data with clear group separation, meaning that future datasets with more overlap or noise may slightly reduce accuracy. Future work should focus on expanding the test dataset and further validating the model's performance across different samples. Although the current results were exceptional, additional testing will help confirm the model's robustness and ensure consistent accuracy under varying data conditions.