Documentation on AI driven loan recovery model

By

Urielle Silvana Megnokam Fotso

Table of Contents

Executive Summary

0.1. Executive Introduction

Unrecovered high-risk loans, resource misallocation, and ineffective recovery strategies drain organizational profits by prioritizing low-value efforts, such as costly legal notices ($500) over efficient calls ($10), resulting in less loans recovered. This inefficiency wastes resources and misses opportunities to recover high-potential loans, undermining financial performance. Our AI-driven loan recovery project addresses these challenges by leveraging predictive modeling to identify recoverable loans and optimize recovery strategies, ensuring resources are allocated effectively to maximize returns.

0.2. Executive Objective

The objective is to improve loan recovery efficiency by increasing recovery rates and reducing costs through a data-driven approach. The project uses a two-step predictive model to:

- Identify high-risk loans likely to be Written Off versus those that can be Recovered (Step 1).
- Among Recovered loans, distinguish between Fully Recovered and Partially Recovered loans (Step 2). Success is measured by high precision to minimize costly false positives, such as overestimating full recovery, which could lead to inappropriate resource allocation.

0.3. Executive Model Description

The project analyzes a dataset of 500 loan records, including borrower demographics, loan characteristics, and repayment behavior, with no missing values. A two-step logistic regression model, optimized with forward feature selection, predicts:

- **Step 1**: Written Off (50 loans) vs. Recovered (450 loans), achieving a ROC AUC of 0.76 and precision of 0.37/0.95.
- **Step 2**: Among Recovered loans, Partially Recovered (296 loans) vs. Fully Recovered (154 loans), achieving a ROC AUC of 0.81 and precision of 0.81/0.85. Key predictors include Payment History On-Time (3.2x higher odds of recovery, 98% confidence, p-value = 0.021), Collection Attempts (42.8% lower odds of full recovery per attempt), and

Collection Method Settlement Offer (2.1x higher odds of full recovery). Data preparation involved encoding categorical variables (e.g., Loan Type, Employment Type), addressing skewness (e.g., Monthly EMI, collateral value, number of missed payments and collection attempts), and balancing classes with SMOTE. Sensitive features like Gender were excluded to ensure FCAC/PIPEDA compliance, maintaining ethical and transparent predictions.

0.4. Executive Recommendations

To address unrecovered high-risk loans, resource misallocation, and ineffective recovery strategies, we recommend operationalizing the two-step model's predictive insights to enable a proactive, data-driven, and segmented recovery approach:

- **Loyalty-Based Retention Programs**: Offer interest rate incentives or fee waivers for loans with strong Payment History (3.2x higher recovery odds, p-value = 0.021) to maintain high repayment rates while minimizing collection costs (e.g., avoiding $500 legal notices).
- **Settlement-Based Recovery Strategies**: Prioritize settlement offers (2.1x higher odds of full recovery) for delinquent accounts with moderate recovery potential (50–70% probability), increasing resolution rates and reducing prolonged collection costs.
- **Stricter Credit Assessments**: Introduce enhanced credit evaluations and collateral-backed lending for new business loans, as Loan Type Business showed lower recovery odds (OR = 0.118), to mitigate future write-offs.
- **Early Settlement Negotiations**: Apply early settlement negotiations for existing high-risk accounts (e.g., high missed payments, OR = 1.940) to reduce write-offs, noting that this counterintuitive finding requires further analysis. This approach shifts from reactive, one-size-fits-all tactics to a targeted strategy, optimizing resource allocation (e.g., favoring $10 calls or $50–100 settlements) and enhancing recovery rates while maintaining customer trust through FCAC/PIPEDA-compliant operations.

- Finally Deploy the model to identify high-risk loans early and optimize recovery strategies, reducing losses from unrecovered loans and improving resource allocation with an implementation of 3 months pilot

INTRODUCTION

1.0 Background

Loan recovery is a critical component of financial risk management, particularly for institutions managing diverse portfolios of personal, business, and secured loans. Inconsistent recovery efforts, resource misallocation, and delayed identification of high-risk accounts can lead to increased write-offs and reduced profitability. With regulatory frameworks such as FCAC and PIPEDA in Canada, recovery processes must be both effective and compliant.
This project leverages AI-driven predictive modeling to address these challenges, integrating machine learning techniques with historical loan performance data to improve decision-making in recovery strategies.

2.0 Problem Statement

The organization currently faces challenges in:

- Accurately predicting the recovery outcome of delinquent loans (Fully Recovered, Partially Recovered, or Written Off).
- Prioritizing accounts for recovery efforts, leading to inefficient resource allocation.
- Lacking a data-driven, standardized approach for tailoring recovery strategies to borrower profiles.

Impact: These gaps result in missed recovery opportunities, higher operational costs, and increased financial losses due to write-offs.

3.0 Objectives & Measurement

Primary Objective:
Develop and deploy an AI-driven predictive model capable of:

- Forecasting loan recovery status with high accuracy.
- Recommending cost-effective and compliance-ready recovery strategies.

Key Performance Indicators (KPIs):

- Model Accuracy & ROC AUC: Evaluate predictive performance.
- Recovery Rate Increase: Target a measurable improvement (e.g., +25%) in overall recovery rates.
- Cost Reduction: Aim for operational cost savings (e.g., -30%) in recovery processes.
- Regulatory Compliance: Ensure outputs meet FCAC and PIPEDA standards.

4.0 Assumptions and Limitations

Assumptions:

- Historical loan data accurately reflects borrower behavior and is representative of current market conditions.
- Recovery processes and strategies remain consistent during the project period.
- Regulatory requirements remain unchanged within the implementation timeframe.

Limitations:

- Data Size: Some loan categories have smaller sample sizes, potentially affecting model generalization.
- Class Imbalance: Certain recovery categories (e.g., Fully Recovered) may dominate, requiring oversampling or resampling methods.
- External Factors: Macroeconomic changes, policy shifts, or unforeseen events (e.g., pandemics) may impact loan recovery patterns beyond model scope.
- Interpretability vs. Complexity: While advanced models (e.g., Random Forest, Logistic Regression with stepwise selection) can improve accuracy, interpretability must be maintained for stakeholder trust and regulatory alignment

Data Sources

5.0. Data Set Introduction

This dataset contains information about 500 loan borrowers, including demographics, loan characteristics, repayment behavior, and recovery status. The objective is to analyze the factors affecting loan recovery, and build models to predict whether a loan is:

- Written Off vs. Recovered (Step 1), and
- Fully Recovered vs. Partially Recovered (Step 2).

Target variables (Is_Recovered for Step 1, Recovery_Level for Step 2) are derived from Recovery_Status to support this two-step classification approach.

6.0. Exclusions

- No records were excluded due to missing values, as the dataset contains no nulls.
- Loan_ID and Borrower_ID were excluded as they lack predictive utility.
- For Step 2 modeling (Fully vs. Partially Recovered), the dataset was filtered to include only Recovered loans (450 records), excluding Written Off loans (50 records).

7.0. Data Dictionary

Variables	Description	Data type
Gender	Gender of the borrower Male (0) or Female (1)	Nominal
Employment type	Type of employment of the borrower 1-Salaried 2-Self-employed 3-Business Owner	Nominal
Monthly Income	Salary obtained per month	Float
Num_Dependents	Number of children below 17 years old financially depending	Integer

Loan_ID	unique identifier for each loan record	Nominal
Loan_Amount	principal amount of the loan (in Canadian Dollars, CAD) borrowed by the individual.	Integer
Interest_rate	annual interest rate (expressed as a percentage) applied to the loan	Float
Loan_type	type or category of the loan taken by the borrower 1-Auto 2-Business 3-Home 4-Personnal	Nominal
Collateral_value	monetary value (in Canadian Dollars, CAD) of the asset(s) pledged as security for the loan	Float
Monthly_EMI	The fixed monthly payment (in Canadian Dollars, CAD) that the borrower is required to make to repay the loan, including both principal and interest components	Float
Payment_History	Track record of making loan payments 1-on time 2-Delayed 3-Missed	Nominal

Num_Missed_payments	Number of times payments were missed	Integer
Days_past_due	Number of days a borrower's loan payment is overdue at the time of data collection	Integer
Collection_attempts	The number of times the bank has attempted to contact the borrower or take action to recover the outstanding loan amount	Nominal
Collection_method	The type of method used by the lender to recover the outstanding loan amount from a delinquent borrower. 1-Calls 2-Debt collectors 3-Legal notice 4-Settlement Offer	Nominal
Legal_Action	Indicating whether the banker has initiated formal legal action to recover the outstanding loan amount from a delinquent borrower. 0-No 1-Yes	Binary
Recovery_status	The outcome of the loan recovery process Written_off (0), fully recovered (1), partially recovered (2)	Nominal

Extra columns which are the target variable for step 1 and 2 obtained from the variable recovery status

Is recovered	Indicate whether a loan has been Written off (0) vs Recovered (1)	Integer
Recovery level	Indicates the extent to which a recovered loan was repaid partially recovered (0) vs fully recovered (1)	Integer

Data exploration

8.0. Data Exploration Techniques

The goal of data exploration is to understand the dataset's structure, identify relationships between variables, assess data quality, and inform feature engineering and modeling decisions. The following techniques were applied to analyze the loan-recovery dataset.

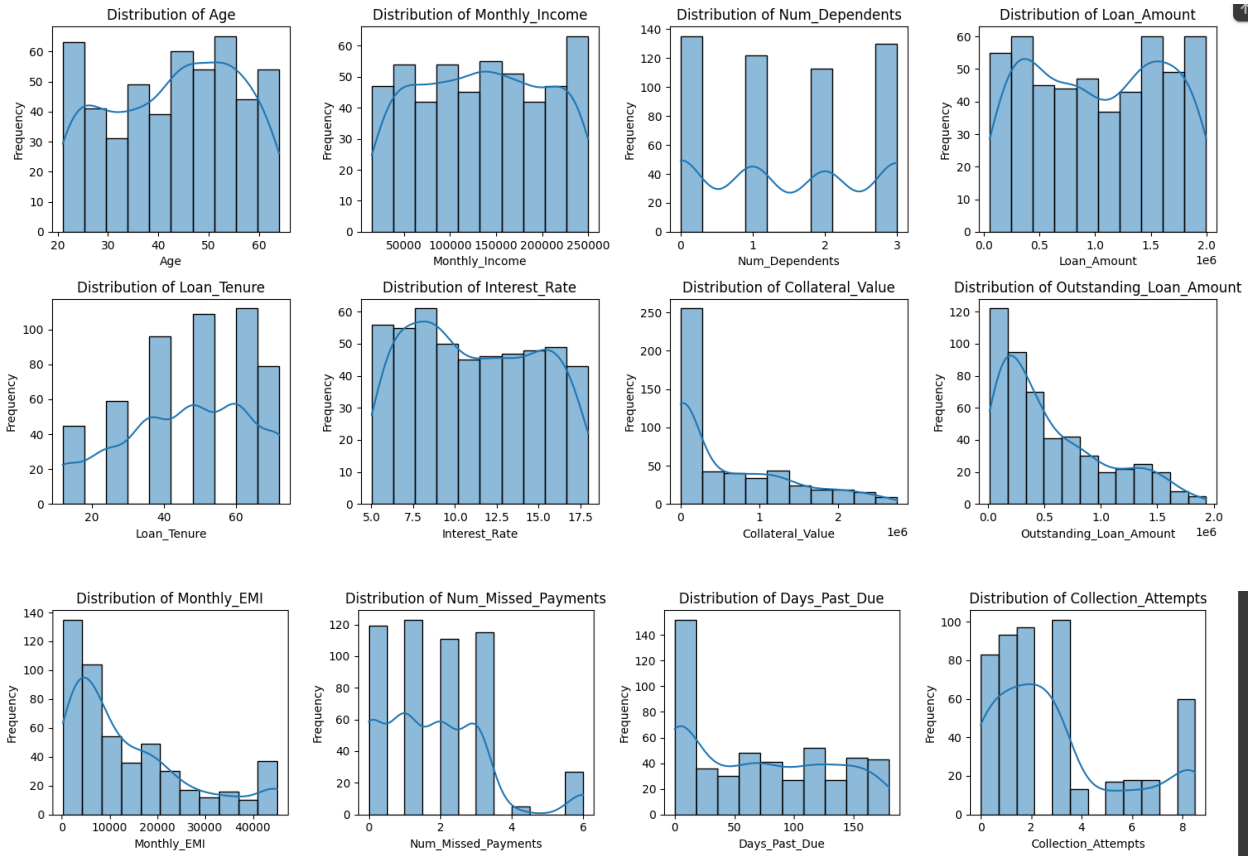
8.1. Data Exploration Techniques

1. Descriptive statistics; This was conducted to examine the distribution, central tendency (mean, median), and variability (standard deviation, range) of the numerical variables. This step provided a foundational understanding of the data by highlighting how each variable behaves, identifying potential skewness, and revealing irregularities such as extreme values or inconsistencies

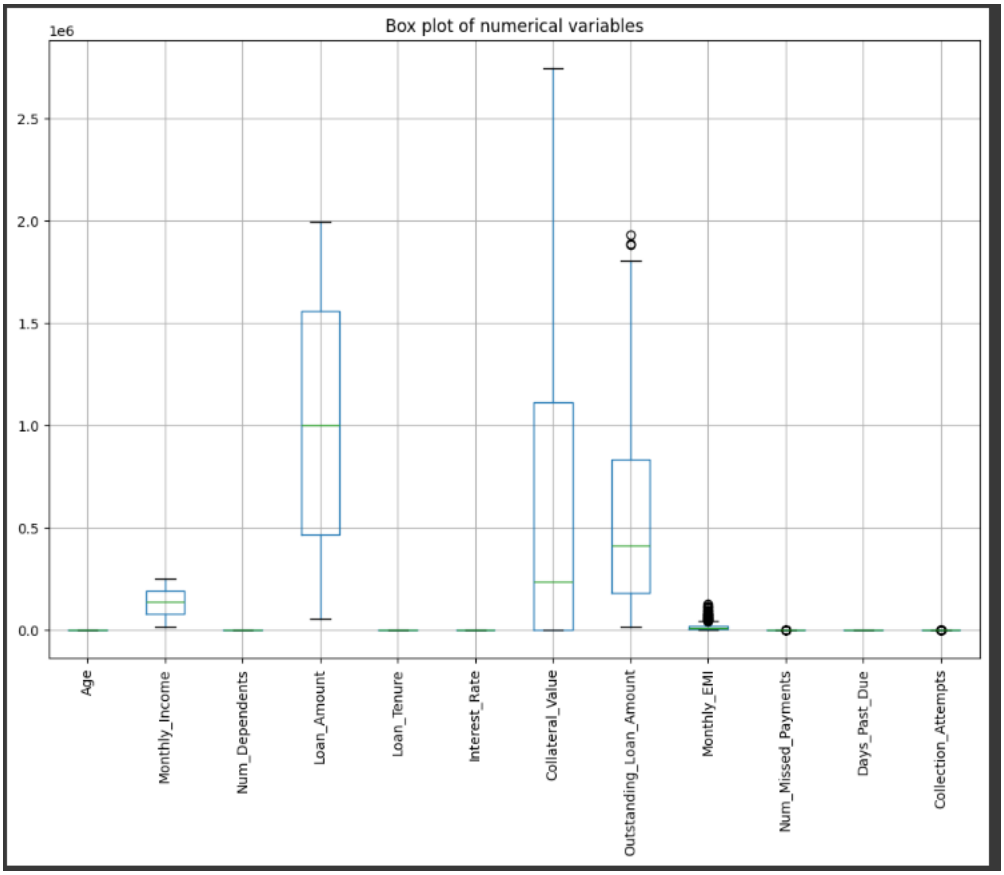
	Age	Monthly_Income	Num_Dependents	Loan_Amount	Loan_Tenure	Interest_Rate	Collateral_Value	Outstanding_Loan_Amount	Monthly_EMI	Num_Missed_Payments	Days_Past_Due	Collection_Attempts
count	500.000000	500.000000	500.000000	5.000000e+02	500.00000	500.000000	5.000000e+02	5.000000e+02	500.000000	500.000000	500.000000	500.000000
mean	43.116000	134829.920000	1.476000	1.024907e+06	46.10400	11.192820	6.032240e+05	5.627260e+05	15861.536020	1.912000	70.678000	3.000000
std	12.733217	68969.356746	1.145447	5.907556e+05	18.23706	3.775209	7.457131e+05	4.723581e+05	18709.231315	2.110252	60.211038	2.807805
min	21.000000	15207.000000	0.000000	5.413800e+04	12.00000	5.020000	0.000000e+00	1.571283e+04	261.880000	0.000000	0.000000	0.000000
25%	32.000000	76343.250000	0.000000	4.629848e+05	36.00000	7.907500	0.000000e+00	1.822072e+05	4039.097500	1.000000	4.000000	1.000000
50%	44.000000	134929.500000	1.000000	9.971240e+05	48.00000	10.915000	2.327684e+05	4.133240e+05	9330.170000	2.000000	66.500000	2.000000
75%	53.000000	193086.250000	3.000000	1.557952e+06	60.00000	14.577500	1.111106e+06	8.324787e+05	20439.485000	3.000000	122.250000	4.000000
max	64.000000	249746.000000	3.000000	1.995325e+06	72.00000	17.970000	2.744395e+06	1.932396e+06	127849.230000	12.000000	180.000000	10.000000

2. Visual Exploration:

- Objective: Visualize feature distributions
- Method: Use boxplots and histograms (numerical variable) and bar plots (categorical variable) via seaborn and matplotlib.



The histogram of numerical variable



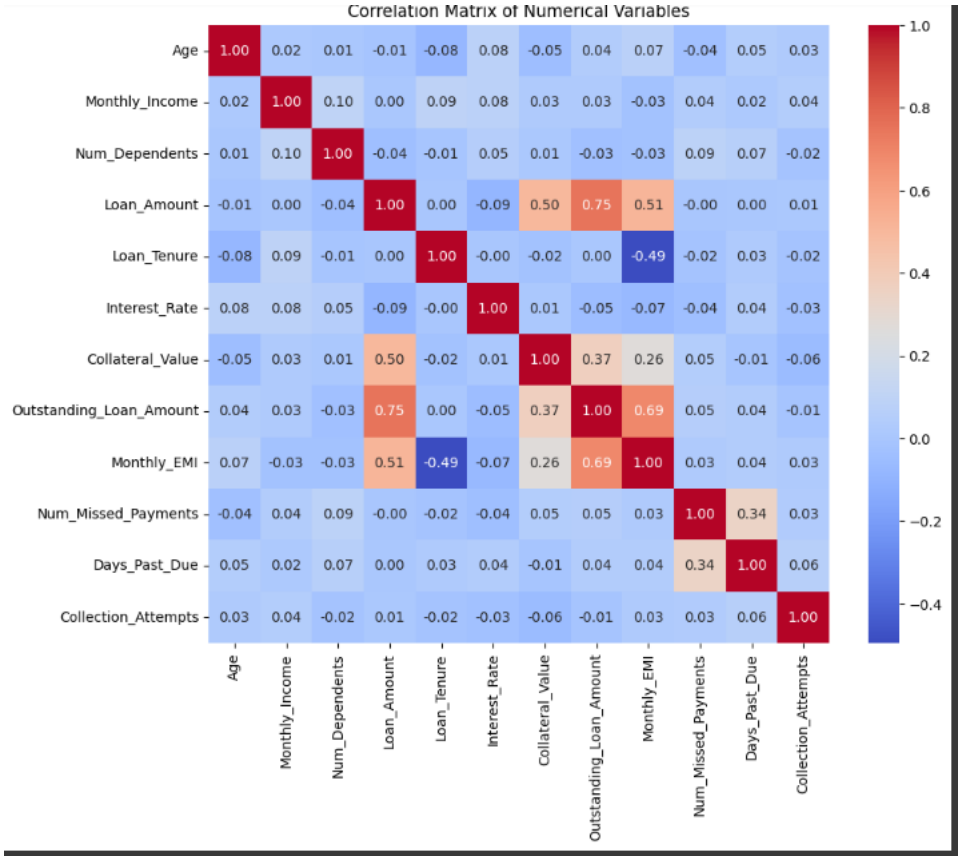
The boxplot of numerical variable

- Analysis; box plots revealed outliers in variables such as Monthly EMI, Number of Missed Payments, Collateral Value, and Collection Attempts. Further preprocessing was involved examining the skewness of these variables to determine the most appropriate treatment — whether to handle the outliers directly or apply transformations to correct skewed distributions.

3. Feature-Target Relationship (correlation analysis):

   - Objective: Identify predictors strongly associated with *Recovery Status* to inform feature selection. For the correlation matrix, a Pearson correlation coefficient greater than or equal to 0.75 was considered indicative of a strong relationship. For Chi-square and ANOVA tests, a p-value less than or equal to 0.05 was used to determine statistical significance. This also helps to avoid multi collinearity.

- Method: A correlation matrix was generated for numerical variables. The results showed that Loan Amount and Outstanding Loan Amount were highly correlated (0.75). An ANOVA test was conducted with Recovery Status to guide variable selection during feature engineering. Interestingly, Loan Amount showed a stronger statistical association with the target. Chi-square tests were conducted to assess the association between categorical variables and Recovery Status. Among these, Legal Action Taken and Collection Method showed strong associations with each other (p-value 0.002). In particular, Legal Action Taken and recovery status had a highly significant p-value (0.000), indicating a strong relationship.



The correlation matrix of numerical variables

```
ANOVA results for numerical variables and Recovery_Status:
Age: F-value = 0.6735, P-value = 0.5104
Monthly_Income: F-value = 0.2657, P-value = 0.7668
Num_Dependents: F-value = 1.4455, P-value = 0.2366
Loan_Amount: F-value = 2.1390, P-value = 0.1189
Loan_Tenure: F-value = 0.1629, P-value = 0.8497
Interest_Rate: F-value = 0.0123, P-value = 0.9878
Collateral_Value: F-value = 2.5418, P-value = 0.0798
Outstanding_Loan_Amount: F-value = 1.6824, P-value = 0.1870
Monthly_EMI: F-value = 0.8309, P-value = 0.4362
Num_Missed_Payments: F-value = 1.0828, P-value = 0.3394
Days_Past_Due: F-value = 0.9984, P-value = 0.3692
Collection_Attempts: F-value = 136.4253, P-value = 0.0000
```

ANOVA results for numerical variables and recovery status

```
Association between categorical variables (excluding Recovery_Status) using Chi-Squared test:
Association between Gender and Employment_Type: Chi-Squared = 0.8896, P-value = 0.6409
Association between Gender and Loan_Type: Chi-Squared = 4.2559, P-value = 0.2351
Association between Gender and Payment_History: Chi-Squared = 2.1310, P-value = 0.3446
Association between Gender and Collection_Method: Chi-Squared = 1.6640, P-value = 0.6450
Association between Gender and Legal_Action_Taken: Chi-Squared = 0.4600, P-value = 0.4976
Association between Employment_Type and Loan_Type: Chi-Squared = 7.1028, P-value = 0.3114
Association between Employment_Type and Payment_History: Chi-Squared = 2.1310, P-value = 0.7117
Association between Employment_Type and Collection_Method: Chi-Squared = 10.6147, P-value = 0.1010
Association between Employment_Type and Legal_Action_Taken: Chi-Squared = 1.1012, P-value = 0.5766
Association between Loan_Type and Payment_History: Chi-Squared = 2.3409, P-value = 0.8858
Association between Loan_Type and Collection_Method: Chi-Squared = 10.6204, P-value = 0.3026
Association between Loan_Type and Legal_Action_Taken: Chi-Squared = 1.5860, P-value = 0.6626
Association between Payment_History and Collection_Method: Chi-Squared = 8.8655, P-value = 0.1813
Association between Payment_History and Legal_Action_Taken: Chi-Squared = 0.9867, P-value = 0.6106
Association between Collection_Method and Legal_Action_Taken: Chi-Squared = 14.8000, P-value = 0.0020
```

Chi square test for categorical variables

```
Chi-Squared test results for categorical variables and Recovery_Status:
Gender: Chi-Squared = 1.7605, P-value = 0.4147
Employment_Type: Chi-Squared = 1.9137, P-value = 0.7516
Loan_Type: Chi-Squared = 6.3189, P-value = 0.3884
Payment_History: Chi-Squared = 1.1955, P-value = 0.8788
Collection_Method: Chi-Squared = 2.2663, P-value = 0.8937
Legal_Action_Taken: Chi-Squared = 226.8908, P-value = 0.0000
```

Chi square test for categorical variables and recovery status

- Conclusion; Loan amount showed a stronger association to the target. However, from a business perspective, Outstanding Loan Amount was deemed more important, as it reflects current financial exposure — recovery efforts are based on what is still owed, not what was originally borrowed. Loan Amount and Collection Method will be dropped to support more effective feature engineering.

4. Skewness Analysis:

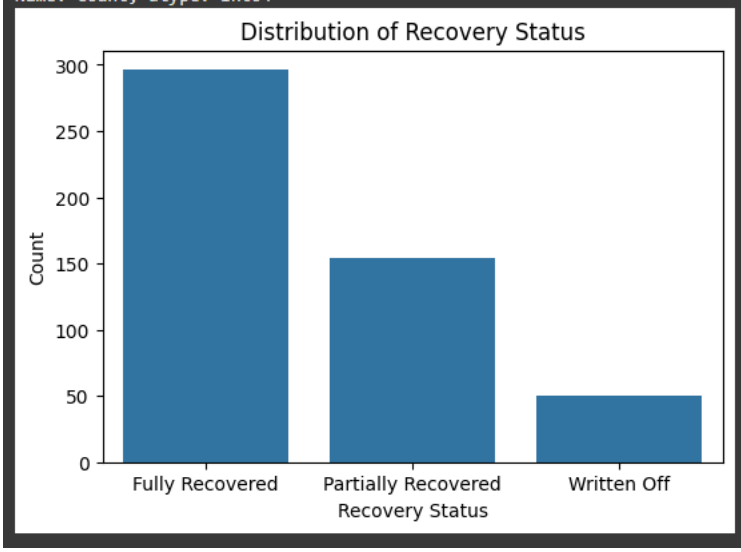
- Objective: Evaluate numerical features for skewness to inform outlier handling and transformations.
- Findings: Monthly_EMI, Num_Missed_Payments, Collateral_Value, and Collection_Attempts showed significant skewness, necessitating transformations as shown below

Skewness of numerical variables:

	0
Age	-0.162218
Monthly_Income	-0.002889
Num_Dependents	0.039021
Loan_Amount	0.001857
Loan_Tenure	-0.281297
Interest_Rate	0.111805
Collateral_Value	1.053850
Outstanding_Loan_Amount	0.930591
Monthly_EMI	2.625062
Num_Missed_Payments	2.576384
Days_Past_Due	0.258856
Collection_Attempts	1.114189

5. Target Variable Distribution:
    - Objective: Assess the distribution of Recovery status to confirm class imbalance for the two-step model.
    - Findings
        - Recovery status; written off (50), fully recovered (296). and partially recovered (154)
        - Is recovered (Step 1): Written Off (50), Recovered (450, combining Fully and Partially Recovered).
        - Recovery level (Step 2, among Recovered): Fully Recovered (296), Partially Recovered (154).

```
Distribution of Recovery_Status:
Recovery_Status
Fully Recovered        296
Partially Recovered    154
Written Off             50
Name: count, dtype: int64
```



9.0. Data Cleansing

The dataset was clean with no missing numbers, no duplicates and no modification on column names.

To ensure modeling readiness and compliance with analytic standards, the following data cleansing steps have been undertaken

- Converted all applicable fields (e.g., Gender, Loan type) to categorical types to optimize model compatibility.
- Irrelevant Fields Dropped: Fields such as Borrower ID and Loan ID were excluded due to lack of predictive utility.
- Skew and Outlier Adjustment: Variables with pronounced skewness and outliers are candidates for log transformation, winsorization.
- Encoding Strategy: Multilevel categorical variables are to be converted using one-hot encoding (get_dummies), allowing seamless integration into modeling algorithms.
- Class Imbalance Strategy: Given the imbalance in Recovery status, oversampling (e.g., SMOTE) will be used to support model fairness and predictive performance

10.0. Summary

The data exploration confirms a clean and rich dataset suitable for building a classification-based prediction model. Key insights—such as payment behavior, loan terms, and borrower characteristics—demonstrate strong alignment with recovery outcomes and support the overall business objective of improving efficiency in debt recovery operations.

To enhance model performance and integrity:

- Features with a data type of objec*t* will be converted to categorical.
- Highly correlated variables will be dropped to prevent multicollinearity.
- Skewed numerical variables will be transformed to reduce bias in model predictions.
- The target variable will be balanced during training to ensure the model performs effectively across all outcome classes

Data Preparation and Feature Engineering

11.0. Data Preparation Needs

Several steps were undertaken to prepare the dataset for modeling:

- **Excluded Columns**: Non-informative or identifier columns such as Borrower ID and Loan ID were removed, as they do not contribute predictive value.
- All object-type variables were converted to categorical formats to ensure compatibility with machine learning models.
- **Handling Multicollinearity**: Highly correlated variables are dropped precisely loan amount and collection to reduce multicollinearity.
- **Skewness and Outlier Treatment**: Numerical variables such as Monthly EMI, Number of Missed Payments, and Collateral Value were skewed. To address this, capping and flooring techniques were applied to limit extreme values while preserving the overall distribution shape.

- **Categorical Encoding**: One-hot encoding (via get_dummies) was used to convert categorical variables into numerical format except the recovery status, making them suitable for modeling and dropped one.
- **Missing Values**: The dataset was confirmed to have no missing values; thus, no imputation was necessary.

11.1. (Build out as appropriate: censored records, excluded columns, imputations, transformations, etc.)

Columns were created based on the Recovery Status variable to serve as target variables for the two-step modeling approach.

- For Step 1, a new column named Is_Recovered was generated. It categorized records as either Written Off or Recovered, where Recovered included both Fully Recovered and Partially Recovered cases.
- For Step 2, a column named Recovery_Level was created, containing only the Fully Recovered and Partially Recovered records from Step 1. This allowed the model to further distinguish the degree of recovery among the recovered loans.

11.2. Upsampling, Downsampling, SMOTE

The target variable, Recovery Status, exhibited class imbalance—most loans were fully recovered, with fewer cases of partial recovery or write-offs. This imbalance can bias model predictions toward the majority class.

To address this, oversampling strategies was used only on the train dataset after splitting.

12.0. Feature Engineering

12.1. (Build out as appropriate: new variables, etc.)

Feature selection was guided by domain relevance and statistical techniques including correlation analysis, ANOVA, and Chi-square tests and random forest was ran to get feature importance which were used for some models

Model Exploration

13.0. Modeling Approach/Introduction

To identify key factors influencing loan recovery and to build a robust prediction system, a two-step classification modeling approach was applied:

- Step 1: Predict whether a loan is Written Off (0) or Recovered (1).
- Step 2: Among recovered loans, predict whether the loan was Partially Recovered (0) or Fully Recovered (1).

The modeling process involved multiple experiments with:

- Sampling strategies: Random Oversampling
- Varying test size: 40%
- The evaluation metrics used were Precision, F1 Score and recall

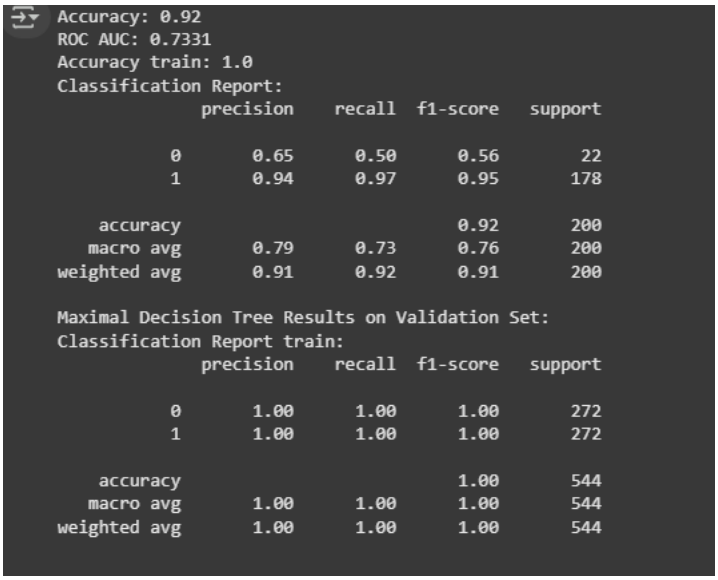
The models run included a maximal decision tree, grid search-tuned decision tree, random forest, and logistic regression. These were applied to both Step 1 and Step 2 of the analysis.

13. A Step 1

14.0. Maximum tree

Maximum tree; after splitting and balancing the train data set, a maximum decision tree was ran using a random state of 42. The classification report for both r=train and valid was obtained

showing that it overfits by performing well on the train set and poorly on the valid set as shown below



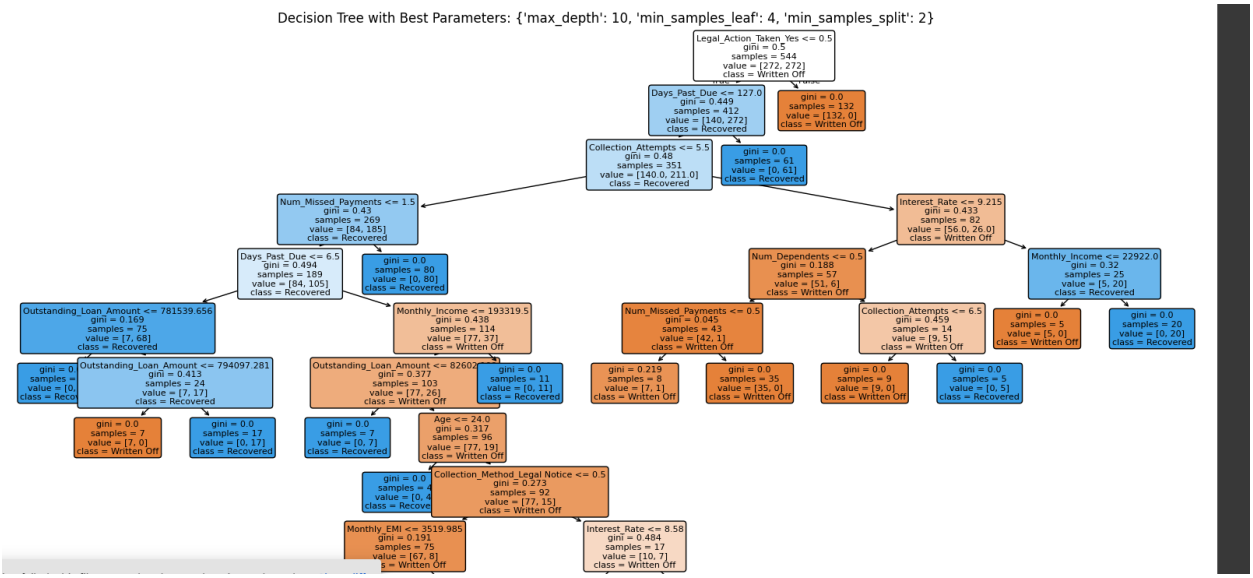
```
Accuracy: 0.92
ROC AUC: 0.7331
Accuracy train: 1.0
Classification Report:
              precision    recall  f1-score   support

           0       0.65      0.50      0.56        22
           1       0.94      0.97      0.95       178

    accuracy                           0.92       200
   macro avg       0.79      0.73      0.76       200
weighted avg       0.91      0.92      0.91       200

Maximal Decision Tree Results on Validation Set:
Classification Report train:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       272
           1       1.00      1.00      1.00       272

    accuracy                           1.00       544
   macro avg       1.00      1.00      1.00       544
weighted avg       1.00      1.00      1.00       544
```

15.0 Grid search tree

A hyperparameter tuning of a Decision Tree model was performed using GridSearchCV to identify the best combination of model parameters. The parameter grid included different values for max_depth, min_samples_split, and min_samples_leaf. A 5-fold cross-validated grid search was then conducted to evaluate all parameter combinations.

Once the optimal parameters were identified, the model was retrained using these settings and evaluated on a validation dataset. The results included the best hyperparameters, accuracy, classification report, and ROC AUC score to assess how well the tuned model performed on unseen data.

This approach optimized model performance while helping to reduce overfitting. The visualization of the tuned tree highlighted the top predictors for recovery prediction, which included **Legal_Action_Taken_Yes**, **Days_Past_Due**, **Collection_Attempts**, **Num_Missed_Payments**, **Outstanding_Loan_Amount**, and **Monthly_Income**

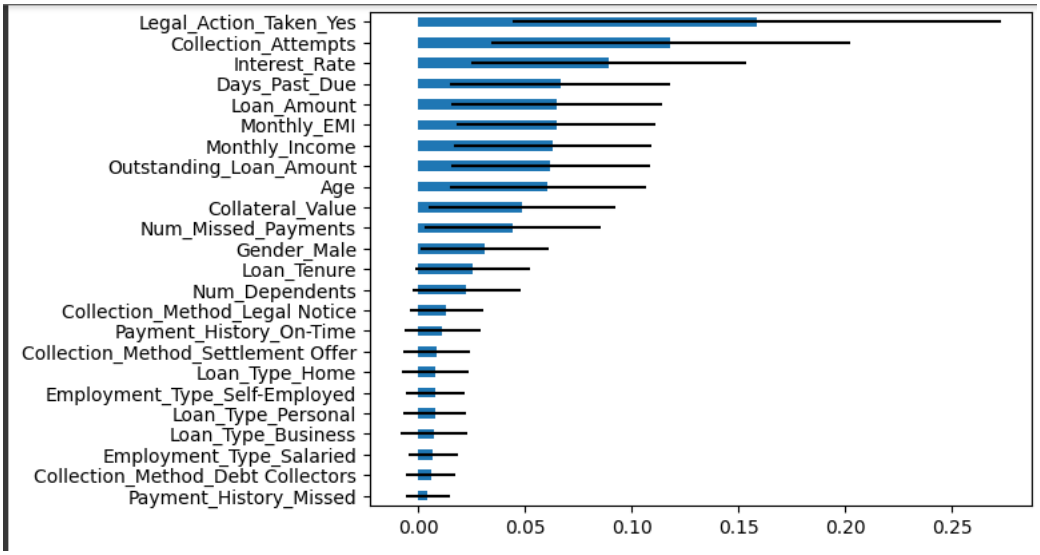


16.0 Random Forest

A Random Forest Classifier was trained using 500 decision trees (n_estimators=500) and a fixed random state to ensure reproducibility. The model was fitted on the resampled training dataset to ensure it is learning effectively. Predictions were then made on the validation set.

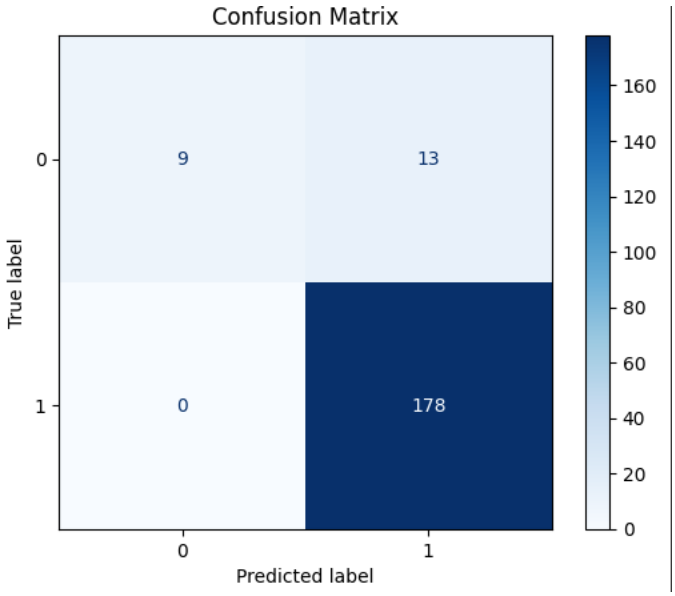
Feature importances were extracted to identify the most influential variables in the model's decision-making process. These importances, along with their standard deviations across all trees, were visualized using a horizontal bar chart.

A confusion matrix was generated to evaluate how well the model classified recovered and written-off loans. Additionally, a classification report was produced, summarizing the model's performance in terms of precision, recall, F1-score, and accuracy. Finally, the **ROC AUC score** was calculated to assess the model's ability to distinguish between recovery outcomes



Classification of feature importance from random forest

From this classification we see that legal action taken yes is the strongest feature followed by collection attempts.



The confusion matrix

From the confusion matrix, it shows that recovered was correctly predicted for all the instances as compared to written off where 13 out 22 were wrongly predicted.

```
Classification Report:

              precision    recall  f1-score   support

           0       1.00      0.41      0.58        22
           1       0.93      1.00      0.96       178

    accuracy                           0.94       200
   macro avg       0.97      0.70      0.77       200
weighted avg       0.94      0.94      0.92       200

accuracy: 0.935
ROC AUC: 0.7045454545454546
```

The classification report

17.0 Logistic regression

This was modelled using different features precisely

i) Logistic regression with only important features from random forest

ii) Logistic regression without correlated features

I) Logistic regression with only important features from random forest

Before running the models, data preprocessing was performed. Skewed variables were transformed to reduce their skewness to an acceptable level. The independent variable X was defined using only the important features identified by the random forest model, with categorical variables encoded using get_dummies. The target variable y was defined using a newly created column named is_recovered. The dataset was then split into 60% for training and 40% for validation due to the relatively small sample size. To address class imbalance, the training data was balanced using oversampling

a)Forward regression ; this was developed using forward feature selection. The Sequential Feature Selector was applied with 5-fold cross-validation and optimized based on the ROC AUC score, selecting the best combination of predictive features. The maximum number of iterations was set to 1000 to help the solver converge. The liblinear solver was chosen, as it is well-suited for small datasets.

The model was then trained on the resampled training data using only the selected features, with balanced class weights to address class imbalance. The floating parameter was set to true, ensuring that once a feature was added, it could not be removed later. The optimal number of features (k_features) was determined during selection.

Predictions were generated for the validation set. Model performance was evaluated using the ROC AUC score, accuracy, and a classification report detailing precision, recall, and F1-score. These results were used to assess the effectiveness of the selected features and the model's ability to distinguish between recovered and written-off loans in Stage 1. The hyperparameter tuning applied here was consistent with that used for all forward selection models in this project

```
Stage 1 (Recovered vs. written off) Results:
ROC AUC: 0.7377
Accuracy: 0.8250
Classification Report:
              precision    recall  f1-score   support

   Recovered       0.94      0.86      0.90       178
 Written Off       0.32      0.55      0.41        22

    accuracy                           0.82       200
   macro avg       0.63      0.70      0.65       200
weighted avg       0.87      0.82      0.84       200
```

b)Backward regression; the model was constructed using backward feature selection to determine the most influential variables for predicting whether a loan would be recovered or written off, with k_features set to the optimal value. Unlike forward selection, this method began with all available features and progressively eliminated the least impactful ones, guided by cross-validated ROC AUC performance. The floating parameter was set to False.

The model used the liblinear solver, which is well-suited for smaller datasets and binary classification tasks. Once the optimal subset of features was identified, the model was trained on the resampled dataset using balanced class weights to mitigate the effects of class imbalance. The trained model was then used to predict recovery outcomes on the validation set, producing both class probabilities and predictions.

Model performance was evaluated using the ROC AUC score, accuracy, and a classification report summarizing precision, recall, and F1-score. The hyperparameter tuning applied here was consistent with that used for all backward selection models in this project.

```
Stage 1 ( Recovered vs. written off) Results:
ROC AUC: 0.7480
Accuracy: 0.8350
Classification Report:
              precision    recall  f1-score   support

   Recovered       0.93      0.88      0.90       178
 Written Off       0.33      0.50      0.40        22

    accuracy                           0.83       200
   macro avg       0.63      0.69      0.65       200
weighted avg       0.87      0.83      0.85       200
```

c)Stepwise selection: A logistic regression model was developed using stepwise feature selection with forward=True and floating=True, enabling the process to dynamically add or remove features to optimize performance. The selection process was based on ROC AUC using 5-fold cross-validation.

The liblinear solver was employed due to its suitability for small datasets and binary classification. The maximum number of iterations (max_iter) was set to 1000 to ensure convergence. The best-performing subset of features was identified and used to train the model on the resampled training dataset with balanced class weights to address class imbalance.

Predictions and class probabilities were generated for the validation set, and model performance was assessed using ROC AUC, accuracy, and a classification report with precision, recall, and F1-score. The hyperparameter tuning applied here matched that used for all stepwise selection models in this project.

```
Stage 1(Recovered vs. written off) Results:
ROC AUC: 0.7474
Accuracy: 0.8300
Classification Report:
              precision    recall  f1-score   support

   Recovered       0.93      0.88      0.90       178
 Written Off       0.31      0.45      0.37        22

    accuracy                           0.83       200
   macro avg       0.62      0.67      0.64       200
weighted avg       0.86      0.83      0.84       200
```

II) Logistic regression without correlated features

The same data preprocessing steps described earlier were applied for this stage, with the key difference occurring at the selection of the X variables. Highly correlated features Loan Amount and Collection Method were excluded to avoid multicollinearity. All remaining features, except Recovery Status and IsRecovered, were included as predictors. Categorical variables were encoded using the get_dummies() function to ensure proper handling in the model.

The Y target variable was set to IsRecovered. The same hyperparameter tuning process used in the earlier regression models was applied here for consistency. Notably, there was a variation in model performance compared to the previous setup, highlighting the impact of feature selection on predictive accuracy

a)Forward regression ;

```
Stage 1 (Recovered vs. written off) Results:
ROC AUC: 0.7612
Accuracy: 0.8450
Classification Report:
              precision    recall  f1-score   support

   Recovered       0.95      0.88      0.91       178
 Written Off       0.37      0.59      0.46        22

    accuracy                           0.84       200
   macro avg       0.66      0.73      0.68       200
weighted avg       0.88      0.84      0.86       200
```

b) Backward selection

```
Stage 1 ( Recovered vs. written off) Results:
ROC AUC: 0.7634
Accuracy: 0.8350
Classification Report:
              precision    recall  f1-score   support

   Recovered       0.94      0.87      0.90       178
 Written Off       0.35      0.59      0.44        22

    accuracy                           0.83       200
   macro avg       0.65      0.73      0.67       200
weighted avg       0.88      0.83      0.85       200
```

c) Stepwise selection

```
Stage 1(Recovered vs. written off) Results:
ROC AUC: 0.7634
Accuracy: 0.8350
Classification Report:
              precision    recall  f1-score   support

   Recovered       0.94      0.87      0.90       178
 Written Off       0.35      0.59      0.44        22

    accuracy                           0.83       200
   macro avg       0.65      0.73      0.67       200
weighted avg       0.88      0.83      0.85       200
```

B) STEP 2 (fully recovered and partially recovered)

The classification within the recovered loan group was further refined by creating a new target variable, Recovery_Level, in the stage 2. This variable was defined as:

- 1 for Fully Recovered loans, and
- 0 for Partially Recovered loans.

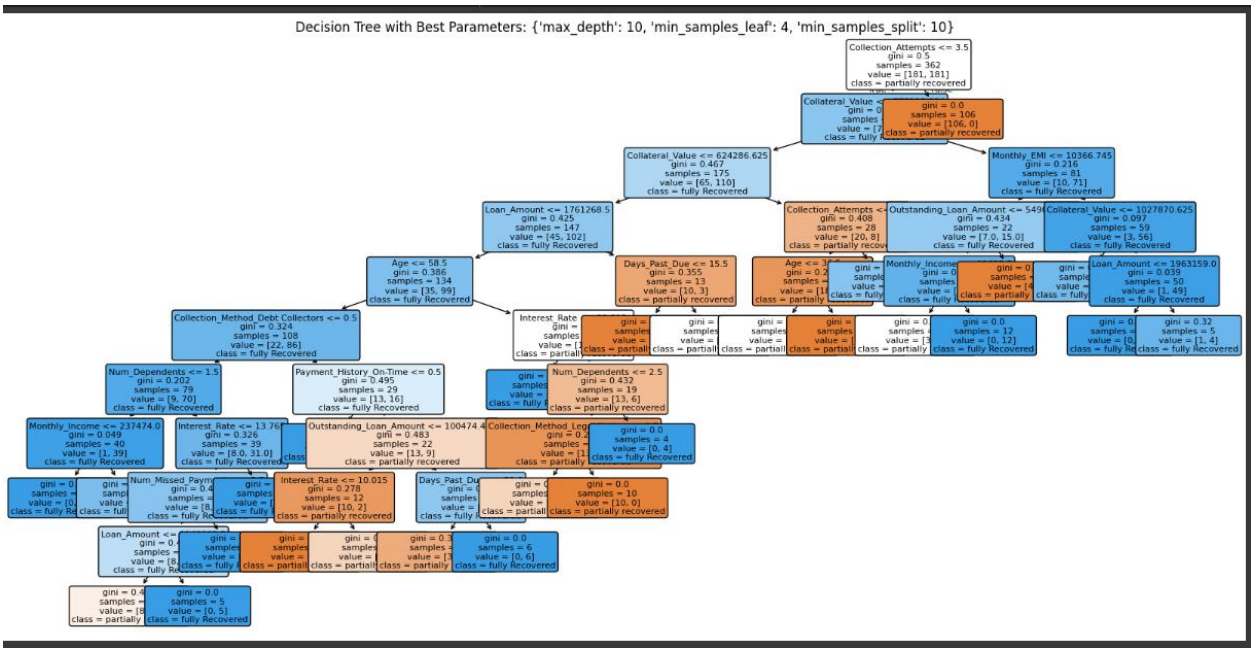
This resulted in 296 records labeled as Fully Recovered (1) and 154 records labeled as Partially Recovered (0). The dataset was then split into training and validation sets using a test size of 40%. To address class imbalance in the training set, oversampling was applied prior to model training

18.0 Maximum tree; the hyperparameter tuning as in step 1 was applied here. For the X variable the following field were excluded recovery status, is recovered, loan ID, Borrower ID and recovery level. The classification of valid data was obtained.

```
Accuracy: 0.78
ROC AUC: 0.7535
Accuracy train: 1.0
Classification Report:
              precision    recall  f1-score   support

           0       0.72      0.65      0.68        65
           1       0.81      0.86      0.84       115

    accuracy                           0.78       180
   macro avg       0.77      0.75      0.76       180
weighted avg       0.78      0.78      0.78       180
```

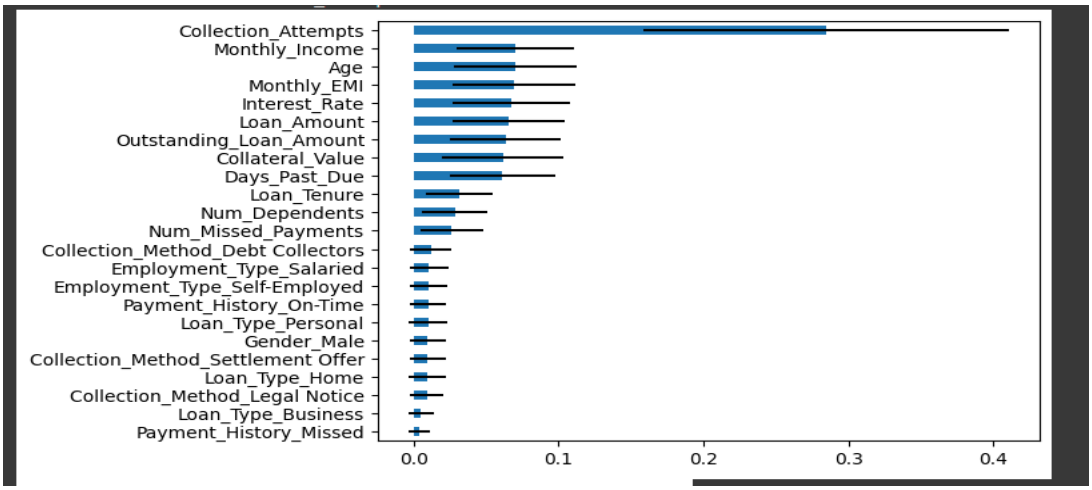
19.0) Grid search cv ; the model was also tuned exactly as in step 1 . the best parameters obtained were max depth as 10, minimum sample leaf as 4 and minimum sample split as 10.

Here the splitting starts at collection attempts less than or equal to3.5 .Below is the classification report of the valid set is as follow




```
Decision Tree Accuracy: 0.8111

Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.66      0.72        65
           1       0.82      0.90      0.86       115

    accuracy                           0.81       180
   macro avg       0.80      0.78      0.79       180
weighted avg       0.81      0.81      0.81       180

ROC AUC score: 0.7785953177257525
```

20.0) Random Forest; The model was tuned exactly as in Step 1. Analysis of the feature importance rankings showed that Collection_Attempts was ranked first in descending order of importance, while Legal_Action_Taken had no predictive significance—an encouraging result from a business perspective, as it suggests that effective recovery outcomes can be achieved without resorting to costly and time-consuming legal actions. The graph below illustrates the ranked feature importances, followed by the corresponding classification report




```
Classification Report:

              precision    recall  f1-score   support

           0       0.93      0.63      0.75        65
           1       0.82      0.97      0.89       115

    accuracy                           0.85       180
   macro avg       0.88      0.80      0.82       180
weighted avg       0.86      0.85      0.84       180

accuracy ; 0.85
ROC AUC score: 0.8023411371237459
```

21 I)Logistic regression with important features from random forest

As this is a new step before running the models, data preprocessing was performed. Skewed variables were transformed to reduce their skewness to an acceptable level. The independent variable X was defined using only the important features identified by the random forest model, with categorical variables encoded using get_dummies. The target variable y was defined using a newly created column named recovery level. The dataset was then split into 60% for training and 40% for validation due to the relatively small sample size. To address class imbalance, the training data was balanced using oversampling. The same hyper parameter tuning for all the regressions in step 1 were also applied here for their respective type. The results of the models are as below

1)forward regression; the selected features are Age, Num_Dependents, Interest_Rate, Days_Past_Due, Collection_Attempts, Loan_Type_Business, Payment_History_Missed, Collection_Method_Settlement Offer

```
Stage 2 (Fully Recovered vs. Partially Recovered) Results:
ROC AUC: 0.8107
Accuracy: 0.8333
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.71      0.75        65
           1       0.85      0.90      0.87       115

    accuracy                           0.83       180
   macro avg       0.83      0.81      0.81       180
weighted avg       0.83      0.83      0.83       180
```

2)Backward regression; selected features are Age, Monthly_Income, Num_Dependents, Loan_Amount, Interest_Rate, Days_Past_Due, Collection_Attempts, Gender_Male, Employment_Type_Salaried, Employment_Type_Self-Employed, Loan_Type_Business, Loan_Type_Home, Payment_History_Missed, Collection_Method_Debt Collectors, Collection_Method_Settlement Offer

```
Stage 2 (Fully Recovered vs. Partially Recovered) Results:
ROC AUC: 0.8146
Accuracy: 0.7611
Classification Report:
              precision    recall  f1-score   support

           0       0.68      0.65      0.66        65
           1       0.81      0.83      0.82       115

    accuracy                           0.76       180
   macro avg       0.74      0.74      0.74       180
weighted avg       0.76      0.76      0.76       180
```

3) Step wise regression ; selected features are Age, Num_Dependents, Interest_Rate, Days_Past_Due, Collection_Attempts, Loan_Type_Business, Payment_History_Missed, Collection_Method_Settlement Offer

```
Stage 2 (Fully Recovered vs. Partially Recovered) Results:
ROC AUC: 0.8107
Accuracy: 0.8333
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.71      0.75        65
           1       0.85      0.90      0.87       115

    accuracy                           0.83       180
   macro avg       0.83      0.81      0.81       180
weighted avg       0.83      0.83      0.83       180
```

II)Logistic regression without the correlated features

The exact data preprocessing was performed as in the previous step, with the difference occurring at the level of the X variables. The correlated features, loan_amount and collection_method, were excluded from the model, along with recovery_status, recovery_level, and is_recovered. All remaining categorical variables were encoded using the get_dummies function.

The target variable y was defined as recovery_level. The same hyperparameter tuning procedures used for all regressions in Step 1 were also applied here for their respective model types. The results of the models are presented below.

A. forward regression; the selected features were Age, Interest_Rate, Days_Past_Due, Collection_Attempts, Gender_Male, Employment_Type_Self-Employed, Loan_Type_Home, Payment_History_Missed, Payment_History_Missed

```
Stage 2 (Fully Recovered vs. Partially Recovered) Results:
ROC AUC: 0.7988
Accuracy: 0.7611
Classification Report:
              precision    recall  f1-score   support

           0       0.68      0.63      0.66        65
           1       0.80      0.83      0.82       115

    accuracy                           0.76       180
   macro avg       0.74      0.73      0.74       180
weighted avg       0.76      0.76      0.76       180
```

B) Backward regression ; the selected features were Age, Interest_Rate, Days_Past_Due, Collection_Attempts, Gender_Male, Employment_Type_Salaried, Loan_Type_Home, Payment_History_Missed, Legal_Action_Taken_Yes, Payment_History_Missed

```
Stage 2 (Fully Recovered vs. Partially Recovered) Results:
ROC AUC: 0.7972
Accuracy: 0.7500
Classification Report:
              precision    recall  f1-score   support

           0       0.66      0.63      0.65        65
           1       0.80      0.82      0.81       115

    accuracy                           0.75       180
   macro avg       0.73      0.72      0.73       180
weighted avg       0.75      0.75      0.75       180
```

C) Step wise selection; the selected features are Age, Interest_Rate, Days_Past_Due, Collection_Attempts, Gender_Male, Employment_Type_Self-Employed, Loan_Type_Home, Payment_History_Missed, Payment_History_Missed

```
Stage 2 (Fully Recovered vs. Partially Recovered) Results:
ROC AUC: 0.7988
Accuracy: 0.7611
Classification Report:
              precision    recall  f1-score   support

           0       0.68      0.63      0.66        65
           1       0.80      0.83      0.82       115

    accuracy                           0.76       180
   macro avg       0.74      0.73      0.74       180
weighted avg       0.76      0.76      0.76       180
```

22.0. Model Comparison

For step 1 predicting recovered and written off below is a table summarizing the metrics of all the models

Model	Accuracy	ROC AUC	Precision (0/1)	Recall (0/1)	F1-Score (0/1)
Maximal Tree	0.92	0.7331	0.65 / 0.94	0.50 / 0.97	0.56 / 0.95
Grid Search	0.915	0.7331	0.65 / 0.94	0.50 / 0.97	0.56 / 0.95
Random Forest	0.935	0.7045	1.00 / 0.93	0.41 / 1.00	0.58 / 0.96
Regression with Important Features					
Forward Reg	0.825	0.7377	0.32 / 0.94	0.55 / 0.86	0.41 / 0.90
Backward Reg	0.835	0.748	0.33 / 0.93	0.50 / 0.88	0.40 / 0.90
Stepwise Reg	0.83	0.7474	0.31 / 0.93	0.45 / 0.88	0.37 / 0.90
Regression without Correlated Features					
Forward Reg	**0.845**	**0.7612**	**0.37 / 0.95**	**0.59 / 0.88**	**0.46 / 0.91**
Backward Reg	0.835	0.7634	0.35 / 0.94	0.59 / 0.87	0.44 / 0.90
Stepwise Reg (No Corr)	0.835	0.7634	0.35 / 0.94	0.59 / 0.87	0.44 / 0.90

The focus of the precision metric is on 1 (recovered) saying that when the model predicts recovered, it is 95 % correct. The ROC AUC reveals that the model classifies the written off and recovered 76.12% correctly.

Step 2 summary metrics

The precision metric is centered on 1 (fully recovered) saying that when the model predicts fully recovered, it is 85 % correct. The ROC AUC reveals that the model classifies the written off and recovered 81.07% correctly.

Model	Accuracy	ROC AUC	Precision (0/1)	Recall (0/1)	F1-Score (0/1)
Maximal Tree	0.78	0.755	0.72 / 0.81	0.65 / 0.86	0.68 / 0.84
Grid Search	0.716	0.708	0.59/ 0.80	0.68 / 0.74	0.63/ 0.77
Random Forest	0.83	0.782	0.91 / 0.81	0.60 / 0.97	0.72/0.88
Regression with Important Features					
Forward Reg	**0.833**	**0.8107**	**0.81/0.85**	**0.71/0.9**	**0.75/0.87**
Backward Reg	0.76	0.8146	0.68/0.81	0.65/0.83	0.66/0.82
Stepwise Reg	0.833	0.8107	0.81/0.85	0.71/0.9	0.75/0.87
Regression without Correlated Features					
Forward Reg	0.76	0.798	0.68 / 0.80	0.63/ 0.83	0.66 / 0.82
Backward Reg	0.75	0.7972	0.66/0.80	0.63/0.82	0.65/0.81
Stepwise Reg	0.7611	0.7988	0.94 / 0.35	0.87 / 0.59	0.90 / 0.44

23.0 Model Selection

The model selection was based on precision, supported by ROC AUC. In this project, a false positive—either false recovered for Step 1 or false fully recovered for Step 2—will be costly, precisely leading to overestimation of financial status and potential loss of high loan amounts due to inappropriate use of strategy. This justifies our first choice of precision as metrics. For Step 1, the best model was forward regression using only important features from Random Forest, with a precision of 0.37 and 0.95 for 0 and 1, and ROC AUC of 0.76. For Step 2, both forward and stepwise regression built with no correlated features were the best, having the same results. So, forward was chosen with a precision of 0.81 and 0.85 for both 0 and 1, and ROC AUC of 0.810719.

Below are their respective odds ratio and their P values of the selected features. From their respective P values, we could see that not all are statistically significant with their value being greater than 0.05

Step1

	coef	std err	z	P>|z|	[0.025	0.975]
Num_Dependents	0.2574	0.112	2.299	0.021	0.038	0.477
Monthly_EMI	1.206e-05	1.26e-05	0.955	0.340	-1.27e-05	3.68e-05
Num_Missed_Payments	0.6627	0.155	4.277	0.000	0.359	0.966
Days_Past_Due	-0.0019	0.003	-0.708	0.479	-0.007	0.003
Collection_Attempts	-0.1566	0.047	-3.329	0.001	-0.249	-0.064
Gender_Male	-1.4198	0.329	-4.312	0.000	-2.065	-0.775
Employment_Type_Salaried	0.5791	0.414	1.400	0.161	-0.231	1.390
Employment_Type_Self-Employed	0.0094	0.435	0.022	0.983	-0.843	0.861
Loan_Type_Business	-2.1396	0.669	-3.200	0.001	-3.450	-0.829
Loan_Type_Home	-2.0122	0.603	-3.337	0.001	-3.194	-0.830
Loan_Type_Personal	-1.8641	0.595	-3.135	0.002	-3.030	-0.699
Payment_History_On-Time	1.1633	0.262	4.435	0.000	0.649	1.677
Legal_Action_Taken_Yes	-33.2974	6.63e+05	-5.03e-05	1.000	-1.3e+06	1.3e+06
const	1.8958	0.705	2.689	0.007	0.514	3.278

	Odds Ratio	Conf. Int. (2.5%)	Conf. Int. (97.5%)
Num_Dependents	1.293545e+00	1.038706	1.610907
Monthly_EMI	1.000012e+00	0.999987	1.000037
Num_Missed_Payments	1.940079e+00	1.431956	2.628508
Days_Past_Due	9.980580e-01	0.992700	1.003445
Collection_Attempts	8.550725e-01	0.779774	0.937642
Gender_Male	2.417570e-01	0.126801	0.460930
Employment_Type_Salaried	1.784468e+00	0.793455	4.013241
Employment_Type_Self-Employed	1.009403e+00	0.430521	2.366655
Loan_Type_Business	1.177026e-01	0.031743	0.436445
Loan_Type_Home	1.336916e-01	0.040999	0.435945
Loan_Type_Personal	1.550341e-01	0.048338	0.497239
Payment_History_On-Time	3.200317e+00	1.913888	5.351426
Legal_Action_Taken_Yes	3.460329e-15	0.000000	inf
const	6.658112e+00	1.671581	26.520067

Gender Male is statistically significant but won't be selected as a predictor because this is against the Canadian regulatory policy (FCAC/PIPEDA)

Step 2

	coef	std err	z	P>|z|	[0.025	0.975]
Age	-0.0092	0.010	-0.929	0.353	-0.029	0.010
Num_Dependents	0.2178	0.113	1.930	0.054	-0.003	0.439
Interest_Rate	0.0296	0.035	0.857	0.392	-0.038	0.097
Days_Past_Due	0.0035	0.002	1.603	0.109	-0.001	0.008
Collection_Attempts	-0.5586	0.066	-8.505	0.000	-0.687	-0.430
Loan_Type_Business	0.0364	0.434	0.084	0.933	-0.815	0.888
Payment_History_Missed	1.2431	0.555	2.242	0.025	0.156	2.330
Collection_Method_Settlement Offer	0.7436	0.341	2.184	0.029	0.076	1.411
const	0.8623	0.589	1.464	0.143	-0.292	2.017

Odds Ratios for forkward Regression (Stage 2):

	Odds Ratio	Conf. Int. (2.5%)	Conf. Int. (97.5%)
Age	0.990877	0.971896	1.010229
Num_Dependents	1.243316	0.996622	1.551073
Interest_Rate	1.030010	0.962627	1.102110
Days_Past_Due	1.003484	0.999226	1.007759
Collection_Attempts	0.572030	0.502938	0.650614
Loan_Type_Business	1.037068	0.442753	2.429145
Payment_History_Missed	3.466495	1.169075	10.278708
Collection_Method_Settlement Offer	2.103442	1.079152	4.099949
const	2.368579	0.746673	7.513553

24.0Model theory

24.1 Model Assumptions and Limitations

Limitations

-Though the logistic regression performs well on small data set there is still some reservation on generalizing perfectly to new, unseen data

-Risk of overfitting

- The model is static, based on historical data. Changes in loan policies, collection strategies, or economic conditions over time could affect its predictive power.

- Some findings (like Num_Missed_Payments increasing the odds of recovery) are counter-intuitive. While the model may be picking up complex relationships, it highlights a limitation in direct interpretability without further investigation into potential interactions or confounding factors.

Assumptions

-It assumes that the observations (loans in this case) are independent of each other.

- The selected features are optimal for this model and the chosen evaluation metric (precision)

-The combined effect of predictors is assumed to be the sum of their individual effects

25.0 Model Sensitivity to Key Drivers

The key drivers were selected based on their statistical significance at a 95% confidence interval

For step 1 the key drivers are

- **Num Dependents (Odds Ratio = 1.294):**
    - **Sensitivity:** The model is sensitive to the number of dependents. A one-unit increase in the number of dependents increases the odds of recovery by about 29.4%. This translates to a higher predicted probability of recovery as the number of dependents increases
- **Num Missed Payments (Odds Ratio = 1.940):**
    - **Sensitivity:** The model is sensitive to the number of missed payments. A one-unit increase in missed payments nearly doubles the odds of recovery (an increase of 94%). This implies a higher predicted probability of recovery for loans with more missed payments according to this model. As noted before, this finding is counter-intuitive and warrants further investigation.
- **Collection Attempts (Odds Ratio = 0.855):**
    - **Sensitivity:** The model is sensitive to the number of collection attempts. Each additional collection attempt decreases the odds of recovery by about 14.5%. This

means a higher number of collection attempts leads to a lower predicted probability of recovery

- **Gender Male (Odds Ratio = 0.242):**
    - **Sensitivity:** The model is very sensitive to the borrower's gender. Being male (compared to female) decreases the odds of recovery by about 75.8%. This indicates a significantly lower predicted probability of recovery for male borrowers
- **Loan Type (Business, Home, Personal) (Odds Ratios < 1, e.g., Business = 0.118)**
    - **Sensitivity:** The model is sensitive to the loan type. Having a Business, Home, or Personal loan (compared to the reference category) significantly decreases the odds of recovery. This translates to a much lower predicted probability of recovery for these loan types
- **Payment History On-Time (Odds Ratio = 3.200):**
    - **Sensitivity:** The model is highly sensitive to having an on-time payment history. An on-time history (compared to the reference, likely Missed) increases the odds of recovery by 3.2 times. This results in a substantially higher predicted probability of recovery for loans with a good payment history

The key drivers for step 2 are collection attempts, collection method settlement offer and payment history missed which have a substantial impact on the predicted probability of a loan being fully recovered.

- **Collection Attempts:**
    - **Odds Ratio:** 0.572
    - **Interpretation:** For each additional collection attempt, the odds of a loan being Fully Recovered decrease by about 42.8%.
    - **Sensitivity:** The model is sensitive to the number of collection attempts. An increase in collection attempts leads to a *lower* predicted probability of full recovery. This suggests that loans requiring more intervention are harder to bring to full recovery. The model captures this negative relationship, indicating that

focusing resources on loans earlier in the collection process might yield better results for achieving full recovery.

- **Payment History Missed:**
    - **Odds Ratio:** 3.466
    - **Interpretation:** Loans with a missed payment history have about 3.47 times higher odds of being Fully Recovered compared to the reference (likely On-Time).
    - **Sensitivity:** The model is quite sensitive to whether a loan has a missed payment history. Having a missed history (as opposed to the reference) significantly *increases* the predicted probability of full recovery according to this model. As discussed before, this is a counter-intuitive finding. If this relationship holds true, it could imply that loans with a history of missed payments are more actively pursued with effective recovery strategies within the 'recovered' pool, leading to a higher likelihood of ultimately being fully recovered. The model strongly weighs this factor.
- **Collection Method Settlement Offer:**
    - **Odds Ratio:** 2.103
    - **Interpretation:** Using a Settlement Offer has about 2.10 times higher odds of leading to a Fully Recovered loan compared to the reference collection method (likely Calls).
    - **Sensitivity:** The model is sensitive to the collection method used. When the method is 'Settlement Offer', the predicted probability of full recovery is significantly *higher* compared to the reference method. This suggests that, based on the data, offering a settlement is a more effective strategy for achieving full recovery. The model assigns a notable positive weight to this collection method.

Conclusion and Recommendations

26.0. Impacts on Business Problem

The combined scope of these models is to provide predictive insights at two critical junctures in the loan recovery process. This allows for:

- Early identification of high-risk-of-write-off loans.
- Informed decision-making on whether to pursue intensive recovery efforts.
- Optimization of recovery strategies for loans deemed recoverable.
- Better prediction of the likely recovery amount (full vs. partial).
- More efficient allocation of limited collection resources.

By operationalizing the predictions and insights from these models, the business can move from reactive, one-size-fits-all recovery approaches to a more proactive, data-driven, and segmented strategy, directly addressing the issues of inefficiency, misallocation, and high unrecovered loans.

27.0. Recommendations

Implement loyalty-based retention programs (e.g., interest rate incentives, fee waivers) to maintain high repayment rates while minimizing collection cost, incorporate settlement-based recovery strategies, especially for delinquent accounts flagged as having moderate recovery potential, to increase resolution rates while reducing prolonged collection costs and introduce stricter credit assessments and collateral backed lending for new business loans. For existing high-risk accounts, apply early settlement negotiations to mitigate write-offs.