Technical Report: LendSmart Credit Risk Analysis

1. Introduction

This report summarizes the methodology, analysis, statistical validation, and modeling results from the notebook **discriminant_analysis.ipynb**, which evaluates credit risk using Linear and Quadratic Discriminant Analysis (LDA/QDA). The project aims to classify borrowers as *default* or *non-default* based on financial and demographic variables.

2. Data Overview

The dataset contains typical credit-risk variables including: - Income\

- Credit score\
- Debt ratio\
- Loan amount\
- Employment status\
- Education level\
- Default indicator

Initial inspection revealed: - Class imbalance (more non-defaults than defaults).\

- Several continuous predictors suitable for discriminant models.

3. Exploratory Data Analysis (EDA)

Key observations: - Non-default borrowers dominate the dataset, which affects model sensitivity.\

- Loan approval and credit attributes show clear separation between default and non-default groups.\

- Strong correlations exist between income, credit score, and loan approval metrics, which may influence discriminant boundaries.\
- Boxplots revealed that defaulting borrowers tend to have: – Lower income\
- Higher debt\
- Lower credit score

4. Data Preprocessing

Steps performed: – Categorical variables were transformed into dummy variables.\

- Redundant or highly collinear features were reviewed.\
- The dataset was standardized to comply with LDA assumptions.

5. Statistical Assumption Testing

5.1 Multivariate Normality

- Visual inspection (boxplots & distributions) suggests partial deviations from strict normality.
- However, LDA is generally robust to moderate violations.

5.2 Equality of Covariance Matrices

- Variance differences between classes indicate that QDA may outperform LDA, as it does not assume equal covariance.

6. Model Implementation

6.1 Linear Discriminant Analysis (LDA)

- Generated discriminant coefficients indicating the most influential predictors (income, debt ratio, credit score).
- Produced classification boundaries assuming equal covariance matrices.

6.2 Quadratic Discriminant Analysis (QDA)

- Allowed non-linear class boundaries.
- More flexible in modeling heterogeneity between risk groups.

7. Model Evaluation

Both LDA and QDA results included: – Perfect classification accuracy on the test set.\

- AUC = **1.0** for both models.\
- Confusion matrices with **zero misclassifications**.

These results strongly suggest: – The dataset may be too small, too clean, or possibly synthetic.\

- The models may be overfitting due to limited variability.

8. Discussion

Although QDA benefits from flexible covariance modeling, LDA is more stable and easier to interpret. Given similar performance, LDA is preferred for operational credit scoring, while QDA serves as a useful comparative model.

9. Conclusion

- The discriminant analysis workflow successfully demonstrated the application of LDA and QDA for credit-risk classification.
- Due to unusually perfect model performance, additional validation with a larger or real-world dataset is recommended.
- LDA provides strong interpretability and should be chosen as the primary model for deployment, while QDA remains a valuable benchmark.