Executive Summary: Credit Risk Model Development

Luis Alan Morales Castillo A01659147, Paulina Díaz Arroyo A010295932, Rodrigo Jiménez Ortiz A01029623

This report presents the findings from a comprehensive credit risk analysis conducted on 2,500 loan applications to address LendSmart's 26.56% default rate. Through rigorous statistical modeling using Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA), we have identified five critical predictors that accurately distinguish between high-risk and low-risk borrowers: payment history score, job stability score, credit utilization, debt-to-income ratio, and credit score. Both models achieved perfect classification accuracy on the test dataset. We recommend immediate deployment of the LDA model due to its superior interpretability and operational simplicity, while maintaining identical predictive power to QDA.

1 Business Problem

LendSmart Financial Services faces a significant business challenge with a loan default rate of 26.56%, meaning that more than one in four borrowers fail to repay their loans as agreed. This high default rate directly impacts profitability, increases operational costs associated with collections and write-offs, and poses reputational risks in the competitive lending marketplace.

Our team was tasked with building a robust statistical model to identify high-risk applicants during the loan approval process. The primary objective was to develop a predictive tool that would enable LendSmart's underwriting team to make more informed decisions, thereby reducing default rates while ensuring qualified borrowers are not unnecessarily rejected. The analysis utilized a comprehensive dataset of 2,500 loan applications from 2022 to 2024, encompassing 18 variables spanning financial metrics, credit history, employment stability, and demographic factors.

2 Key Findings & Insights

2.1 Critical Risk Factors Identified

Our analysis revealed five dominant predictors that distinguish defaulters from reliable borrowers. These factors, ranked by their statistical importance, provide a clear risk profile framework:

- **Payment History Score** Emerging as the single most powerful predictor of loan default. Borrowers with poor repayment patterns on previous credit obligations are significantly more likely to default on new loans.
- **Job Stability Score** Ranking as the second most important factor. Applicants with unstable employment histories—characterized by frequent job changes or short tenure—demonstrate higher default risk due to income uncertainty and limited ability to weather financial disruptions.
- **Credit Utilization** Representing the third critical indicator. High credit utilization ratios signal that borrowers are already heavily dependent on existing credit lines, leaving minimal financial cushion to absorb additional loan payments.
- **Debt-to-Income Ratio** Which measures the proportion of an applicant's income committed to existing debt obligations. Borrowers with high ratios have limited capacity for new loan payments, making them substantially more vulnerable to default.
- **Credit Score** Remaining as a reliable and well-established predictor. Lower credit scores correlate strongly with increased default risk, reflecting a history of credit mismanagement and late payments.

2.2 Profile of High-Risk Applicants

Based on our comprehensive analysis, the typical high-risk borrower exhibits the following characteristics: inconsistent payment history with missed or late payments on previous obligations; unstable employment with frequent job changes or gaps in employment; credit utilization exceeding 60% of available credit limits;

debt-to-income ratio above 50%, indicating limited disposable income; and credit scores below 650, suggesting past credit difficulties.

Conversely, low-risk applicants demonstrate consistent on-time payment patterns, stable employment tenure exceeding 5 years, conservative credit utilization below 30%, manageable debt-to-income ratios under 35%, and strong credit scores above 700.

2.3 Additional Insights

The exploratory data analysis revealed important patterns in categorical variables. Education level shows a modest correlation with default rates, with doctorate and master's degree holders exhibiting slightly lower default rates than those with high school education only. Marital status demonstrated that married applicants tend to have marginally lower default rates compared to single, divorced, or widowed individuals, potentially reflecting greater financial stability or dual-income households.

3 Model Performance & Selection

3.1 Models Evaluated

We developed and rigorously tested two discriminant analysis models: Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). Both models were trained on 2,000 applications and validated on a held-out test set of 500 applications, maintaining the same 26.56% default rate distribution to ensure unbiased evaluation.

3.2 Performance

Both models achieved perfect classification performance on the test dataset, correctly identifying every single defaulter and non-defaulter with zero errors. The models demonstrated:

- **Accuracy (100%):** All 500 test cases were correctly classified, with 367 non-defaulters and 133 defaulters properly identified.
- **Precision (100%):** Every applicant flagged as high-risk was indeed a defaulter, meaning the model generates no false positives.
- **Recall (100%):** The model successfully identified every actual defaulter in the test set, meaning no false negative predictions were made.
- **AUC Score (1.0):** The ROC curves for both models reached the ideal top-left corner, indicating complete separation between risk classes across all probability thresholds.

3.3 Model Selection: LDA

While both models achieved identical perfect performance, we recommend deploying the **Linear Discriminant Analysis (LDA)** model for the following strategic reasons:

- **Interpretability:** LDA produces straightforward, linear relationships between predictor variables and default risk. The model coefficients can be easily explained to underwriting staff, supporting transparent and defensible lending decisions.
- **Operational Simplicity:** LDA's simpler mathematical structure makes it easier to implement, maintain, and troubleshoot in production systems.
- **Robustness:** Simpler models typically generalize better to new data and are less prone to overfitting, making LDA a safer choice for long-term deployment.

4 Final Recommendation

4.1 Deployment Recommendation

We strongly recommend that LendSmart immediately deploy the LDA credit risk model into production for all new loan applications. This is a must to have recommendation based on the exceptional model performance and substantial business value.

4.2 Expected Business Impact

- **Default Rate Reduction:** Deploying the model will enable LendSmart to reduce the current 26.56% default rate to near-zero levels on properly screened applications, assuming future applicants exhibit similar characteristics to the training data.
- **Financial Savings:** With an average loan amount of approximately $155,000 and assuming a 50% recovery rate on defaulted loans, each prevented default saves roughly $77,500 in losses. If LendSmart processes 10,000 applications annually with a 26.56% baseline default rate, preventing even 80% of these defaults would save approximately $164 million per year.
- **Competitive Advantage:** Lower default rates enable more competitive pricing for qualified borrowers, improved investor confidence, and stronger market positioning.

4.3 Risk Management

- **Real-World Validation:** The perfect 100% accuracy observed in testing is exceptionally rare and may not fully replicate in live production environments. continuous monitoring has to be implemented to validate performance on new applicants.
- **Market Changes:** Economic conditions, regulatory changes, or shifts in applicant demographics may impact model performance over time.
- **False Positive Management:** Although the current model shows zero false positives, any degradation in performance could result in rejecting qualified borrowers, LendSmart should establish a manual review process for borderline cases

4.4 Future Enhancements

- Expand the model to incorporate additional data sources such as bank account transaction patterns, rental payment history, or alternative credit data to further refine risk assessment.
- Investigate machine learning approaches as complementary tools to validate and potentially enhance the discriminant analysis results.
- Build a continuous learning pipeline that automatically retrains the model monthly as new default outcomes are observed, ensuring the model adapts to changing market conditions.

In conclusion, the LDA credit risk model represents a powerful tool that will enable LendSmart to dramatically reduce default losses while maintaining fair and transparent lending practices. With proper monitoring and periodic recalibration, this model will provide sustained business value and competitive advantage in the lending marketplace.