BDA 620 Data Mining

Project Title: Predictive Modeling for Loan Default Risk Management

Prepared By:
Jun Ming Li – MUID: 11022435
Jairo Onate – MUID: 11062465
Mazen Alhaffar – MUID: 11054744

Submitted To:
Ehsan Ahmadi, Ph.D.
Assistant Professor of Management Science
Stetson-Hatcher School of Business
Mercer University

December 12, 2024

Table of Contents

1. Introduction 3

2. Literature Review 3

3. Identifying data, data sources, and data characteristics 3

3.1. Dataset collection 3

3.2. Cleaning and Preprocessing the Data 4

3.3. Data Dimension Reduction 4

4. Methodology 5

4.1. Logistic Regression 5

4.2. Random Forest 5

5. Empirical Results 5

6. Conclusions and Recommendations 6

7. Appendix 8

8. References 13

1. Introduction

The project, “Predictive Modeling for Loan Default Risk Management,” aims to predict customer loan defaults in the banking industry, enabling data-driven decision-making. By identifying high-risk borrowers, banks can mitigate financial losses and improve credit assessment processes. Predictive modeling streamlines operations, reducing costs while enhancing customer relationships through personalized solutions. Additionally, these insights empower banks to gain a competitive edge in managing risks effectively, ensuring smarter resource allocation and more robust financial stability. This project emphasizes the role of advanced analytics in driving efficiency and optimizing risk management strategies.

2. Literature Review

The literature review explores key factors influencing loan defaults, integrating insights from various studies to inform predictive modeling. Demographic vulnerabilities such as race, gender, and education are highlighted by Pew Charitable Trusts [6], underscoring the need for targeted support for at-risk groups. Himberg and Wang [5] demonstrate the predictive power of machine learning algorithms like XGBoost, emphasizing critical variables such as employment history and loan-to-value ratio. Chan et al. [3] reveal that high loan-to-value ratios and nonrecourse loans increase default risks, where borrowers often prioritize liquidity by paying off credit card debt and auto loan payments over housing loans. Akomas [2] highlights geographical disparities, with regions offering institutional support showing reduced default rates, while Agrawal and Maheshwari [1] focus on industry-specific risks, introducing the "industry beta" metric—calculated from the relationship between a firm's stock returns and its industry index—to assess corporate default probabilities. Together, these studies provide a comprehensive framework for identifying predictors of loan default, enabling data-driven strategies for risk mitigation and improved credit assessment.

3. Identifying data, data sources, and data characteristics

3.1. Dataset collection

This case study uses data that has been sourced from **Kaggle** [4], a platform recognized for being a dataset repository that can be used for data analysis and research.

The dataset used for the research is entitled ***Loan Defaulter*** by Gaurav Dutta and it aims to showcase the importance of understanding how risk analytics can be used to minimize the odds of a lender losing money if customers default on their financial compromise. All files from the source were downloaded in CSV format and subsequently imported into R Studio for further exploration. These files originally have

been divided into current application and previous application datasets, where the former one contains all the information about the client at the time of the application and whether it has payment difficulties. The previous application file encloses information about the client's previous loan data in terms of the status of the application as if it has been Approved, Cancelled, Refused, or Unused offer.

A more detailed explanation of the collection process and data characteristics can be found in the sections below. In addition, the dataset dictionary is included as an appendix which can be used for future reference (Table 5).

3.2. Cleaning and Preprocessing the Data

Initially from the application data, all records were divided into two groups, one containing the categorical variables and another with the numerical variables. From the numeric variables, the dataset contained 8,388,094 missing values out of 22,755,814 cells distributed in 307,511 records along 74 variables. All the negative values stored in the original dataset "application data" were converted to positive by applying the absolute function to each. In addition, numeric variables containing missing values excluding those being binary categorical variables were standardized by computing the mean and replacing it respectively in each variable.

For the previous application dataset, it contained 1,670,214 records and 37 variables with 10,288,585 missing values out of 61,797,918 cells, representing 16.64% of the total cells. Applying the procedure of saving the index of the numeric missing values and replacing it with the mean, led to an 86.09% reduction of the missing data equivalent to only 1,430,816 cells that needed further preprocessing as they were still missing. The processing of missing categorical values was done by retrieving the levels of the classes belonging to each categorical variable and subsequently saving the index of the missing values. Using random sampling to avoid bias, the missing records were shuffled, reclassified into one of the existing classes of that categorical variable, and equally added to that class's total.

3.3. Data Dimension Reduction

To reduce the dimension of the variables, similar classes were combined into a major group. This adheres to the principle of parsimony, as each variable is easier to interpret, and the records tend to be more homogeneous among each other. Then, using our literature review and common knowledge of financial metrics that banks typically base creditworthiness on, an initial group of 45 predictors and our target variables were chosen for an initial run of the Logistic regression.

4. Methodology

The current and previous application datasets were joined together using the SK ID found in both datasets. After joining the datasets, a new variable, Loan to Goods ratio was developed to measure how much over or under an applicant's loan was in comparison to the cost of the item being purchased. Using histograms, boxplots, and the data statistics, the appropriate values were chosen to replace missing values and outliers. This was done to preserve the robustness of the data set. The dataset was then partitioned into a training data set (70%) and testing data set (30%). The training data set was then balanced using "under" sampling. "Under" sampling was chosen due to the limitation of current hardware to run and compute more record intensive models.

4.1. Logistic Regression

For the Logistic regression model, the data was scaled with z-score standardization to minimize the effects of disproportionately large range values among the different variables. The accuracy of the initial model was run against the testing data set using a confusion matrix. To improve the predictability of the model, a second round of data reduction was conducted after further review of the variables. Moreover, variables that banks cannot feasibly filter or have control over, such as age and sex, were also removed. Finally, a third model was created after using forward stepwise regression to remove statistically insignificant variables, lowering the total count of predictors to 15. A summary of the Logistic model statistics is summarized in Table 1.

4.2. Random Forest

Random Forest using the variables from the 15-predictor Logistic model was run under the following conditions: a limit of 100 trees, with 25 nodes per bucket, and a random selection of 4 variables. To increase the robustness of our research, additional models were also run with 5, 6, 8, 10, and 16 variables selected. The results are shown in Table 3 and the trend of their AUC and accuracy are also shown in Figure 2 and Figure 3. Variable importance was also recorded for each model, and number of days an individual was employed, their age, and the loan-to-goods ratio were consistently the top 3 most important features in the additional models as well. Figure 1 is an example of our variable importance plots from each run and Table 4 shows the odds of defaulting.

5. Empirical Results

The initial Logistic model using 45 predictors has a 27.43% accuracy rate, with an AUC of 0.6199, showing that it is better than a random guess. The second model used 24 predictor variables and had an

accuracy of 58.02% and an AUC of 0.6552. The third model resulted in 56.65% accuracy and 0.6445 AUC. The sensitivity was 0.5584 and the specificity was 0.6509.

The Logistic Regression model with 23 predictors had a fair performance, with an accuracy of 58.02% in detecting loan defaulting. It is important to mention that an AUC of 0.6552 indicates that the model can predict better than a random guess. Moreover, the 23-predictor model has better accuracy, sensitivity, and specificity than the other 2 models. However, we ultimately chose to continue forward with the 15- predictor model and use those are the basis for our random forest. The conclusion section will elaborate on this further.

The Random Forest Model with 4 variables selected has an accuracy of 76.71%, AUC of 0.7637, sensitivity of 0.7678, and specificity of 0.7596. When compared to the additional models ran (Table 3), there was a general trend of higher variable selection per node split correlating with higher AUC and accuracy. However, there is a plateau effect past the 8-variable selection. The gain in AUC has a more horizontal slope and the effect is even more pronounced in the accuracy graph (Figure 2 and Figure 3). For the important variables, it was shown that a one unit increase in Age decreased the odds of defaulting by nearly 77%. A one unit increase in days employed resulted in the odds of defaulting decreasing by 96.6%, and a one unit increase in loan-to-good ratio increased the odds of defaulting by 120%.

6. Conclusions and Recommendations

After running the Logistic Regression 3 times with a combination of 45 predictors for the initial run, 23 predictors for the second run, and 15 predictors for the third run, the model's best performance in terms of accuracy, sensitivity, specificity, and Area Under the Curve (AUC) was with 23 predictors. While the Logistic model using 23 predictors proves to be slightly better in all 4 fields of measurement, we ultimately decided to pick the 15-predictor model as our final Logistic model and to use those 15 predictors as the basis for our random forest analysis. Adhering to the parsimony principle, having a smaller set of predictors will reduce the complexity of the model and enhance its interpretation from a managerial and client perspective. Thus, we recommend that banks use the 15-predictor model when using a Logistic approach to classifying default vs. non-defaulting clients.

When identifying loan defaulting versus non-defaulting transactions, the Random Forest model with 15 predictors achieved the best results overall, with an accuracy of 79.84%, sensitivity of 79.87%, specificity of 78.56%, and AUC of 79.71%. Following the principle of parsimony, the final Random Forest model chosen has 8 predictors with an admirable performance, attaining 79.71% accuracy, 79.82%

sensitivity, 78.56% specificity, and 79.19% AUC. Including more predictors does not significantly increase the performance of the model. Reducing the complexity of the model is more advantageous for evaluating real-time applications where transparency and computing efficiency are crucial. Implementing this classification model is a useful tool for commercial applications since it strikes a compromise between interpretability and accuracy, particularly in situations when decision-making clarity is required. Thus, we recommend that banks use the model that randomly selects 8 variables when using a random forest approach to classifying default vs. non-default clients.

To enhance the robustness of our research and refine our results, we propose integrating additional models into our analysis, particularly XGBoost. This model, utilized by Himberg and Wang [5], demonstrated its effectiveness in identifying key predictors of loan defaults, with employment history and loan-to-goods ratio emerging as the most significant variables, which aligns closely with the findings of our study.

7. Appendix

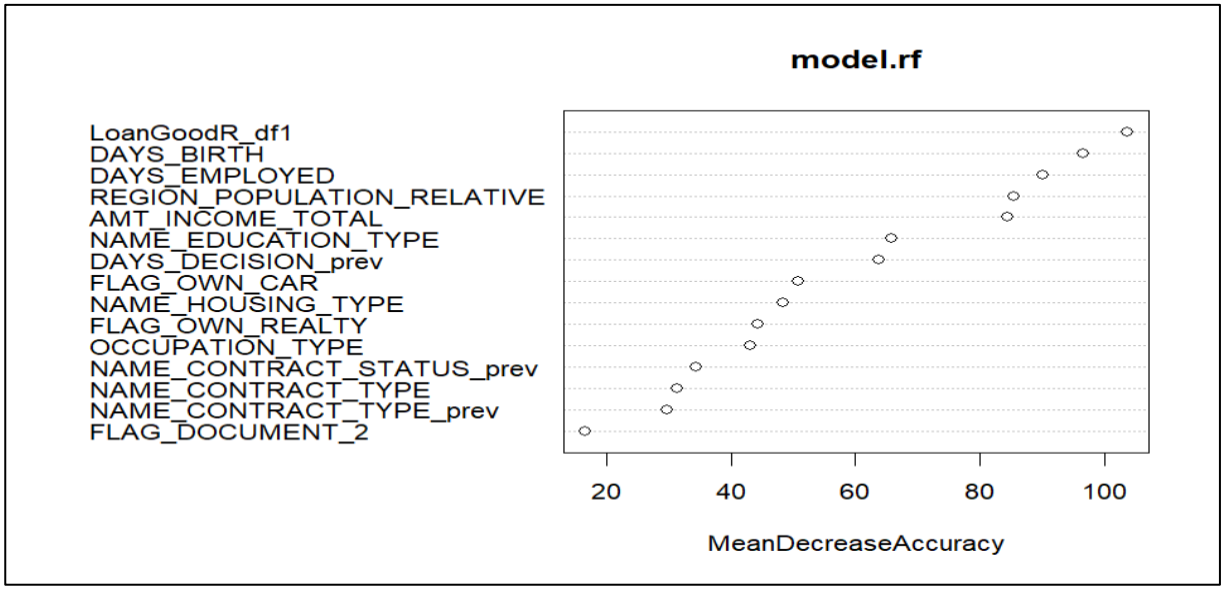

Figure (1): Variable Importance, 4 Variable Selection Random Forest

Logistic Regression			
	Initial Run, 45 predictors	2nd Run, 23 predictors	3rd Run, 15 predictors
Accuracy	0.2743	0.5802	0.5665
Sensitivity	0.214	0.5732	0.5584
Specificity	0.9094	0.6538	0.6509
AUC	0.6199	0.6552	0.6445

Table (1): Performance of Logistic Regression Runs

Variable	Coef	P-Value
(Intercept)	-0.2471	1.40905E-19
NAME_CONTRACT_TYPERevolving loans	-0.3511	3.42946E-57
FLAG_OWN_CARY	-0.1473	4.05604E-41
FLAG_OWN_REALTYY	0.0187	0.105418779
AMT_INCOME_TOTAL	0.0577	0.013808451
NAME_EDUCATION_TYPEHigh School	0.3851	3.2926E-201
NAME_EDUCATION_TYPEMiddle School	0.6015	9.76548E-41
NAME_HOUSING_TYPEOwned	-0.2273	4.12473E-22
NAME_HOUSING_TYPERenting	-0.1164	0.000111177
REGION_POPULATION_RELATIVE	-0.0992	1.93536E-84
DAYS_BIRTH	-0.2663	0E+00
DAYS_EMPLOYED	-0.0350	6.74062E-08
OCCUPATION_TYPEGeneral Labor	0.2574	2.1266E-128
FLAG_DOCUMENT_21	2.7635	1.93157E-07
NAME_CONTRACT_TYPE_prevConsumer loans	-0.1206	6.7909E-20
NAME_CONTRACT_TYPE_prevRevolving loans	0.0343	0.032976937
NAME_CONTRACT_STATUS_prevCanceled	0.0596	0.000123008
NAME_CONTRACT_STATUS_prevRefused	0.4016	7.6118E-191
NAME_CONTRACT_STATUS_prevUnused offer	0.0658	0.10641626
DAYS_DECISION_prev	-0.0894	4.61086E-53
LoanGoodR_df1	0.1818	4.7443E-251

Table (2): Performance of 15-Predictor Logistic Regression Run

	Random Forest					
	4 Variables Selected	5 Variables Selected	6 Variables Selected	8 Variables Selected	10 Variables Selected	16 Variables Selected
Accuracy	0.7671	0.7836	0.7877	0.7971	0.8004	0.7984
Sensitivity	0.7678	0.7851	0.7888	0.7982	0.8013	0.7987
Specificity	0.7596	0.7685	0.777	0.7856	0.7904	0.7954
AUC	0.7637	0.7768	0.7829	0.7919	0.7959	0.7971

Table (3) Performance of Random Forest Models

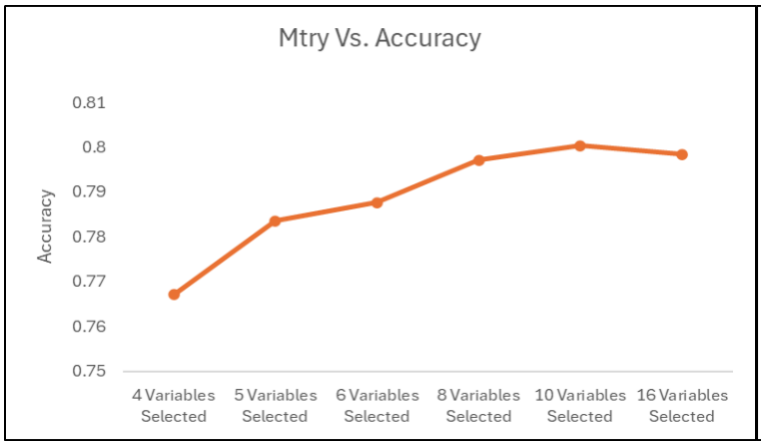

Figure (2) Variable Selected vs. Accuracy

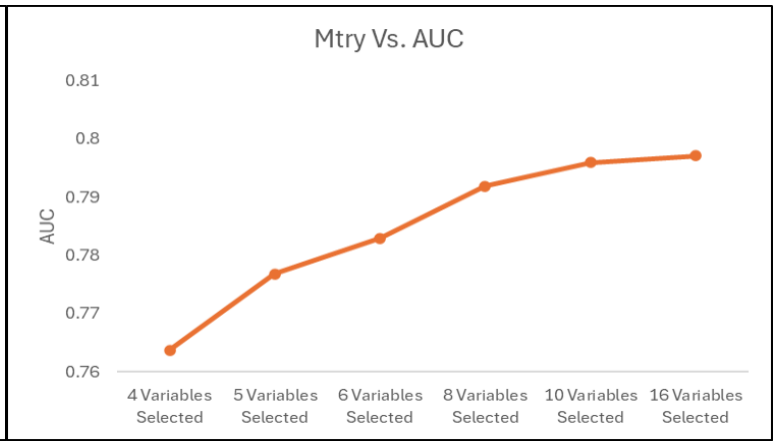

Figure (3) Variable Selected vs. AUC

Variable	Odds
Age	-76.62
Days Employed	-96.56
Loan to Good Rati	119.94

Table (4): Performance of Random Forest Models

Data Dictionary	
NAME_CONTRACT_TYPE	Identification of loan is cash or revolving
FLAG_OWN_CAR	Flag if the client owns a car
FLAG_OWN_REALTY	Flag if client owns a house or flat
AMT_INCOME_TOTAL	Income of the client
NAME_EDUCATION_TYPE	Level of highest education the client achieved
NAME_HOUSING_TYPE	Housing situation of the client (renting, living with parents, ...)
REGION_POPULATION_RELATIVE	Normalized population of region where client lives (higher number = more populated region)
DAYS_BIRTH	Client's age in days at the time of application
DAYS_EMPLOYED	Days before the application the person started current employment
OCCUPATION_TYPE	Client's occupation type
FLAG_DOCUMENT_2	Indicates if document 2 was provided
NAME_CONTRACT_TYPE_prev	Contract product type (Cash loan, consumer loan [POS], ...) of the previous application
NAME_CONTRACT_STATUS	Contract status (approved, cancelled, ...) of the previous application
DAYS_DECISION	Relative to current application when the decision about previous application was made
LoanGoodR_df1	Loan to Goods Price Ratio for current applications
TARGET	Target variable (1 = payment difficulties, 0 = no difficulties)

Table (5): Data Dictionary

	DAYS_EMPLOYED	Previous application DAYS_DECISION	Loan Good Ratio
Min	0.00	1.00	0.15
Mean	72663.47	880.37	1.12
Max	365243.00	2922.00	6.00

Table (6): Summary Statistics 1

	AMT_INCOME_TOTAL	REGION_POPULATION_RELATIVE	DAYS_BIRTH
Min	25650.00	0.00	7489.00
Mean	173316.04	0.02	16321.05
Max	117000000.00	0.07	25201.00

Table (7): Summary Statistics 2

Variable	Categories	Count
Target	0	1,291,341
	1	122,360
Missing document 2	0	1,413,601
	1	100
Realty Ownage	N	389,609
	Y	1,024,092
Car Ownage	N	937,176
	Y	476,525
Education	College+	358,635
	High School	1,037,902
	Middle School	17,164
Housing Type	Doesn't Own	61,614
	Owned	1,269,341
	Renting	82,746
Occuputation	Coporate	586,105
	General Labor	827,596
Previous Contract type	Cash Loans	626,867
	Consumer Loan	625,360
	Revolving Loans	161,474
Previous Loan Status	Approved	886,099
	Canceled	259,441
	Refused	245,390
	Unused Offer	22,771
Contract Type	Cash Loans	1,307,115
	Revolving Loans	106,586

Table (8): Summary Statistics 3

8. References

1) Agrawal, K., & Maheshwari, Y. (2018, September 10). Efficacy of Industry Factors for Corporate Default Prediction. IIMB Management Review, 31(1), 71–77. https://doi.org/10.1016/j.iimb.2018.08.007

2) Akomas. (2018, June). Effects of Geographical Location on MFI Lending Behaviour in Developing Countries. Doctoral thesis, University of Huddersfield. http://eprints.hud.ac.uk/id/eprint/34683/

3) Chan, S., Haughwout, A., Hayashi, A., & Van Der Klaauw, W. (2016, March). Determinants of Mortgage Default and Consumer Credit Use: The Effects of Foreclosure Laws and Foreclosure Delays. Journal of Money, Credit and Banking, Vol. 48 (2/3), 393–413. https://www.jstor.org/stable/pdf/43862617?saml_data=eyJzYW1sVG9rZW4iOiI1ZDA0NTc2My03MjM5LTQ1Y2YtYjBmOS02Mzg4MDMwMjZjM2YiLCJpbnN0aXR1dGlvbklkcyI6WyIxODg0NWVhNS1jNmU0LTQzY2ItODdkMy03MzMwMmJlNjUwYjEiXX0

4) Dutta, G. (2020). Loan Defaulter. Retrieved December 12, 2024, from Kaggle.com website: https://www.kaggle.com/datasets/gauravduttakiit/loan-defaulter

5) Himberg, T., Wang, X. (2021, March 12). Loan Default Prediction with Machine Learning. https://www.doria.fi/bitstream/handle/10024/182846/himberg_tomi.pdf

6) Who Experiences Default? (2024, March). https://www.pewtrusts.org/en/research-and-analysis/data-visualizations/2024/who-experiences-default