ALPEREN UNAL

OPTIMIZING CREDIT DEFAULT PREDICTION: ENHANCING FINANCIAL SECURITY THROUGH ADVANCED ANALYTICS

1. INTRODUCTION

1.1 Problem Statement and Project Goal

The core objective of this project is to develop a predictive model for accurately identifying potential credit card defaulters. This task is critical for financial institutions to prevent substantial monetary losses. The model will classify individuals into two groups: '0' for non-defaulters and '1' for defaulters, with 'Defaulter' being the target variable. The primary focus is on achieving high recall, ensuring the model effectively captures as many defaulters as possible. This emphasis is due to the significant financial risk posed by missing defaulters, as each undetected defaulter could lead to considerable financial loss. Therefore, the project's success hinges on creating a robust model that prioritizes the detection of defaulters while maintaining an acceptable level of precision to minimize the misclassification of non-defaulters. The overarching goal is to safeguard the financial health of the credit card issuer by implementing a reliable and efficient predictive tool for credit default risk assessment.

1.2Data Preparation and Overview

Setting the Seed: Explanation of why setting a seed (set.seed(123)) is critical for reproducibility in statistical simulations and random number generation.

Dataset Assignment: Description of assigning imported datasets to df_train and df_test.

Data Dimension Analysis: Analysis of the number of inputs (rows) and variables (columns) in the training and test sets, using nrow() and ncol() functions.

2. DATA CLEANING, PREPROCESSING AND EDA

2.1 Data Cleaning and Preprocessing

Missing Value Analysis: Overview of the approach to check for missing values using colSums(is.na(df_train)) and its implications.

Variable Type Inspection: Understanding variable types in the dataset (sapply(df_train, class)) is important.

Renaming Variables: Rationalization of renaming variables for better understandability and clarity.

Calculation of Proportions: A function calculate_proportion is defined to calculate the proportion of specific values across repayment history columns. This is applied to each column to gain insights into the distribution of values.

Correlation Analysis: Correlation analysis is fundamental in any predictive modeling process as it helps identify variables that have a strong linear relationship with the target variable. The correlations are sorted by their absolute values, providing a clear picture of which variables are most strongly related to the target.

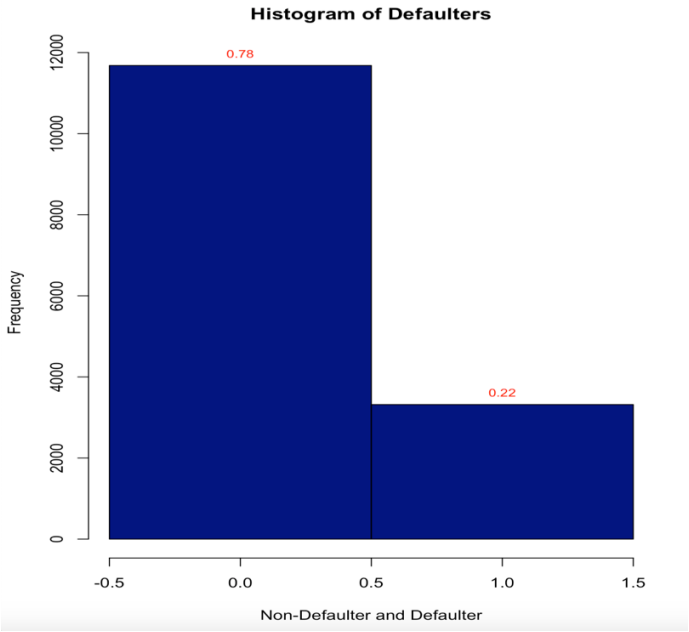
Categorical and Numerical Variable Separation: This distinction is important because categorical and numerical data require different types of processing and analysis. Correctly identifying the nature of each variable ensures that appropriate modeling techniques are used.

Align the factor levels: In the training and test datasets and address discrepancies in unseen categories, the preprocessing step harmonizes the REPAY_AUG, REPAY_MAY, and REPAY_APR variables. Categories not present in both datasets are consolidated, with excess levels dropped to ensure consistency across the data, facilitating accurate model training and prediction.

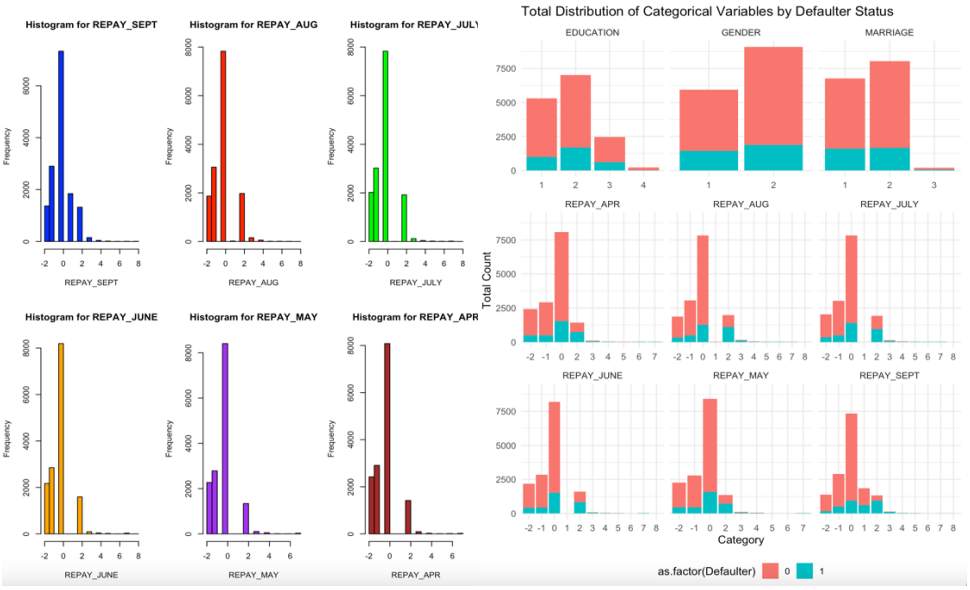
2.2 Exploratory Data Analysis

Histogram Analysis: Explanation of creating a histogram to visualize the distribution of defaulters and non-defaulters in the training set. Importance of visual data inspection.

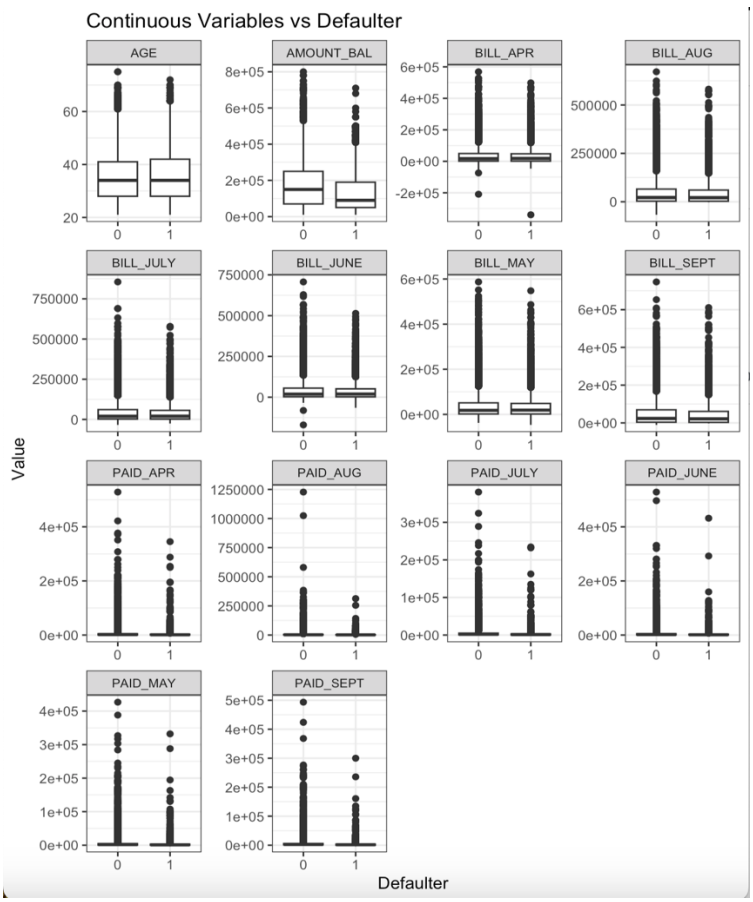
Default Payment Proportion: Discussion on the calculated proportions of defaulters and non-defaulters and implications of data imbalance on model training and prediction accuracy.



Exploring 'History of Past Payment': Exploring variables related to repayment history (REPAY_SEPT, REPAY_AUG, etc.) through histograms.



Continuous varibles box-plot:



The box plots reveal a trend where defaulters generally exhibit higher median balances and bills, hinting at a link between increased financial obligations and the likelihood of defaulting. Age appears to have no distinctive impact on defaulting behavior. The presence of outliers underscores the diverse financial habits among individuals.

2.3 Data Transformation

Handling Anomalous Values in 'EDUCATION': The project identifies undocumented values in the EDUCATION variable (0, 5, 6) and appropriately recodes them as '4', which is assumed to represent 'other'. This same method is applied to both the training and test datasets.

Addressing Issues in 'MARRIAGE': Similar to EDUCATION, the MARRIAGE variable contains undocumented '0' values, which are then recoded to '3'.

3. FURTHER ANALYSIS

Possibly the beginnings of a predictive model. This would involve statistical tests to understand correlations, feature selection methods to identify significant predictors, and preliminary steps towards building a machine learning model.

3.1 Deep Dive into Repayment History Variables

Analysis of '0' Value:

- A subset of the data where all repayment variables are '0' is sampled.
- It's observed that these individuals often have a large discrepancy between bill statements and previous payments, yet most are categorized as non-defaulters.
- This leads to the interpretation that a '0' value in the repayment history likely signifies the *minimum amount of payment.*

Exploring '-2' Value:

- Another subset is created for cases where all repayment variables are '-2'.
- Analysis of this subset reveals that bill statements and payment amounts are closely aligned, and most individuals in this sample are non-defaulters.
- The '-2' value is thus interpreted as indicative of *fully paid balances or inactive credit cards.*

Analysis of '-1' Value:

- The script notes that for the '-1' value in repayment variables, there is a consistency between bill statements and payment amounts, akin to what was observed for the '-2' value.
- This observation suggests that the '-1' value, like '-2', indicates a situation where the client has made sufficient payments on their credit card.

Redefining 'Duly Payment':

- Based on the analysis, the values '-1', '-2', and '0' are all indicative of 'duly payment', meaning the clients have paid sufficient money towards their credit cards.

- To simplify the model and improve its interpretability, these values are combined under a single category. In the context of this project, '0' is redefined to mean 'pay duly'.

3.2 Deeply Variable Evaluation with Backward Selection Algorithm

Logistic regression with backward selection refines the predictive model for credit defaults.
Key predictors identified:
REPAY_SEPT: Strong positive correlation with defaults.
AMOUNT_BAL: Higher balances are inversely related to default risk.
Demographics: Gender and marital status contribute to prediction.

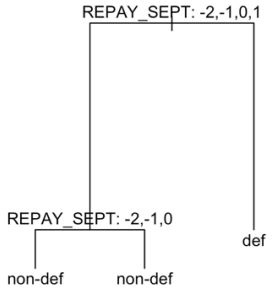
The reduced model achieves:

- A balance between simplicity and predictive accuracy.
- An interpretable framework for risk assessment.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.265e-01  1.684e-01  -3.721 0.000198 ***
AMOUNT_BAL  -8.525e-07  2.220e-07  -3.841 0.000123 ***
GENDER      -1.000e-01  4.349e-02  -2.300 0.021430 *
EDUCATION   -7.290e-02  3.134e-02  -2.326 0.020013 *
MARRIAGE    -1.842e-01  4.492e-02  -4.101 4.12e-05 ***
AGE          5.581e-03  2.538e-03   2.199 0.027848 *
REPAY_SEPT   5.891e-01  2.513e-02  23.436  < 2e-16 ***
REPAY_AUG    1.180e-01  2.449e-02   4.817 1.46e-06 ***
REPAY_MAY    8.647e-02  2.299e-02   3.761 0.000169 ***
BILL_SEPT   -6.743e-06  1.614e-06  -4.177 2.95e-05 ***
BILL_AUG     4.080e-06  1.874e-06   2.177 0.029452 *
BILL_JUNE    2.032e-06  1.047e-06   1.941 0.052275 .
PAID_SEPT   -1.189e-05  3.029e-06  -3.924 8.70e-05 ***
PAID_AUG    -1.378e-05  3.160e-06  -4.361 1.30e-05 ***
PAID_JULY   -7.565e-06  2.755e-06  -2.746 0.006028 **
PAID_JUNE   -2.875e-06  2.017e-06  -1.425 0.154136
PAID_MAY    -3.169e-06  2.067e-06  -1.533 0.125186
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15853  on 14999  degrees of freedom
Residual deviance: 13868  on 14983  degrees of freedom
AIC: 13902

Number of Fisher Scoring iterations: 6
```

4. MODEL ANALYSIS

4.1. Decision Tree 1 (With Selected Variables)

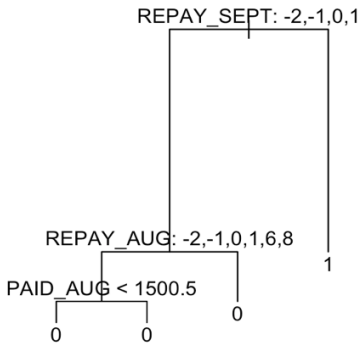


- Utilized a decision tree model on the df_train_new dataset with variables selected via backward elimination.
- Recoded binary and categorical variables for clarity:

Defaulter: Levels set to "non-def" and "def" for non-defaulters and defaulters, respectively.
GENDER: Levels labeled "male" and "female".
MARRIAGE: Categories labeled "married", "single", "others".
Repayment statuses for September, May, and April: Converted to factors to handle categorical nature.

- The decision tree, when plotted, revealed inadequacy due to a lack of sufficient nodes, implying over-simplification.
- Cross-validation for pruning did not yield a more complex tree, leading to a single-node tree.
- This indicates that the variables selected through backward selection may not provide enough depth for tree-based modelling.
- The decision tree's methodology differs from logistic regression, resulting in an overly simplistic model when using the same variables.

4.2 Decision Tree



- Reconstructed a decision tree model (tree_model2) using the original df_train dataset without variable removal.
- Converted multiple variables into factors to properly encode categorical data for the tree algorithm:

Demographic Factors: GENDER, MARRIAGE, EDUCATION.
Repayment Status: REPAY_SEPT, REPAY_MAY, REPAY_APR, REPAY_AUG, REPAY_JUNE, REPAY_JULY.

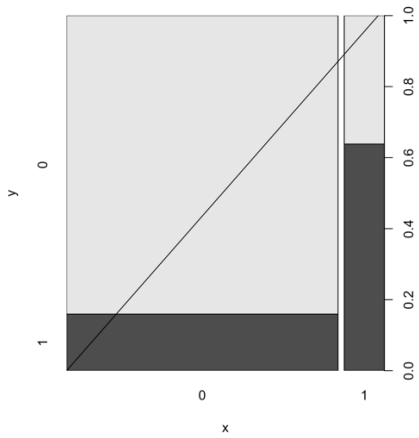
Target Variable: Defaulter coded as a factor.

- The new tree model is plotted, and a summary indicates a more complex structure compared to the pruned model from previous steps.
- Evaluation of tree_model2 with df_test dataset yields:

Accuracy: 81.73%, a promising result indicating a high level of correct predictions.
Precision and Recall: Calculated to assess model performance in identifying defaulters.

- Pruning of tree_model2 performed with cv.tree() identifies an optimal tree size.
- Post-pruning, the tree (pruned_tree2) maintains the same level of accuracy (81.73%), suggesting pruning did not detrimentally affect the model's predictive ability.
- The decision tree, both before and after pruning, demonstrates consistent performance, indicating robustness in the model's ability to predict defaulters with the full set of variables.

4.3 Bagging Model

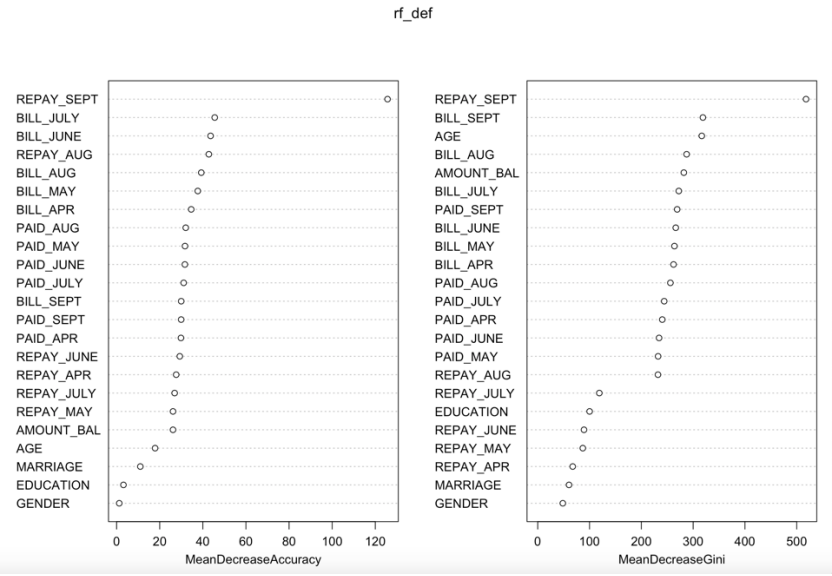


- The bagging model is built using the randomForest package, with mtry = 8 and variable importance enabled.
- The model's prediction accuracy is visualized against actual default statuses, providing a clear comparison.
- Performance assessed via a confusion matrix reveals:

Precision: About 63.82%, indicating moderate accuracy in correctly identifying defaulters.
Recall: Lower at 37.22%, suggesting the model misses a significant number of true defaulter cases.
Accuracy: High overall accuracy of 81.69%, showing effectiveness in general classification.

- Despite its strengths in precision and overall accuracy, the model's lower recall points to a need for further refinement, especially in reliably detecting all potential defaulters.

4.4 Random Forest Model



- A Random Forest model (rf_def) is created using randomForest in R, with mtry = 5 and importance tracking.
- Variable importance is evaluated, highlighting:

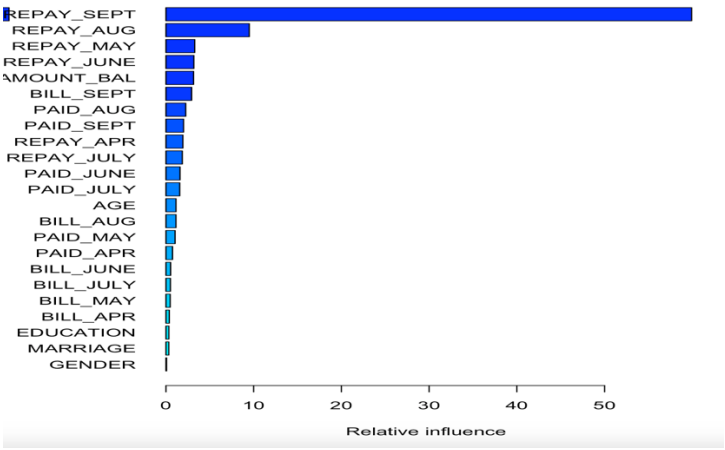
MeanDecreaseAccuracy: Measures the impact on accuracy when a variable is excluded. REPAY_SEPT emerges as a vital predictor, significantly influencing model accuracy.
MeanDecreaseGini: Reflects a variable's contribution to node homogeneity. REPAY_SEPT again stands out, indicating its key role in creating pure nodes.

- Performance metrics from the confusion matrix:

Precision: 64.21%, showing the model's reasonable accuracy in correctly predicting 'Defaulter' instances.
Recall: 37.04%, indicating the model captures just over a third of all actual 'Defaulter' cases.

- The model performs adequately in identifying defaulters but needs improvement in detecting a higher number of true positive cases without increasing false positives.

4.5 Gradient Boosting



- Upgraded to the latest version of Gradient Boosting Model (GBM) using gbm3 package, following issues with the old version.
- Constructed a GBM (gbm_def) for predicting Defaulter status:

Model Configuration: Utilized Bernoulli distribution, 5000 trees (n.trees), and an interaction depth of 4.

- Variable importance analysis from the model highlights:

REPAY_SEPT: Emerges as the most influential predictor, significantly impacting the model's predictions.
Other Factors: REPAY_AUG and REPAY_MAY were also noted as important, albeit to a lesser extent.
Demographic Variables: GENDER, MARRIAGE, and EDUCATION show minimal influence on the model.

- Prediction and performance assessment:
- Applied the model to df_test dataset, with a threshold of 0.5 for binary classification.

The confusion matrix indicates:
Precision: The model accurately predicts 71.23% of Defaulter cases.
Recall: Identifies 27.39% of actual defaulters.
Accuracy: Overall, 81.49% of predictions are correct.

- The GBM model, with its focus on repayment variables, demonstrates effectiveness in default prediction, albeit with room for improvement in recall.

4.5.1 Gradient Boosting with Tuning Learning Rate

A revised Gradient Boosting Model (gbm_def2) was tuned with a learning rate (shrinkage) of 0.2, maintaining other parameters constant. Its performance on df_test data, with a 0.5 threshold, yielded a confusion matrix indicating the modified model's precision, recall, and accuracy rates, providing insight into its predictive effectiveness.

5. MODEL EVALUATION

Model Evaluations and Key Results:

Gradient Boosting Model (With tuned learning rate) (GBM2):

- Accuracy: 79.8%
- Precision: 57.56%
- Recall: 33.06%

Original Gradient Boosting Model (GBM):

- Higher accuracy and precision
- Lower recall than GBM2

Random Forest Model:

- Accuracy: 81.51%
- Precision: 64.21%
- Recall: 37.04%

Bagging Model:

- Comparable accuracy and precision to Random Forest
- Highest recall at 37.16%

Decision Tree Model:

- Best precision rate
- Lower recall compared to other models

F1 Score Comparison:

- Bagging model leads with an F1 score of 0.4702

Top Performing Model:

Bagging model excels in both the highest recall and F1 score, showcasing its superior balance in precision and recall.

Conclusion:

The Bagging model, with its outstanding recall and F1 score, stands out as the most effective choice for predicting credit defaults, crucial for scenarios demanding high accuracy in detecting defaulters, such as in fraud detection.