Predictive Modeling for Loan Default Risk Management

BY: JUN MING LI, JAIRO ONATE, AND MAZEN ALHAFFAR

Table of contents

Introduction

Objective	The objective is to predict the loan default of customers in the banking industry
Risk Mitigation	Identifies high-risk borrowers to reduce financial losses
Improved Credit Assessment	Enhances loan approval decisions with data-driven insights
Cost Efficiency	Reduces operational costs by streamlining the credit evaluation process
Customer Relationships	Enables personalized solutions to support borrowers
Competitive Advantage	Provides a smarter approach to risk management

Variables 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)

Summary Statistics

1

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

2

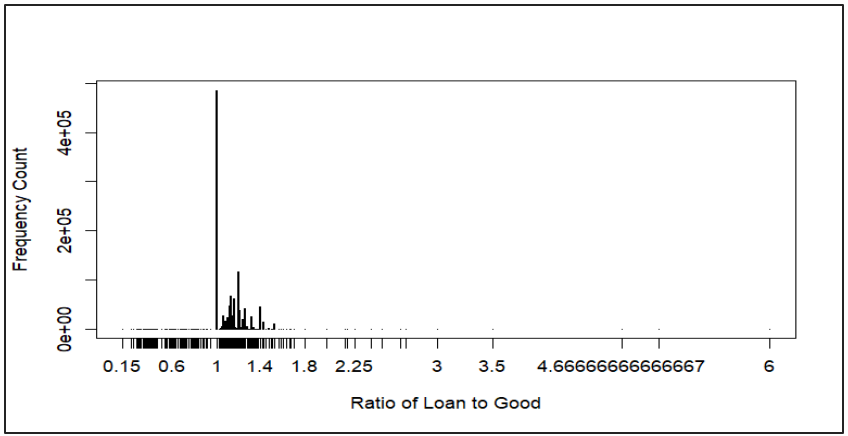
	AMT_INCOME_TOTAL	REGION_POPULATION_RELATIVE	DAYS_BIRTH
Min	25650.00	0.00	7489.00
Mean	173316.04	0.02	16321.05
Max	117000000.00	0.07	25201.00

3

	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

Summary Statistics

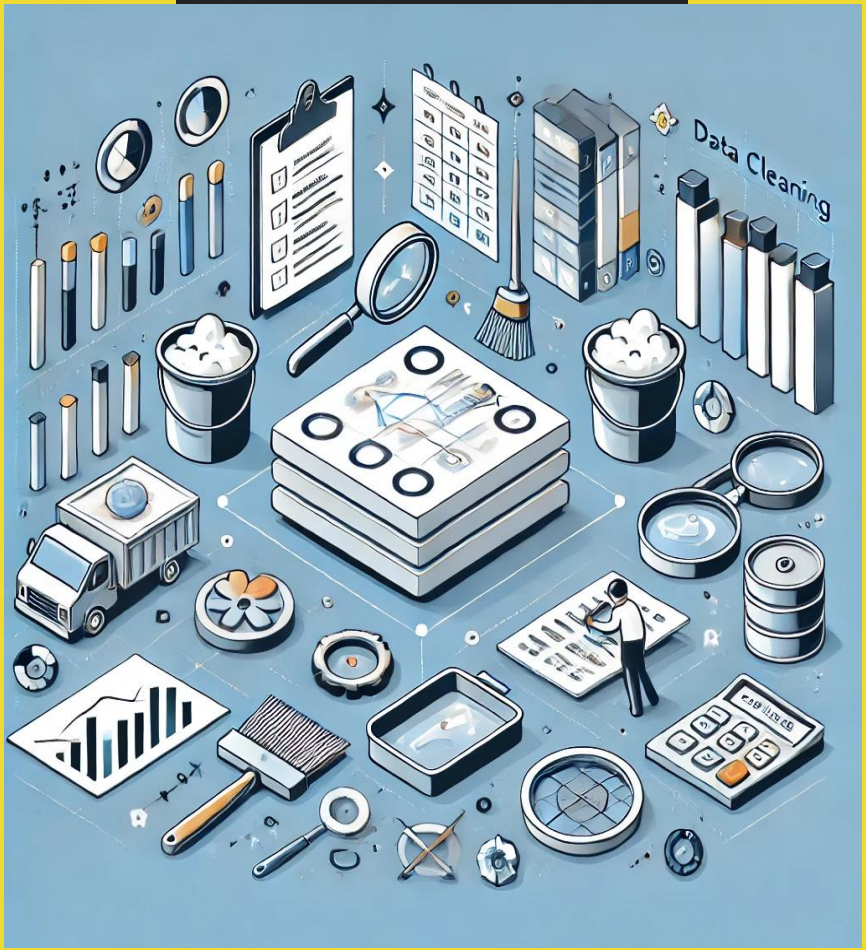
1) **Dataset Overview:** 1,413,701 rows and 16 variables, capturing customer demographics, financial details, and loan application data.
2) **Target Variable:** 91% of loans were repaid without issues (TARGET = 0), while 9% represent loan defaults (TARGET = 1).
3) **Income and Credit:** Mean annual income is $173,316, mean credit amount is $587,554, and mean goods price is $527,728.
4) **Age:** Mean client age is 44 years.
5) **Housing and Ownership:** 89.7% of clients own real estate, while 34% own a car. Majority of clients live in owned housing (90%), with smaller proportions renting or having no housing ownership.
6) **Previous Loans Status:** Most prior loans were approved (63%).
7) **Geographic Trends:** Clients predominantly reside in regions with moderate population densities (mean: 0.02).



The median value of 1.00 indicates that, on average, clients borrow an amount equivalent to the value of the goods they purchase

01

Data Cleaning & Preprocessing



Data Cleaning & Preprocessing

Numerical Data Processing:

- 8.39M missing fields across 307K records and 74 variables.
- Negative values were converted to positive.
- Missing values were replaced with the mean (excluding binary variables).

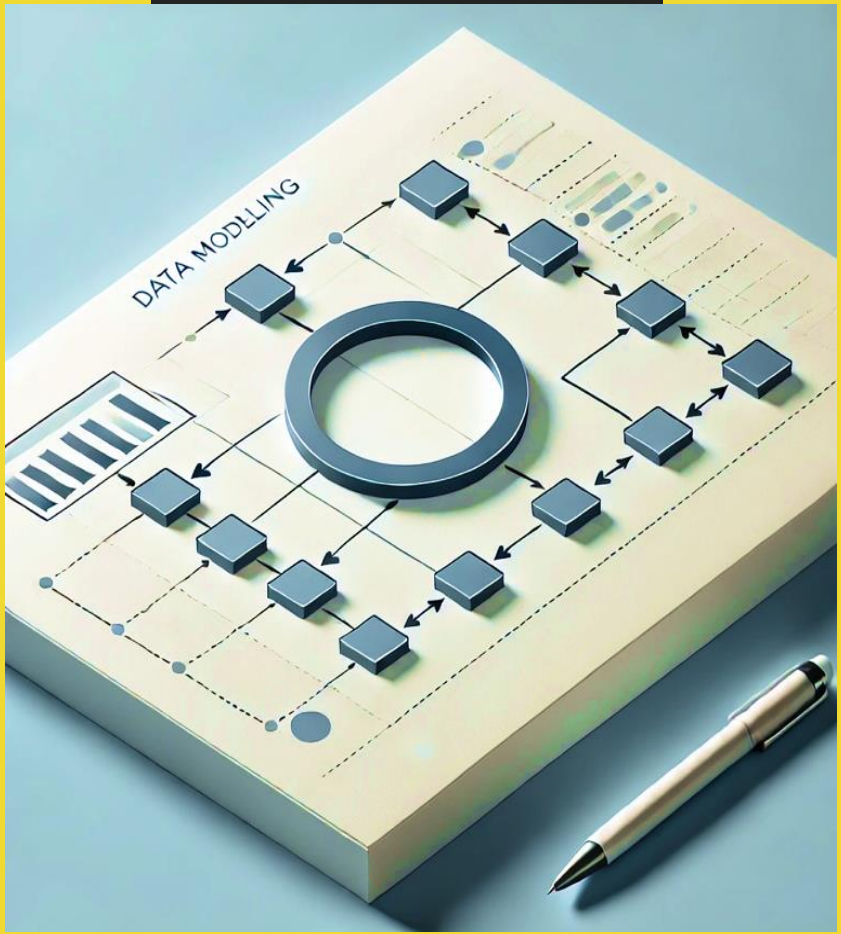
Previous Application Dataset:

- Initially contained 1.67M records, 37 variables, and 16.64% missing fields.
- Mean imputation reduced missing data by 86%, leaving 1.43M fields requiring further processing.

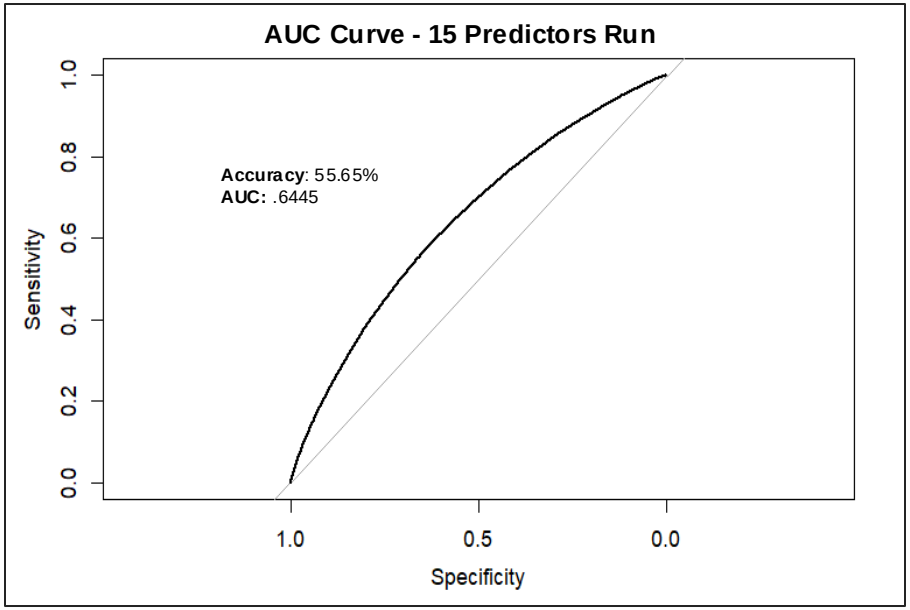
Categorical Data Processing:

- Combined and reduced classes.
- Used imputation to replace NAs, but preserve robustness of data set.

02
Data Modeling



Logistic Regression



AUC is at 0.6445, indicating the model performs better than a random classification

Logistic Regression			
	Initial Run, 45 predictors	2nd Run, 23 predictors	3nd 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

Random Forest

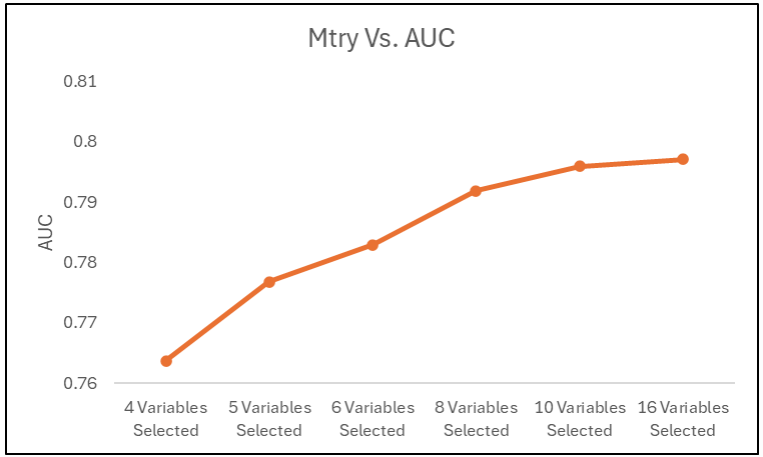
Random Forest						
	4 Variables Selected	5 Variables Selected	6 Variables Selected	8 Variables Selected	10 Variables Selected	16 Variables Selected
Accuracy	0.7671	0.7836	0.7877	0.7971	0.8004	0.7984
Sensitivity	0.7678	0.7851	0.7888	0.7982	0.8013	0.7987
Specificity	0.7596	0.7685	0.777	0.7856	0.7904	0.7954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

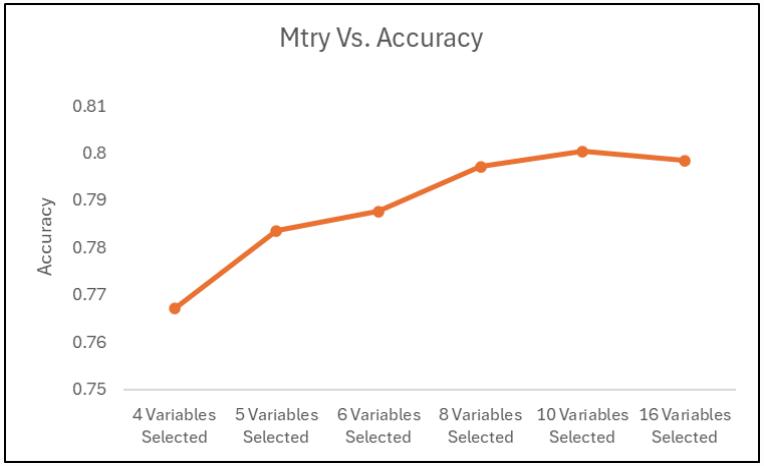
The AUC (Area Under the Curve) improves as the number of variables selected (mtry) increases from 4 to 16.

Accuracy also improves as the number of variables (mtry) increases from 4 to 16, following a similar trend to AUC.

03
Model Evaluation

Random Forest Performance





- Model performance improves consistently with more variables, but the improvement diminishes after 8 variables.
- The highest accuracy and AUC are achieved at 10-16 variables, but 8 variables offer a good balance between performance and efficiency.

Model Performance Comparison

Model	AUC	Accuracy	Sensitivity	Specificity
Baseline	**.6199**	**.2743**	**.214**	.9094
Logistic Regression (15 Predictors)	.6445	.5665	.5584	**.6509**
Random Forest (mtry = 8 run)	**.7919**	**.7971**	**.7982**	**.7856**

Compared to our baseline, our improved logistic and random forest models has significantly improved performance.

05
Key Insights

Key Insights

Variable Importance

"Days employed," "age," and "loan-to-goods ratio" were top predictors of loan default risk.

Model Selection

Most accurate is not always the “best” model.

Feature Contribution

- Days employed odds = decreases by 97%
- Age odds = Decreases by 76%
- Loan-to-goods ratio odds = increases by 120%

Conclusions & Recommendations

Technical

- Recommendations made adhering to the parsimony principle.
- While the Logistic Regression with 23 predictors has the best performance, the 15-predictor model is recommended for banks when using a logistic approach.
- Random Forest with 8 predictors chosen with 9.71% Accuracy, 79.82% Sensitivity, 78.56% Specificity and 79.19% AUC.

Business

- Simplifying the model enhances transparency and computing efficiency, making it ideal for real-time applications.
- The classification model balances interpretability and accuracy, making it valuable for commercial applications requiring clear decision-making.

Thank You!