

Data Science for Finance

Group Project #4

Fernando Reis – 20231535

Luis Ribeiro – nº 20231536

Thiago Bellas – nº 20231131

Renato Moraes – nº 20231135

Saad Islam – nº 20230513

Table of Contents

1. Introduction	4
2. Exploratory data analysis	4
3. Feature Engineering & Selection	6
4. Modelling	8
Logistic Regression	8
Machine Learning Model – Random Forest	9
Machine Learning Model – XGBoost	10
Deep Learning Model	11
5. Discussion	12
6. Conclusion	13

Table of Figures

Figure 1 Text to numbers.....	6
Figure 2 Label Encoder	6
Figure 3 One Hot Encoder	6
Figure 4 Date formatting	7
Figure 5 woe and iv for verification_status	7
Figure 6 Outliers in the dataset.....	7
Figure 7 Scaling of the dataset	7
Figure 8 Smote approach	8
Figure 9 Results of the Logistic Regression Model	8
Figure 10 Confusion Matrix of the Logistic Regression Model.....	9
Figure 11 Results of the Random Forest Model.....	9
Figure 12 Confusion Matrix of the Random Forest Model	10
Figure 13 Results of the XGBoost Model	10
Figure 14 Confusion Matrix of the XGBoost Model	11
Figure 15 Training & Validation Loss Figure 16 Training & Validation accuracy	11

1. Introduction

In this group project, our primary goal was to design and implement a predictive model capable of accurately assessing credit risk and forecasting loan defaults. We utilized advanced machine learning techniques to analyze a rich financial dataset that captured detailed borrower information, including loan amounts, loan terms, and employment titles among other things. This dataset required meticulous preprocessing to prepare for model training, involving comprehensive data cleaning, normalization, and exploratory data analysis. Such steps were crucial to ensure the integrity and accuracy of our predictive models. One of the significant challenges in credit risk modeling is addressing the imbalance typically present in datasets where defaults are much less common than non-defaults. To overcome this, we employed the Synthetic Minority Over-sampling Technique (SMOTE), a method proven to improve model outcomes by balancing class distribution, which in turn enhanced the predictive performance of our models. Our methodological framework was systematic and robust, starting with a basic logistic regression model and progressively incorporating more sophisticated algorithms such as Random Forest and XGBoost and ending up with a deep learning model. This approach allowed us to evaluate and compare the effectiveness of various models in a controlled manner, ensuring that we selected the most suitable model based on performance metrics tailored to credit risk assessment. Through this project, we aimed to contribute valuable insights and tools to the field of financial risk management, supporting better decision-making processes in loan approvals and risk mitigation.

2. Exploratory data analysis

This exploratory data analysis data reveals interesting patterns that can provide insights for financial institutions, investors, companies, and policymakers.

Loan Distribution

In terms of loans, the majority are concentrated between \$10,000 and \$20,000, a range that suggests the customers are seeking moderate sums. Investors tend to finance in this bracket, which could be considered an ideal point of accessibility and appeal. The tenure of financial products is short, implying a demand for quicker settlements. Interest rates, often situated between 10% and 15%, indicate a sought-after balance between risk and return, while higher rates point to products intended for clients with higher risk or less favorable credit.

Payment Regularity

Regarding payments, monthly installments tend to vary from \$300 to \$400, although larger loans, less frequent, are also present. This could signal an opportunity for institutions to adjust their products to different installment sizes. The distribution of credit categories falls in a middle zone, indicating a medium risk profile in the loan portfolio.

Stability in the Job Market and Economic Implications

Analyzing the job market, a great diversity of positions and levels of specialization is observed. There is a predominance of individuals who remain in their jobs for extended periods, primarily in highly

qualified professions, suggesting employment stability. Interestingly, no seasonal pattern in the job market is observed, and most loans tend to be paid on time, with clear distinctions between non-performing and delayed loans, which could reflect efficient collections management or a tendency of borrowers to avoid default.

The correlations between the data suggest that specialized professions are associated with greater employment stability, while less specialized roles may have higher turnover. Finally, borrowers' ability to pay loans may be linked to overall economic health, suggesting a resilient job market.

We observe a decrease in the frequency of job titles as we move from left to right of the graph, suggesting that subsequent positions are progressively less common. The diversity of professions, which includes managers, drivers, nurses, and accountants, reflects a wide range of work fields and the distribution of job types within the studied population. This indicates a variety of available employment opportunities but may also signal fierce competition in certain areas.

Payment Regularity and Economic Health

The analysis of loan status indicates that the majority are paid punctually, with a clear distinction between those who are up-to-date and those with delays. This may be reflective of proactive debt collection management and borrowers motivated to preserve a healthy credit reputation.

Additional insights with Exploratory Data Analysis

The combination of a stable economy, efficient loan management, and a diversified job market suggests a virtuous cycle. Highly qualified professionals tend to enjoy greater job security, contributing to higher repayment rates. The absence of seasonal variation in both employment and loan repayments reinforces an overall stability that benefits both financial institutions and workers. These indicators are essential for the development of effective economic and social policies, ensuring the continuity of economic health.

Finally, the analysis of loan status shows that most loans tend to be fully paid. However, there is a clear distinction between non-performing loans and those that are overdue. This may reflect efficiency in debt management, with borrowers willing to regularize debts before entering complete default, which could be motivated by effective collection policies or the borrowers' desire to maintain a good credit reputation.

Through this exploratory analysis, we can infer that a combination of a stable economy with effective loan management and a diversified job market creates a positive cycle. Highly qualified professionals are more likely to maintain employment stability, which could contribute to a higher loan repayment rate. In contrast, the absence of a seasonal pattern in both jobs and loan repayments suggests a stability that benefits both financial institutions and workers. The ability to pay loans and employment stability can be seen as indicators of economic health and are crucial for planning effective economic and social development policies.

3. Feature Engineering & Selection

Feature Engineering and Feature Selection

Feature engineering is essential to ensure that the data is in a suitable format for the algorithms used. This process involved manipulating and transforming raw data into refined data, adding relevance to our model. By identifying, creating, and selecting the most statistically relevant variables, feature engineering has allowed for more effective capturing of patterns and underlying relationships in the data, contributing to improvements in model performance.

The steps developed in this process were as follows:

Duplicate data verification: by identifying and removing duplicate data, we ensure that each observation contributes uniquely to the analysis, avoiding overestimation of patterns or trends. However, we did not find any duplicate rows.

Missing values verification: identifying columns with missing values allows for appropriate treatment strategies to be adopted, such as value imputation, record exclusion, or even the creation of new variables to indicate the absence of data. We found that the variable "emp_title" has 9.5% missing values, and due to its insignificance, we decided to remove it.

Variable transformation: transforming data according to its typology is essential for modeling as our machine learning algorithms require numerical inputs.

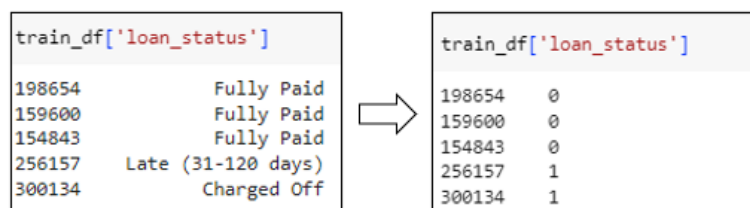


Figure 1 illustrates the transformation of the 'loan_status' variable from text to numerical values. The left table shows the original text values, and the right table shows the corresponding numerical values (0 for 'Fully Paid' and 1 for 'Late (31-120 days)' and 'Charged Off').

train_df['loan_status']	
198654	Fully Paid
159600	Fully Paid
154843	Fully Paid
256157	Late (31-120 days)
300134	Charged Off

train_df['loan_status']	
198654	0
159600	0
154843	0
256157	1
300134	1

Figure 1 Text to numbers

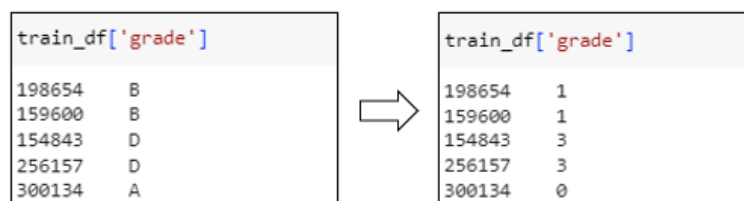


Figure 2 illustrates the transformation of the 'grade' variable from text to numerical values using a Label Encoder. The left table shows the original text values, and the right table shows the corresponding numerical values (1 for 'B', 3 for 'D', and 0 for 'A').

train_df['grade']	
198654	B
159600	B
154843	D
256157	D
300134	A

train_df['grade']	
198654	1
159600	1
154843	3
256157	3
300134	0

Figure 2 Label Encoder



Figure 3 illustrates the transformation of the 'home_ownership' variable from text to numerical values using a One Hot Encoder. The left table shows the original text values, and the right table shows the corresponding numerical values for each category (OWN, MORTGAGE, RENT).

train_df['home_ownership']	
198654	OWN
159600	MORTGAGE
154843	RENT
256157	MORTGAGE
300134	RENT

train_df.iloc[:, -4:]				
	home_ownership_ANY	home_ownership_MORTGAGE	home_ownership_OWN	home_ownership_RENT
198654	0	0	1	0
159600	0	1	0	0
154843	0	0	0	1
256157	0	1	0	0
300134	0	0	0	1

Figure 3 One Hot Encoder

```
def transform_earliest_cr_line(df):
    df['earliest_cr_line'] = df.earliest_cr_line \
        .apply(lambda d: int(d.split("-")[1])) \
        .apply(lambda y: 2000 + y if y < 30 else 1900 + y)
```

Figure 4 Date formatting

Elimination of redundant variables: the presence of redundant variables can cause multicollinearity problems, which harm the model. By eliminating redundant variables, we reduce the dimensionality of the data, making the model simpler.

Evaluation of the relationship between variables: the calculation of Weight of Evidence (WoE) and Information Value (IV) helps identify and select the most relevant and informative variables for model construction, where it is necessary to assess the relationship between independent variables and the dependent variable.

```
Showing WoE and IV for feature verification_status against target loan_status
loan_status      0      1      woe      iv
verification_status
0      0.389873  0.245897 -0.460908  0.066360
1      0.390912  0.431921  0.099761  0.004091
2      0.219215  0.322182  0.385062  0.039649
```

Figure 5 woe and iv for verification_status

Outlier removal: removing outliers can improve the accuracy and generalization of models by reducing the influence of extreme points that may negatively affect model performance.

```
Feature int_rate has 3823 outliers
Feature dti has 1684 outliers
Feature revol_bal has 16376 outliers
Feature revol_util has 16 outliers
Total of 21453 outliers or 8.64% of data
```

Figure 6 Outliers in the dataset

Multicollinearity verification: the presence of multicollinearity can lead to problems such as instability in estimated coefficients, difficulty in interpreting the effects of independent variables, and reduced accuracy of predictions. To check for multicollinearity, we use correlation heatmaps and calculate the VIF (Variance Inflation Factor), which resulted in the elimination of some variables.

Data standardization: standardization is particularly useful when input variables have different scales or when data distributions are non-Gaussian. By standardizing the data, we ensure that all variables have the same scale.

```
X_train = pd.DataFrame(scaler.fit_transform(X_train), columns = X_train.columns)
X_validation = pd.DataFrame(scaler.transform(X_validation), columns = X_validation.columns)
```

Figure 7 Scaling of the dataset

4. Modelling

In this section, we will fit the models, take the results, and briefly analyze them. The first big issue before fitting any models is recognizing we have an unbalanced dataset.

The non-defaulting class is much bigger than the defaulting class, representing over 2/3 of the data. This is problematic since we will have little recall of the positive, defaulting, cases. It is necessary, therefore, to rebalance the data.

There are various techniques to balance the data, namely under-sampling and oversampling. Under-sampling would be reducing the majority cases to match the minority, discarding some of the cases. This is not great since we would be discarding a large number of cases.

Oversampling is creating synthetic data for the minority cases to increase the number of rows. This means we will have more data to work with but at the cost of precision. We will proceed with this method.

To achieve this, we can simply use the SMOTE method (Synthetic Minority Oversampling Technique) which can be achieved with the code in Figure 8:

```
smt = SMOTE()  
X_train, y_train = smt.fit_resample(X_train, y_train)
```

Figure 8 Smote approach

After doing this, we can start with measuring the models.

Logistic Regression

Starting with the classic Logistic Regression we achieve the results seen in Figure 9:

```
Accuracy: 0.6327941223858473  
F1 Score: 0.5101975027401088  
Precision: 0.44093610698365526  
Recall: 0.6052725511192698
```

Figure 9 Results of the Logistic Regression Model

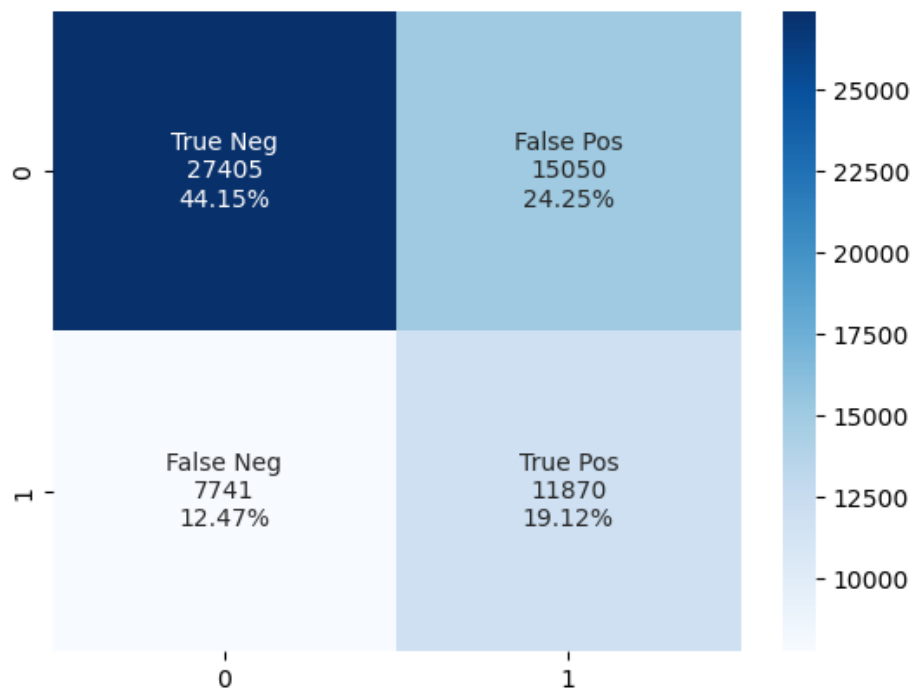


Figure 10 Confusion Matrix of the Logistic Regression Model

The model achieves an overall accuracy of 63%, indicating it correctly classified most but not most data points. Precision and F1-score suggest there's room for improvement, particularly in correctly identifying positive cases. The model performs better on class 0 compared to class 1 (default).

Machine Learning Model – Random Forest

```

Accuracy: 0.6285728095897916
F1 Score: 0.5149289847448711
Precision: 0.438346349502042
Recall: 0.6239355463770333

```

Figure 11 Results of the Random Forest Model

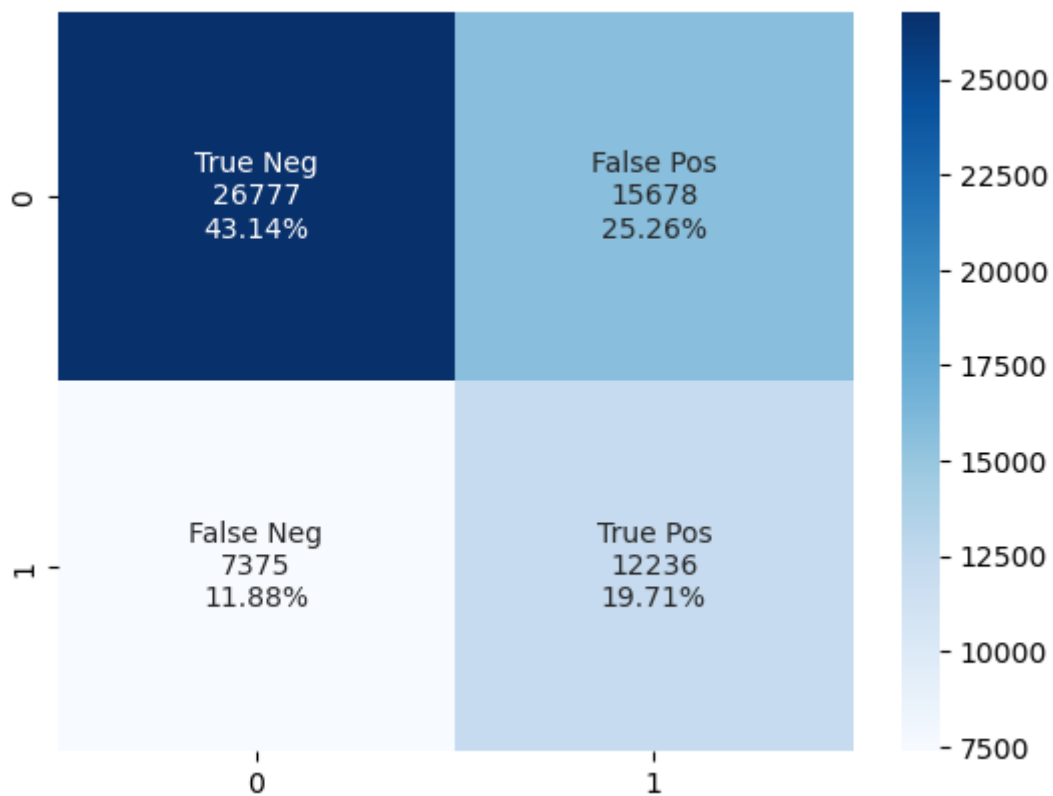


Figure 12 Confusion Matrix of the Random Forest Model

Not many changes in this model, but we do increase slightly the F1 score.

Machine Learning Model – XGBoost

We will now attempt a more sophisticated approach: XGBoost. In Figure 13, it can be seen the results obtained:

```

Accuracy: 0.8045306609093545
F1 Score: 0.7047169352090737
Precision: 0.6741327124563445
Recall: 0.7382081484880935

```

Figure 13 Results of the XGBoost Model

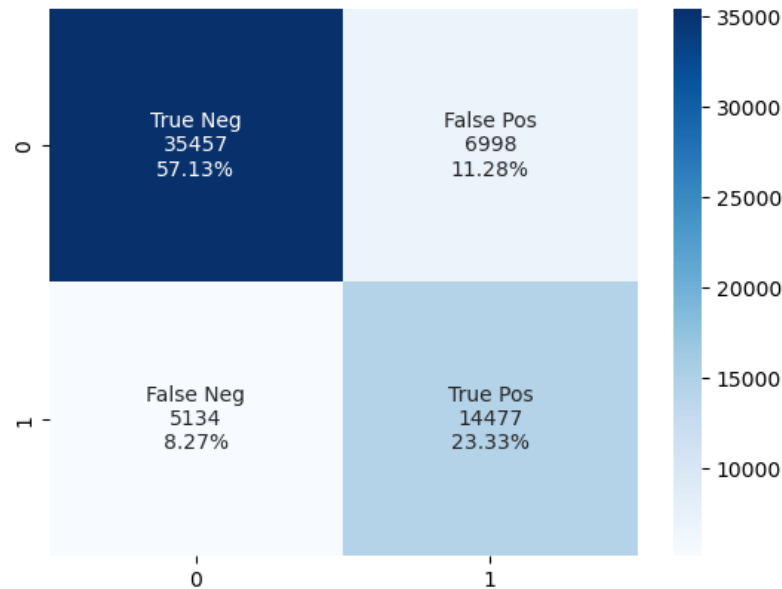


Figure 14 Confusion Matrix of the XGBoost Model

The model achieves an overall accuracy of 80%, indicating it correctly classified a significant majority of data points. F1-score (0.70) is also respectable. The model performs better on class 0 compared to class 1, with higher precision (0.87) for class 0. This will be our primary model.

Deep Learning Model

For the Deep Learning model, we have a simple network with inputs for each feature, a dropout to avoid overfitting, a feature reducing hidden layers, and a sigmoid activation function for prediction.

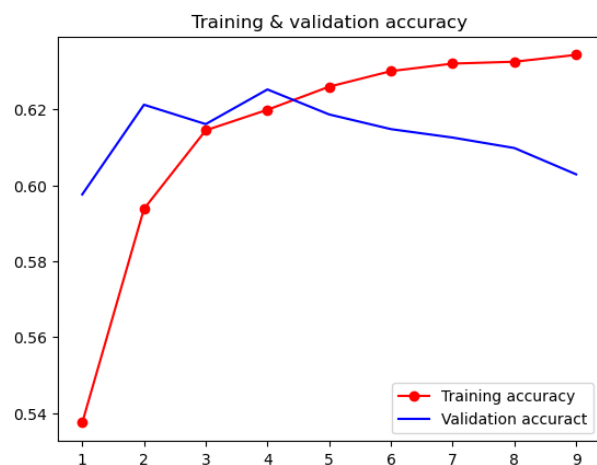


Figure 15 Training & Validation Loss

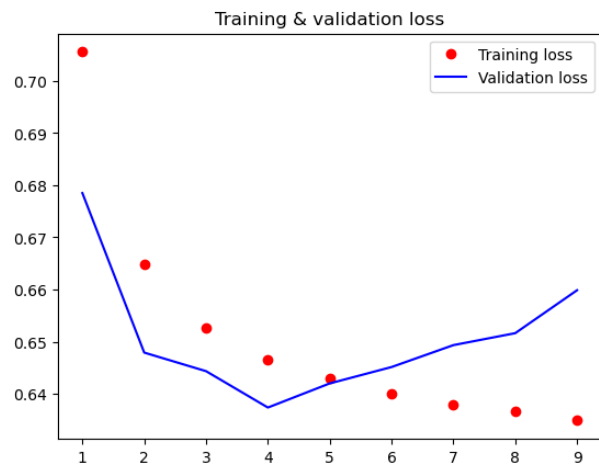


Figure 16 Training & Validation accuracy

The model is slowly learning. While accuracy is going up (from 51% to 63%) and loss is going down (from 0.74 to 0.64) over 9 epochs, the starting performance is weak, and the improvement seems gradual. In short, we have not been able to achieve remarkable results with a deep learning model.

5. Discussion

Our group project delves into the realm of modern financial risk management, where the integration of advanced machine learning techniques with rich datasets holds the promise of valuable insights and practical solutions. Our primary aim is to develop a predictive model capable of accurately assessing credit risk and predicting loan defaults. This task is vital as it directly impacts the stability and performance of financial institutions, making it a significant focus of study in the field.

One of the main challenges we encountered was dealing with imbalanced data, a common issue in credit risk modeling where defaults are much less frequent than non-default instances. To tackle this challenge, we utilized a technique called SMOTE, which helped rebalance the dataset and improve the performance of our predictive models, although introducing a high degree of synthetic data in our dataset.

Our approach was systematic, starting with simpler models like logistic regression and gradually progressing to more complex algorithms like Random Forest and XGBoost, later going to a deep learning model. By comparing and evaluating these models, we were able to identify XGBoost as the most effective for our purposes, consistently outperforming the others in terms of accuracy and reliability.

XGBoost's superior performance on the Lending Club dataset can be attributed to its ability to handle complex relationships efficiently, effectively capturing nonlinear interactions in the data. Its ensemble nature and regularization techniques prevent overfitting, ensuring robust predictions even in noisy financial datasets. While random forest models are powerful, XGBoost optimizes both bias and variance more effectively. Deep learning models may struggle due to dataset size and interpretability requirements in financial applications. Thus, XGBoost emerges as the top choice for predictive modeling in this context.

Overall, our project not only provided us with valuable direct experience in applying machine learning techniques to real-world financial data but also highlighted the potential of these methods in enhancing risk management practices. Our findings could have practical implications for financial institutions, informing better decision-making processes related to loan approvals and risk mitigation. This project represents a significant step in our journey as aspiring data scientists and analysts, equipping us with the skills and knowledge necessary to tackle similar challenges in our future careers.

6. Conclusion

Throughout this project, our group has successfully navigated the complexities of a detailed financial dataset and applied a variety of predictive modeling techniques to assess credit risk effectively. Through rigorous comparison and evaluation, we identified that the XGBoost model consistently outperformed other techniques, offering superior predictive accuracy and robustness. This success was achieved despite facing significant challenges, such as addressing the imbalance in the data which is a common issue in credit risk modeling, and optimizing model parameters to enhance performance. Our collaborative efforts not only led to the development of a highly effective predictive model but also fostered a deep understanding of the practical applications of machine learning in the domain of financial analysis. The direct experience gained through this project has been invaluable, enhancing our analytical skills and preparing us for future challenges in data science and related fields. Moreover, the findings and methodologies refined during this study hold significant potential for real-world application within financial institutions. The insights derived could inform better risk assessment practices and decision-making processes related to loan approvals, thereby reducing financial risks, and facilitating more informed lending practices.