Loan Default Prediction: Data-Driven Insights

Practical Data science
Home Equity Loans

Nassim Rafiefard

December 2024

Data-Driven Decision Making: Project Overview

1

Project Context

Understanding the background and goals of our data analysis project.

2

Data Exploration and Cleaning

Examining raw data, identifying patterns, and preparing it for analysis.

3

Modeling Results

Presenting insights gained from our analytical models and algorithms.

4

Summarize and Recommendations

Proposing actionable steps based on our findings to drive business value.

Project Context and Objectives

Enhance Financial Stability

Develop a reliable model to predict loan default likelihood, ensuring profitability for banks and lending institutions.

Improve Fairness

Enhance transparency in loan approval processes, aligning with regulatory and ethical standards.

Overcome Traditional Challenges

Address issues of labor-intensive manual assessments prone to errors, biases, and inconsistent judgments.

Streamline Processes

Improve accuracy and consistency in loan approvals, building customer trust through data-driven decisions.

Data Exploration

Dataset Characteristics

Our dataset comprises 5,960 customers with 13 features (11 numerical, 2 categorical). Key features include loan status (BAD), approved loan amount (LOAN), outstanding mortgage (MORTDUE), property value (VALUE), loan purpose (REASON), and job type (JOB).

Univariate Analysis

Numeric Data

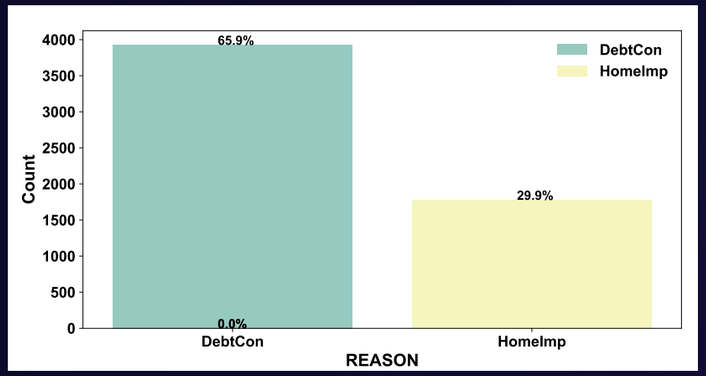
All numerical features show right skewness with outliers present.

Categoric Data

Loan Purpose:

Debt Consolidation (DebtCon): 69.8% of clients.

Home Improvement (HomeImp): Remaining clients

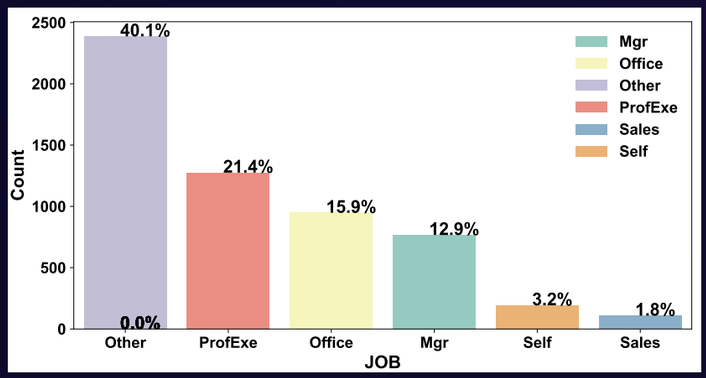


Job Category:

The largest group (44.7%) belongs to the "Other" job category.

Remaining categories include: ProfExe (Professional/Executive)

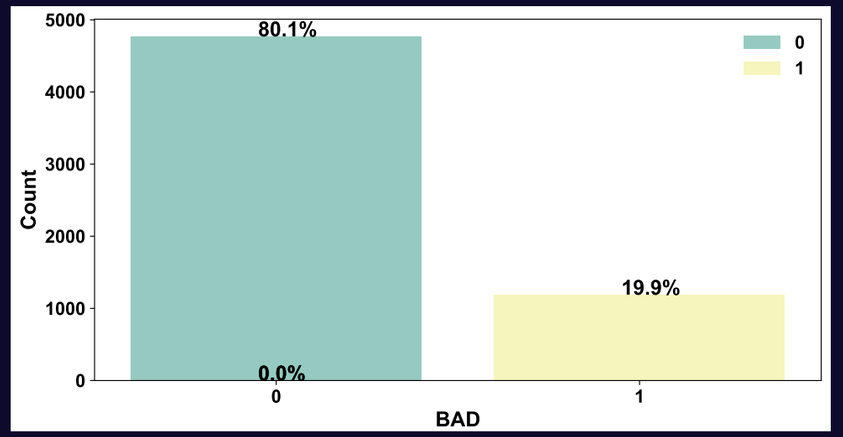
Office, Mgr (Manager), Self, and Sales



Data Exploration

Target Variable Analysis

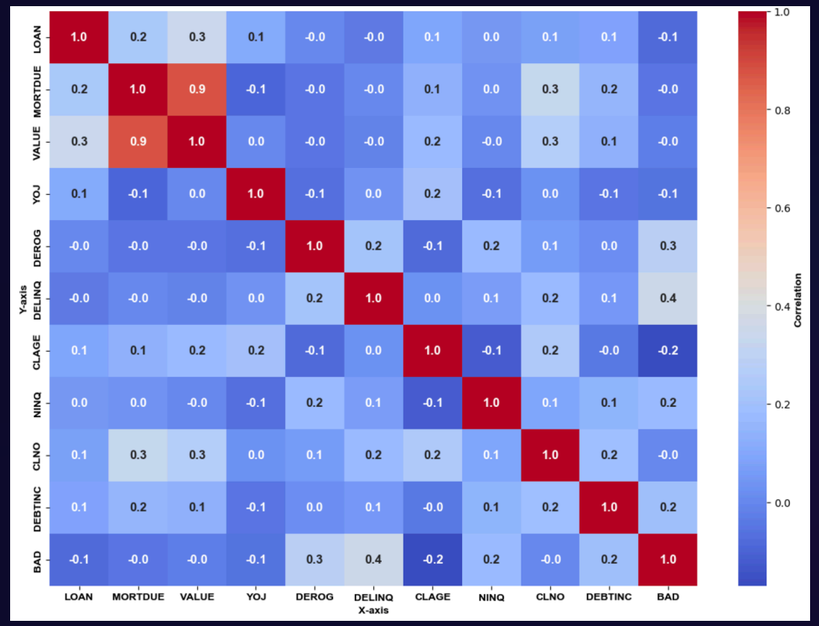
- Loan Defaulters (BAD = 1): Approximately 20% of clients in the dataset.
- Non-Defaulters (BAD = 0): 80% of clients have successfully repaid their loans.



Bivariate Analysis

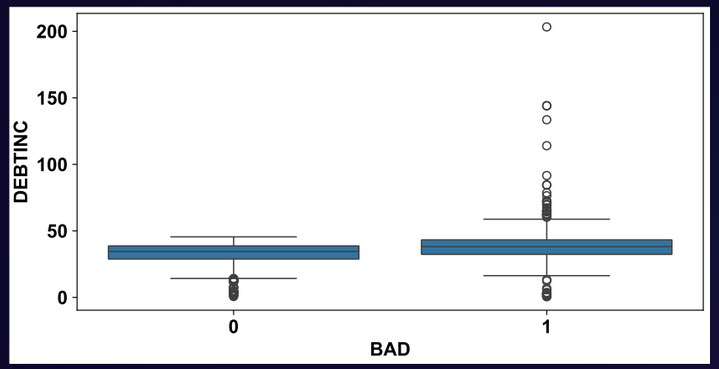
Correlation matrix:

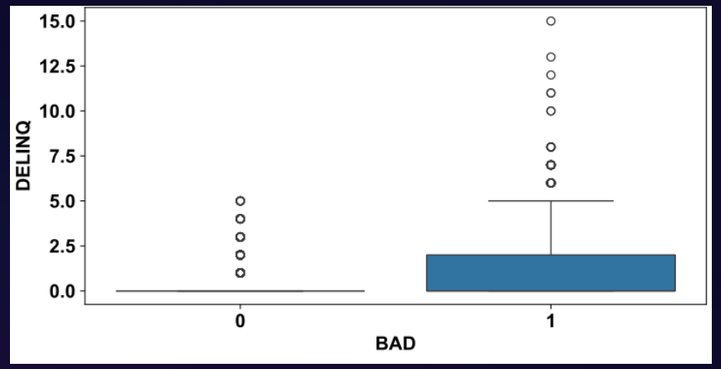
- Feature-to-Feature:
  - Strong correlation observed between VALUE and MORTDUE.
  - Moderate correlation exists between these two features and CLNO.
  - LOAN shows moderate correlation with VALUE and MORTDUE.
- Feature-to-Target:
  - DELINQ and DEROG show a noticeable correlation with the target variable (BAD).

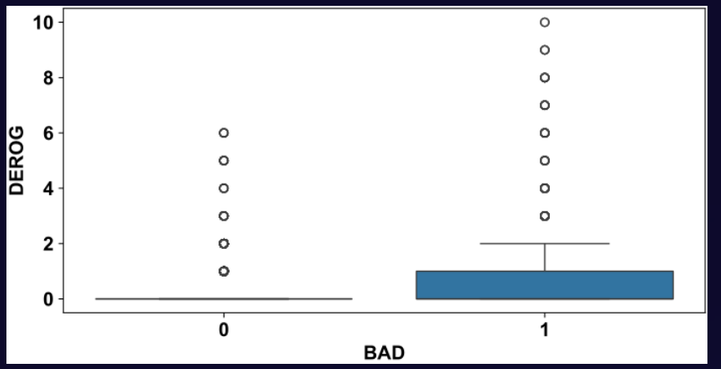


Histogram and Boxplot Analysis For Numerical Features

- Consistent with the correlation findings:
  Features DEBTINC, DEROG, and DELINQ demonstrate a significant association with the likelihood of loan default or severe delinquency.

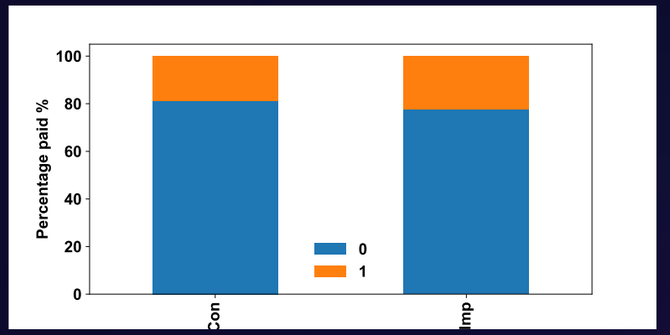




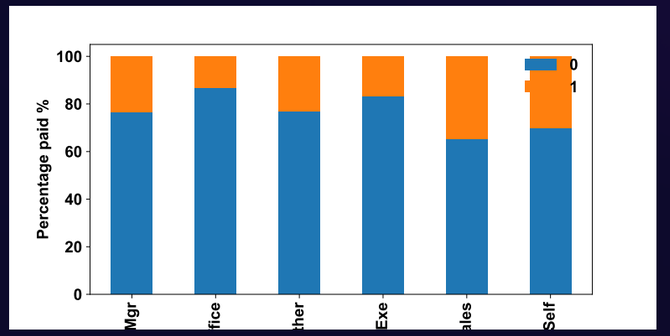


Categorical Features Analysis

- Loan Purpose:
  Clients requesting loans for home repairs and those requesting loans for debt consolidation show similar likelihoods of defaulting or experiencing severe delinquency.



- Job Categories and Default Risk:
  Occupations associated with a higher likelihood of default or severe delinquency (ranked highest to lowest): Sales, Self-employed, Manager, Other and Professional/Executive



Data Cleaning and Preparation

1 Outlier Treatment

Applied Z-score and IQR methods to detecting outliers in numerical features.

	LOAN	MORTDUE	VALUE	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
Zscore	95	0	0	0	0	0	0	0	0	0
IQR	256	234	320	91	725	1201	47	177	219	94
Handle type	null	null	null	null	none	none	mean	Upper Whisker	upper Whisker	null

2 Missing Value Treatment

Features with >6% missing values imputed using range-based methods. MORTDUE removed due to high correlation with LOAN.

Features with <6% Missing Values: Imputed using mean, median, or mode, depending on the feature type (numerical or categorical).

Features	LOAN	MORTDUE	VALUE	Reason	JOB	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
%	4.3	12.6	7.2	4.2	4.7	10.2	11.9	9.7	5.2	10.6	4.4	22.8
Handle type	mean	Deleted	72000-120 000	mode	mode	5.2-11	0.12 - 1.2	0- 0.88	mean	0.45-1.4	median	32-39

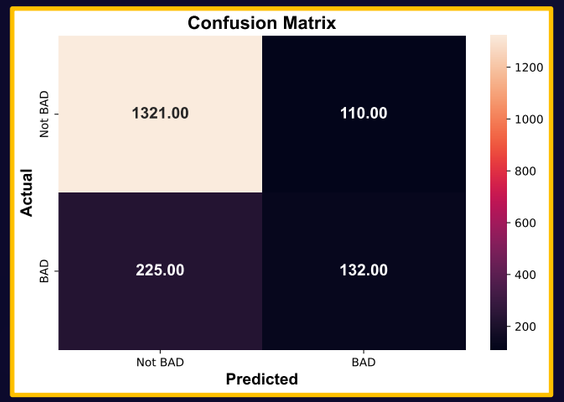
3 Feature Engineering

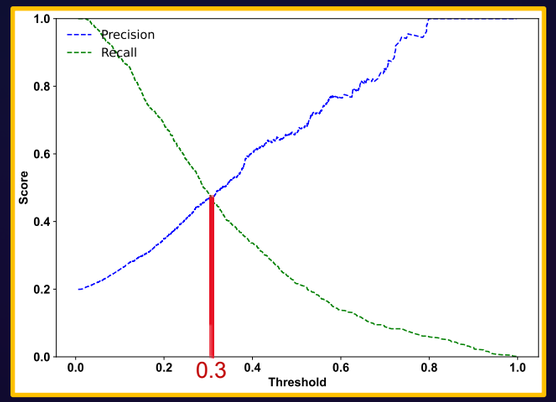
Categorical features encoded and numerical features normalized to prepare data for modeling.

Modeling Results:

Logistic Regression

Threshold=0.3

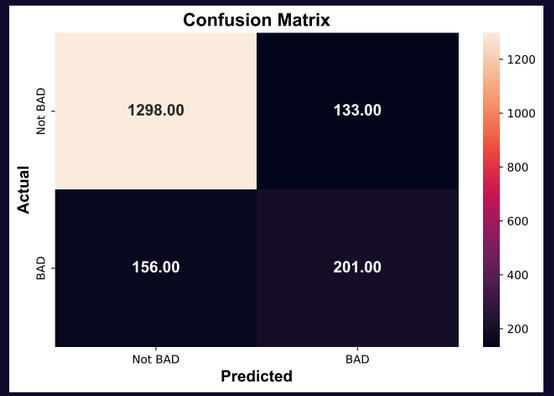


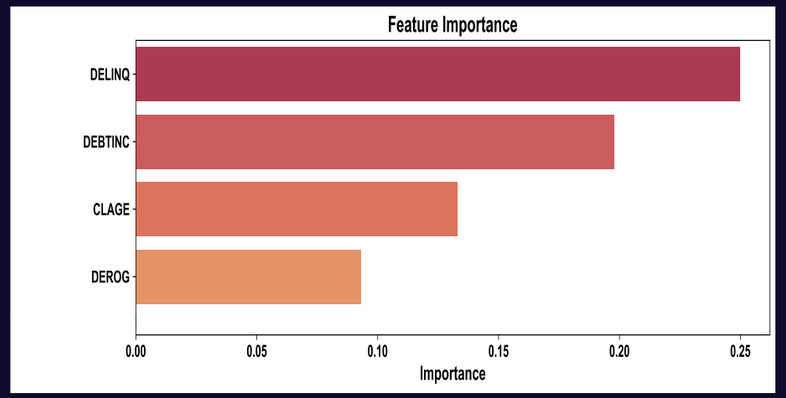


	odds
DELINQ	2.250886
DEBTINC	1.693103
DEROG	1.500510
NINQ	1.275832

	Precision	Recall	Accuracy
Regression	0.78	0.62	0.83

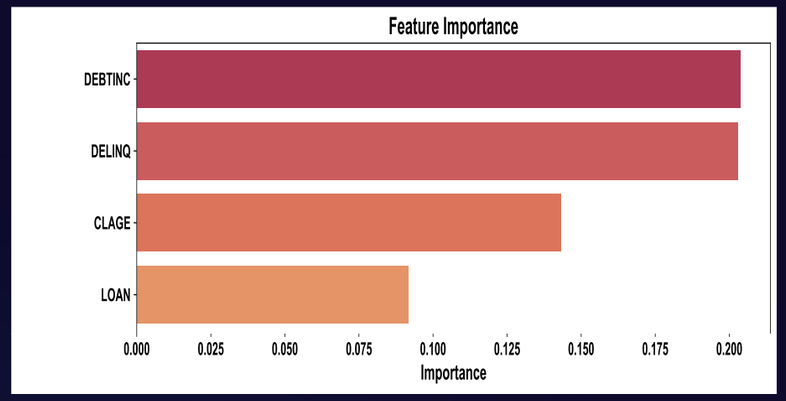
Tuned Decision Tree





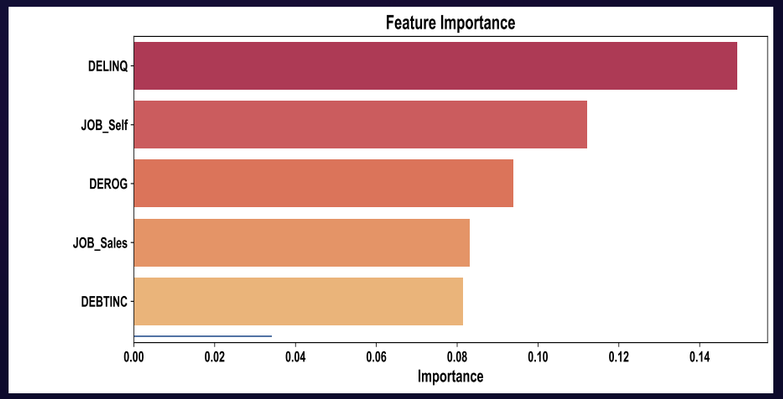
	Precision	Recall	Accuracy
Tuned Decision Tree classifier	0.75	0.74	0.84

Tuned Random Forest



	Precision	Recall	Accuracy
Tuned Decision Tree classifier	0.82	0.77	0.88

XGBoost



	Precision	Recall	Accuracy
XGBoost	0.90	0. 80	0.90

	Logistic Regression	Decision Tree	Tuned Decision Tree	Random Forest	Tuned Random Forest	XGBoost
Precision	0.74	0.79	0.75	0.92	0.82	0.90
Recall	0.69	0.76	0.73	0.76	0.77	0. 80
Accuracy	0.82	0.87	0.83	0.90	0.88	0.90

Summarize:

Key Question:

The primary issue is identifying factors contributing to loan defaults and using these insights to predict which applicants are at the highest risk

Models:

The solution integrates predictive machine learning models to achieve the models which are tuned to prioritize recall beside having high accuracy, ensuring a focus on identifying defaulters while minimizing false negatives.

variety of classification models were tested, including:

- Logistic Regression
- Decision Trees
- Random Forests
- XGBoost Classifier:

Final Model Recommendation:

-The XGBoost Classifier is the recommended final model because of its excellent performance, achieving: 79% recall for identifying defaulters (Class 1), 89% precision, and 90% overall accuracy on the test data. This model strikes an ideal balance between recall and precision, meeting the key evaluation criteria effectively.

-The Tuned Random Forest classifier achieves: 77% recall for identifying defaulters (Class 1), 82% precision, and 88% overall accuracy on the test data. It ranks second among the models. Both this model and the top-ranking model maintain high accuracy and recall without sacrificing much in precision.

Summarize and Recommendations:

Feature Importance:

All features which are financially attributed are important in the model: DEBTING, DELINQ, DEROG, CLAG, …

Key Recommendations for the Bank:

1.Focus on Critical Features

2.Model Selection and Use:
•Pilot Testing: Test the model alongside the current manual process and validate its performance before fully transitioning.
•Handle Missing Data Effectively:
•Leverage Advanced Machine Learning Methods: Explore other algorithms, such as Boosting Techniques, KNN, Support Vector Machines (SVM), or Neural Networks, for further improvement.
•Perform feature engineering to enhance data quality, such as creating new features or dropping irrelevant columns.

3.Expected Benefits:
•Improved Risk Management: Identify high-risk applicants early to reduce financial losses.
•Efficient Loan Processing: Automate decision-making to reduce manual effort and accelerate approvals.
•Enhanced Customer Satisfaction: Faster and fairer loan evaluations improve the applicant experience.

4.Key Risks and Challenges:
•Ethical and Legal Compliance: Ensure the model adheres to all regulations and avoids unintended bias against specific applicant groups.