STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES

Credit Risk Modeling Using Logistic Regression: A Replication and Extension of Costa e Silva, Lopes, Correia, and Faria (2020)

LILIBERT NAMAYENGO - 196543
FLEMY KABALISA - 121580
CYNTHIA KIAMBI - 204843
ESMIE ABAALUK - 204946

6th May 2025

Contents

List of Figures

Chapter 1

Introduction

The Credit Risk project replicates and extends the methodology presented in the Portuguese credit risk modeling paper Costa e Silva et al. (2020), applying logistic regression models to predict the probability of default using the German Credit dataset UCI Machine Learning Repository (1994). The objective of this paper is to understand the significant drivers of credit risk, assess the performance of the model, and provide insights based on statistical tests and evaluation metrics.

1.1 Understanding of the Paper

The original Portuguese study examines the predictive power of the logistic regression model in modeling credit risk using various demographic and financial variables.
The variable of default is a binary variable Y such that $Y = 1$ if defaulted, and 0 otherwise.Using the logistic regression model, the PD is a function of a set of explanatory variables X as follows:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-\beta x}} \tag{1.1.1}$$

Estimation of regression coefficients β Maximum Likelihood Estimation (MLE) was used.
‘

The methodology used:

1. **Univariate Selection - Statistical Inferences**

   To infer the factors that influence credit risk, the authors analyzed using statistical inference techniques, such as the Mann–Whitney–Wilcoxon and Pearson Chi-squared independence tests.

   - The nonparametric Mann-Whitney-Wilcoxon test was used to compare the medians of each variable at a 5% level of significance between the medians in the groups of defaulters and non-defaulters.
   - The Pearson Chi-squared independence test at a 5% level of significance was used to check if the qualitative variables have some influence on the likelihood of probability of default.

2. **Model Training**: The authors trained the model in 80:20. For the selection of the suitable model,the logistic regression model was refined iteratively with the aim of achieving the lowest AIC and a model of good fit using the likelihood ratio test.

3. **Model Validation**: The model was evaluated using:

    - Deviance measures - Pearson and Deviance residuals.
    - Test of good fit using the Hosmer–Lemeshow test that is performed by sorting the n observations by the predicted probabilities, and forming g groups with approximately the same number of subjects in each group (m).
    - Model performance using Confusion matrix and ROC Curve.

1.2 Results

This Portuguese dataset is a simple randoms ample of all the banking institution records, composed of 3221 individuals, where 319 defaulted, making an observed default rate of 10%. The dataset was imbalanced. The explanatory variables found relevant were Spread, Term, Age, Credit Cards, Salary and Tax Echelon.

Default Observed	**Predicted 0**	**Predicted 1**	**Total**
0	2313	251	2564
1	11	2	13
Total	2324	253	2577

TABLE 1.1 Confusion Matrix

The confusion matrix obtained show a higher overall accuracy of 89.9% and a lower default detection rate of 15.4%.

This project mirrors these steps and further investigates interaction effects, threshold optimization, and the trade-off between model complexity and predictive power.

Chapter 2

German Dataset

2.1 Dataset Description

The project analyzes the German Credit Dataset to cluster customers and classify loan applicants as either good or bad credit risks. The dataset comprises 1000 entries with 20 attributes, which include both categorical and numerical data.

The attributes are as describes as:

- Numerical Data (7)
- Categorical Data (13)

We are working with an imbalanced dataset. Bad:Good credit is equivalent to 3:7 respectively.

The attributes are as follows:

- **Age**: Age of the individual
- **Sex**: Gender of the individual(male/ female)
- **Job**: Job type(0-unskilled and non resident, 1- unskilled and resident, 2-skilled , 3-highly skilled)
- **Housing**: Housing status(own, rent, free)
- **Saving accounts**: Saving accounts status(little, moderate, quite rich, rich)
- **Checking account**: Checking account status(little, moderate,rich)
- **Credit amount**: Amount of credit in DM
- **Duration**: Duration of the credit in months
- **Purpose**: Purpose of the credit(car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others)
- The target variable is the **credit risk**, which is binary(good/bad)

The rest of the variables are as documented within the German Credit text file attached in GITHUB.

2.2 Data Pre-processing Techniques

Before modelling, the necessary pre-processing steps were applied, including :

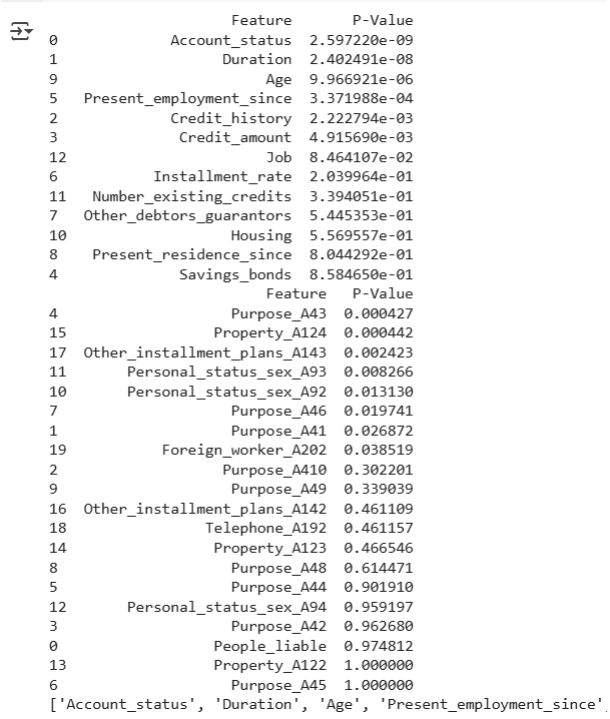
- Data cleaning - The data we had didn't have missing values. However, conversion of the data types to the appropriate data types.
- Encoding categorical variables - For ordinal variables we applied label encoding while for nominal and binary variables we used one hot encoding. This avoids implying any kind of rank or order.
- Exploratory Data Analysis(EDA)
- Normalization - Features are on different scales for instance credit amount, age, number of credit cards.
- Scaling ordinal features.
- Splitting the dataset into training and testing sets. We replicated the ratio used 80:20 respectively.

Chapter 3

Methodology

3.1 Feature Selection

We employed two statistical inference techniques for the initial feature selection, the Mann- Whitney- Wilcoxon and Pearson Chi-squared independence tests for continuous and categorical variables respectively, at a significance level of 5%. The same techniques used in the Portuguese Banking Institution Dataset.
The figure indicates the P-value for the variables.



	Feature	P-Value
0	Account_status	2.597220e-09
1	Duration	2.402491e-08
9	Age	9.966921e-06
5	Present_employment_since	3.371988e-04
2	Credit_history	2.222794e-03
3	Credit_amount	4.915690e-03
12	Job	8.464107e-02
6	Installment_rate	2.039964e-01
11	Number_existing_credits	3.394051e-01
7	Other_debtors_guarantors	5.445353e-01
10	Housing	5.569557e-01
8	Present_residence_since	8.044292e-01
4	Savings_bonds	8.584650e-01

	Feature	P-Value
4	Purpose_A43	0.000427
15	Property_A124	0.000442
17	Other_installment_plans_A143	0.002423
11	Personal_status_sex_A93	0.008266
10	Personal_status_sex_A92	0.013130
7	Purpose_A46	0.019741
1	Purpose_A41	0.026872
19	Foreign_worker_A202	0.038519
2	Purpose_A410	0.302201
9	Purpose_A49	0.339039
16	Other_installment_plans_A142	0.461109
18	Telephone_A192	0.461157
14	Property_A123	0.466546
8	Purpose_A48	0.614471
5	Purpose_A44	0.901910
12	Personal_status_sex_A94	0.959197
3	Purpose_A42	0.962680
0	People_liable	0.974812
13	Property_A122	1.000000
6	Purpose_A45	1.000000

['Account_status', 'Duration', 'Age', 'Present_employment_since'

FIGURE 3.1 Feature Selection

Selection refined based on Wald test results from initial logistic regression. The variables for which the null hypothesis of the Wald test is rejected, at a significance level of 5%, and therefore are significant covariables in the model, are:

Significant variables at 5% level based on Wald test:

	Coef.	Std.Err.	z	P>|z|
Account_status	0.331346	0.085164	3.890664	0.000100
Duration	0.339009	0.106520	3.182573	0.001460
Age	-0.340752	0.098782	-3.449529	0.000562
Purpose_A43	-0.782922	0.208939	-3.747128	0.000179
Property_A124	0.724537	0.246893	2.934616	0.003340
Other_installment_plans_A143	-0.515924	0.209513	-2.462487	0.013798
Purpose_A41	-1.093338	0.323789	-3.376699	0.000734

FIGURE 3.2 Final Model Feature Selection

3.2 Iterative Models

Three iterative logistic regression models were developed:

- Model 1: Full model with all predictors.
- Model 2: Model using variables selected from statistical tests.
- Model 3: Model including interactions between key quantitative and qualitative variables.

3.3 Evaluation Metrics

Goodness-of-fit assessed via:

- Hosmer–Lemeshow Test
- Residual Analysis (Pearson and Deviance residuals)

Predictive performance evaluated using:

- Confusion Matrix
- ROC Curve and AUC Score

Chapter 4

Results & Analysis

4.1 Selected Variables

The final model identified key co-variates that significantly influence credit risk, as determined through the Wald test, with the model specification guided by the Mann-Whitney-Wilcoxon and Pearson Chi-Square variable selection techniques.

Notably,the final variables chosen differ with the Portuguese selected variables. Largely driven by difference in data sources and the macroeconomic conditions like unemployment, inflation, and policy shifts due to difference time collection period. The Portuguese dataset spans January 2008 to December 2009, a period marked by financial crisis impacts, whereas the German dataset was collected in 1994, under different economic circumstances.

Key Significant Predictors of Default:

- *Account_status, Duration & Age (Qualitative Factors)*
- *Purpose_A43 (Television/Radio) and Purpose_A41 (Used Car)*:

  These loan purposes fall under consumption-related borrowing, often not backed by income-generating assets. This suggests that borrowers seeking credit for depreciating goods have a higher likelihood of default, consistent with consumer over-extension theory.

- *Property_A124 (No property)*:

  Borrowers who do not own property lack tangible collateral, which not only weakens their bargaining position with lenders but also reduces recovery prospects in case of default. This aligns with increased credit risk.

- *Other_Installment_Plans_A143 (None)*:

  Surprisingly, individuals without any existing installment plans (i.e., no current borrowing track record) are flagged as riskier. This may reflect thin credit files,

a known concern in retail lending where lack of past credit data limits accurate assessment of repayment behavior.

Question:

For borrowers with other installment plans through banks or stores — does this reflect high financial leverage? A more granular analysis incorporating DTI (debt-to-income) ratios and payment behavior across different credit types could further clarify this relationship.

Interaction Between Variables

We considered interactions between the quantitative and qualitative variables present in the best model found in the previous section. The best model in the previous section is the one with the lowest AIC.

4.2 Model Chosen

Although this reduced model has a slightly higher AIC (873.83) than the full model (872.22), we chose the reduced model due to:

- Greater parsimony: Fewer, more interpretable variables
- Stronger statistical significance across selected co-variates.
- More stable estimation, with reduced multi-collinearity.
- Meaningful insights aligned with economic theory and credit risk frameworks

In line with the Portuguese paper, we favor statistical robustness and interpretability over mere goodness-of-fit. This model serves as a reliable foundation for policy and credit decision-making..

4.3 Model Evaluation

4.3.1 Goodness of Fit

Hosmer–Lemeshow Test:

$\text{Chi}^2 = 15.922$, df = 8, p-value = 0.0435.

The p-value is just below 0.05, which suggests that there is a statistically significant difference between the observed and predicted outcomes.

This may indicate the model does not calibrate perfectly — the predicted probabilities deviate from actual default rates in some bins. However, since the p-value is only marginally below 0.05, the model is still usable, though improvement is possible.

4.3.2 Residual Analysis

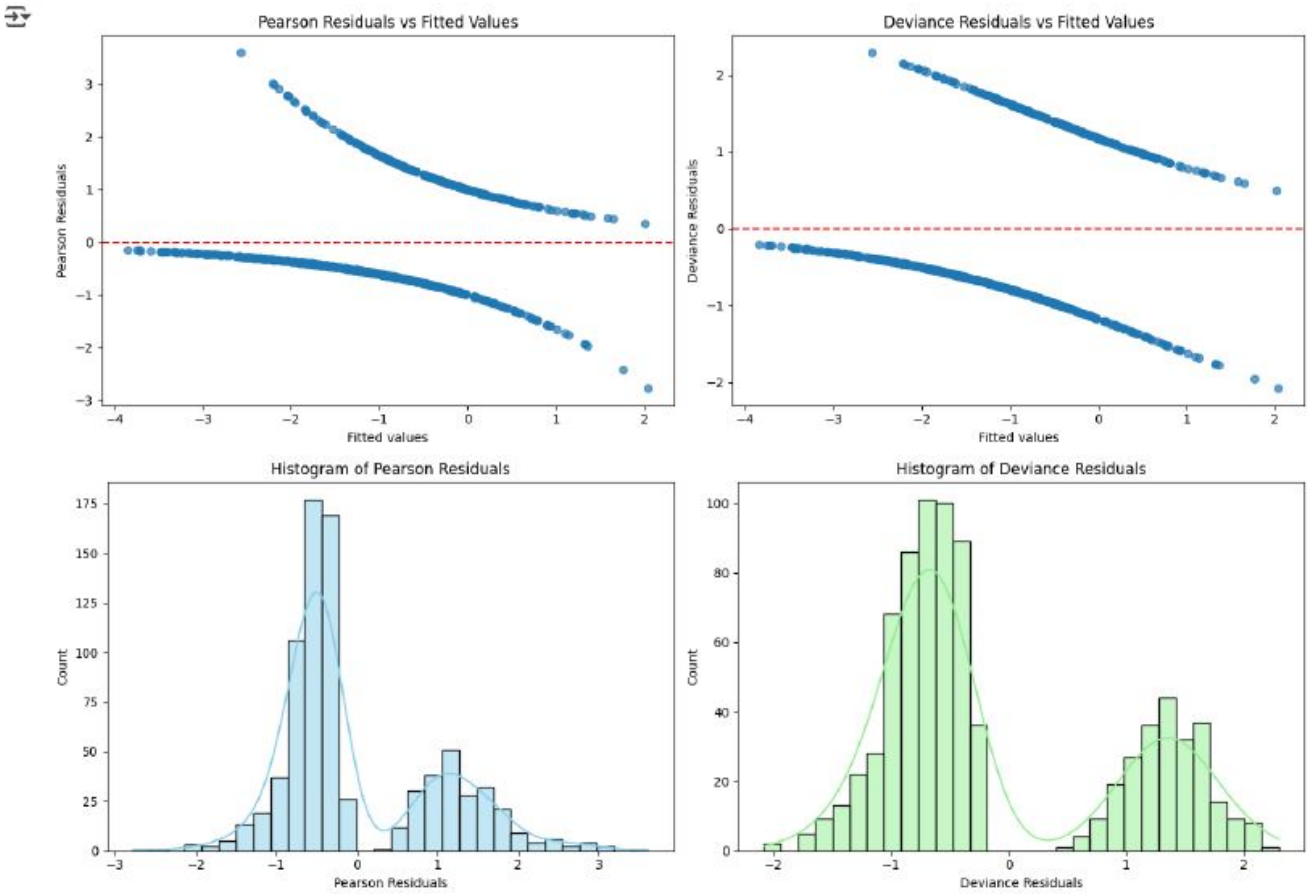


FIGURE 4.1 Residual Analysis

The residuals are mostly centered and reasonably dispersed, but the slightly negative means and small variance deviations might hint at minor model mis-specification or imbalance in the classes.

TABLE 4.1 Summary Statistics of Residuals

Residual Type	Mean	Variance
Pearson Residuals	−0.017	0.943
Deviance Residuals	−0.109	1.044

4.3.3 Precision & Recall

The model is better at identifying non-defaulters than defaulters.

High Recall (0.91) for Class 0 - Most non-defaulters are correctly identified.

Low Recall (0.28) for Class 1 - The model misses many actual defaulters.

Class	Precision	Recall	F1-Score	Support
0	0.75	0.91	0.82	559
1	0.57	0.28	0.38	241

TABLE 4.2 Classification report showing performance metrics for each class.

F1-score (0.38) for defaulters suggests weak performance in detecting risky clients.

In credit risk business, It is worse to class a customer as good when they are bad, than it is to class a customer as bad when they are good. We need to improve the recall rate to a higher value from 38%.

This can be done by adjusting the classification threshold. We adjusted to 0.35 so as to improve the recall rate for defaulters as we are dealing with an imbalanced datasets.

Class	Precision	Recall	F1-Score	Support
0	0.80	0.78	0.79	141
1	0.50	0.53	0.51	59

TABLE 4.3 Classification report showing adjusted classification threshold.

Lowering the threshold from the default 0.5 to 0.35 increased sensitivity (recall) for class 1 (defaulters) from 28
However, precision dropped, meaning more false alarms (non-defaulters flagged as defaulters).

We did a trade-off between recall and precision.

For defaulters (sensitivity) is often prioritized to minimize missed risks in credit risk.

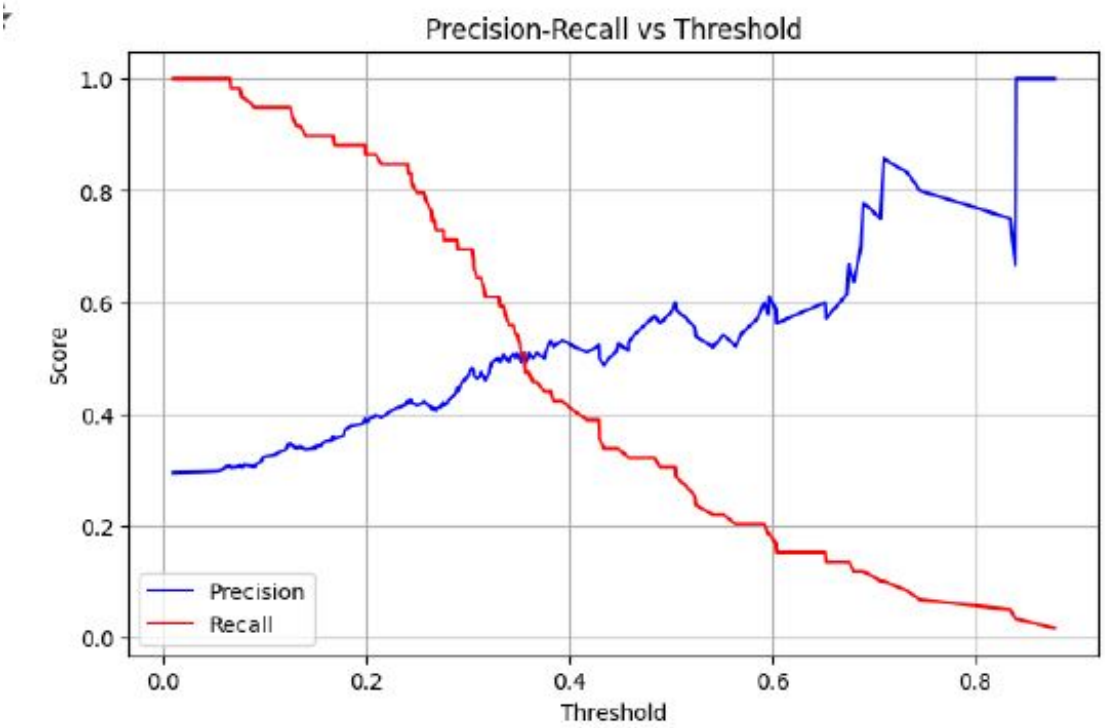


FIGURE 4.2 Precision & Recall Threshold

4.3.4 ROC Curve

AUC of 0.7553 is moderately good and suggests decent discriminatory power. The prior ROC before adjustment of the threshold. After, change in the threshold the AUC went low to 0.72.

[b]0.45

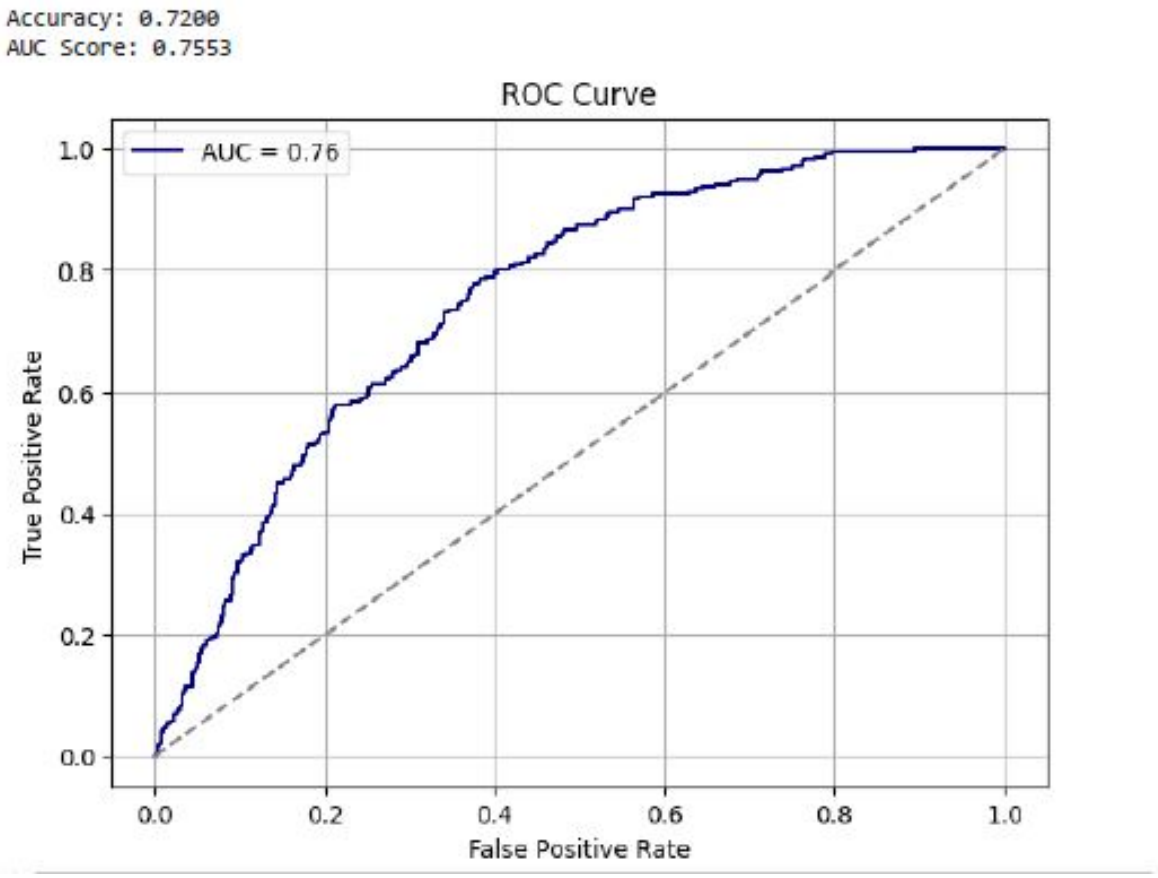


FIGURE 4.3 ROC Before Adjustment

[b]0.45

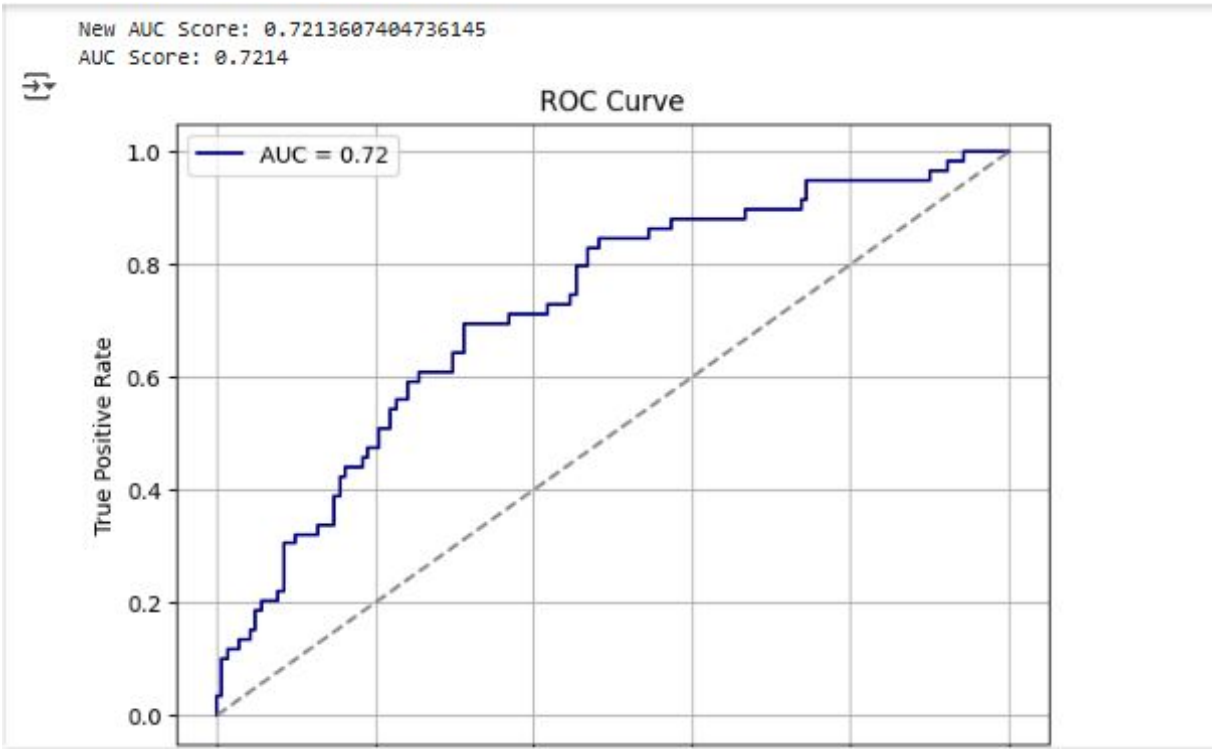


FIGURE 4.4 ROC After Adjustment

FIGURE 4.5 Comparison of ROC Curves Before and After Threshold Adjustment

Chapter 5

Conclusion

5.1 Recommendations

- Future models should explore advanced machine learning techniques like: Support Vector Machines(SVM) and gradient boosting techniques(example, XGBoost)
- Feature engineering and additional data enrichment could further improve model accuracy.
- Alternative sampling strategies.
- Periodic model recalibration to adapt to changing borrower behaviors and economic conditions

Bibliography

Costa e Silva, E., Lopes, I. C., Correia, A., & Faria, S. (2020). A logistic regression model for consumer default risk. *Journal of Applied Statistics*, *47*(13-15), 2879–2894. doi: 10.1080/02664763.2019.1694166

UCI Machine Learning Repository. (1994). *German credit data.* `https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)`, note = Accessed: 2025-05-05, institution = University of California, Irvine, School of Information and Computer Sciences.