Predictive Modeling for Personal Loan Acceptance in Banking

Shubham Singh
dept. of Data Science
University of Massachusetts Dartmouth
North Dartmouth, MA
ssingh21@umassd.edu

***Abstract*—The purpose of this study is to determine how well logistic regression and decision tree classification models predict liability customers' propensity to apply for "Thera" Bank personal loans. The dataset includes 5000 consumers' banking ties and demographic data, as well as their response to a prior personal loan campaign. Data splits of 80/20 and 50/50 are used to train and test the models. Models for categorization using decision trees and logistic regression are used and contrasted. Specifically, in the context of binary classification problems, specificity and sensitivity are used to demonstrate the performance metrics used to assess the effectiveness of classification models. The AUC-ROC curve show the TPR for the sensitivity and TNR for specificity for the interpretation of the metric.**

Index Terms*—*"Personal Loan"*; *"Income"*; *"CCAvg"*; *"Securities Account"*; *"Data Mining"*; *"Classification Techniques" "Logistic Regression" "Decision Tree-Based Classification"*; *"Sensitivity"*; *"Specificity"*; *"ROC-AUC".

I. Introduction

This report explores the domain of data-driven insights in the banking sector, with a particular emphasis on the effective identification of prospective customers who are inclined to accept offers of personal loans. The criteria for the analysis of personal loans are categorised in two models. The first one is Logistic Regression which is used to build the model prediction to categorize consumers according to how likely they are to accept offers for personal loans[1]. The approach with both 80/20 and 50/50 splits in the dataset enables thorough examination of its predictive capabilities, offering insights into its interpretability and performance in predicting customer behavior regarding personal loan acceptance. The second method implemented is CART (Classification and Regression Trees) algorithm for building a predictive model. Though an empirical exploration of CART methodology, executed the three approach i.e. Cross Validation, Pruning, and Evaluation Metrics.

To classify the dataset using the logistic regression technique for classifying customers based on their likelihood of accepting personal loan offers applied three methodologies:

Model Evaluation: A variety of indicators, including accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), are used to assess the performance of the model[2].

Model Comparison: Tests of significance and log likelihood are used to evaluate the goodness-of-fit of various models.

Visualization: Visualization techniques such as S-curve plots are employed to visualize the relationship between predictor variables and the probability of accepting personal loans.

Later, to execute the CART for the classification of the customer's personal loan analysis, here implemented the three methodologies in CART as well:

Cross Validation: Implemented cross-validation to assess the model's performance and generalization ability.

Pruning: Used a pruning strategy to keep decision trees from overfitting. In order to ensure that the final decision tree is both comprehensible and successful in capturing the underlying patterns in the data, pruning helps achieve a balance between model complexity and forecast accuracy.

Evaluation Metrics: Used the misclassification rate, or classification error, as a criterion for evaluation. Evaluation metrics including accuracy, precision, recall, F1-score, and ROC-AUC are used to provide a thorough analysis of the model's performance in relation to various splits and validation techniques.

II. BACKGROUND

Data mining techniques offer a powerful solution to this challenge. Data mining allows banks to extract important insights and patterns that guide decision-making processes from massive amounts of consumer data, including demographics, financial behaviors, and historical transactions with the bank. Specifically, predictive modeling methods like decision tree-based categorization and logistic regression have shown to be useful instruments for assessing consumer behavior and forecasting loan approval.

A. Algorithms and Techniques Used

The purpose of this paper is to show how well decision tree-based classification models and logistic regression can predict consumer behavior with regard to personal loan acceptance. To train and assess our models, we use a dataset comprising customer demographic data, banking relationships, and loan campaign answers [3]. Through the use of these data mining tools, we hope to give banks useful information that will help them make wise decisions, improve their relationships with customers, and succeed commercially in the cutthroat and fast-paced financial sector.

1. Logistic Regression: One popular statistical method for binary classification problems is logistic regression. Logistic regression predicts the probability of a binary outcome (acceptance or rejection of a personal loan offer) based on a collection of predictor variables in the context of forecasting customer behavior regarding personal loan acceptance. Logistic regression determines the chance of loan acceptance by calculating the coefficients for every predictor variable. This approach is especially helpful for examining the correlation between the likelihood of accepting a personal loan and client factors such as age, income, and education.

In order to examine the possibility of clients taking personal loans, we used logistic regression as a predictive modeling tool in our investigation. A popular statistical technique for binary classification tasks in which the response variable is categorical and has two possible values is logistic regression. The response variable in our scenario, "Personal.Loan," shows whether a consumer accepted (1) or declined (0) a personal loan offer.

50/50 Split:

For the purpose of training and testing the logistic regression model, we divided the dataset 50/50. The logistic regression model was fitted using the training dataset, and its performance was assessed using the testing dataset.

The goal of the logistic regression model is to calculate the probability of loan acceptance log-odds. The projected probability are then obtained by applying the logistic function to the log-odds. The formula for logistic regression is as follows:

$$\textbf{logit}(p) = \beta_0 + \beta_1 \times \textbf{Age} + \beta_2 \times \textbf{Experience} + \ldots + \beta_n \times \textbf{CreditCard} + \epsilon$$

Where:

logit(p) represents the log-odds of the probability p of loan acceptance, $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients of the predictor variables, and ϵ is the error term.

We calculated the p-values for each predictor variable in our analysis to determine its significance. The final logistic regression model comprised predictor variables that were deemed statistically significant, with p-values of less than 0.05. Significant predictors of loan acceptance included income, family size (particularly, Family 3 and Family 4), CCAvg, education level (Education 2 and Education 3), owning a credit card, using online banking, and having a CD account.

Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC) for 50/50 Split:

The trade-off between sensitivity and specificity across various thresholds for binary classification models is represented graphically by the ROC curve. At different threshold values, it shows the true positive rate (sensitivity) against the false positive rate (1 - specificity). By calculating the area under the ROC curve, the AUC measures the model's overall performance. Better discrimination ability is indicated by a higher AUC value; a value of 0.5 denotes random guessing and a value of 1 denotes flawless categorization.

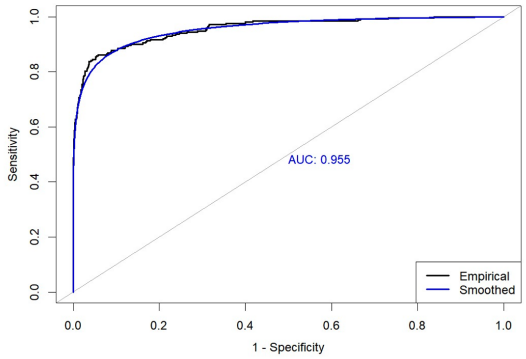

Fig. 1. ROC plot for 50/50 split

80/20 Split:

We must first comprehend the importance and workings of logistic regression in predictive modeling before we can fully discuss the logistic regression model with an 80/20 split for your report. A statistical technique for modeling binary outcomes is called logistic regression. The customer's decision to accept or reject a personal loan is the binary result in your scenario.

The logistic regression equation models the log-odds of the probability of a customer accepting a personal loan. It's given by:

$$\textbf{logit}(p) = \beta_0 + \beta_1 \times \textbf{Age} + \beta_2 \times \textbf{Experience} + \ldots + \beta_n \times \textbf{CreditCard} + \epsilon$$

Here, p represents the probability of accepting a personal loan, $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients of the predictor variables (such as age, income, education, etc.), and ϵ is the error term.

Using R's sample() function, the dataset is initially divided into training and testing sets at random. The training set is used to train the logistic regression model after the data has been separated. The model incorporates predictor variables, like age, income, education, family size, credit card usage, and so on, to forecast the probability of a consumer accepting a personal loan.

Accuracy, sensitivity, specificity, negative predictive value (NPV), positive predictive value (PPV), and true positives and true negatives are just a few of the performance metrics that can be computed using the information obtained from the confusion matrix after the model has been trained.

The given code trains the logistic regression model with the help of R's glm() function, passing "binomial" as the family parameter because the outcome variable is binary. The logistic regression model's summary, which includes the coefficients, standard errors, z-values, and p-values for each predictor variable, is then obtained using the summary()

function.

Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC) for 80/20 Split:

The Receiver Operating Characteristic - Area Under the Curve, or ROC-AUC curve, is a graphic depiction that is frequently used to assess how well binary classification models—like logistic regression—perform. The ROC-AUC curve gives you important information about how well the 80/20 split logistic regression model can distinguish between the positive and negative classes, or loan acceptance and rejection.

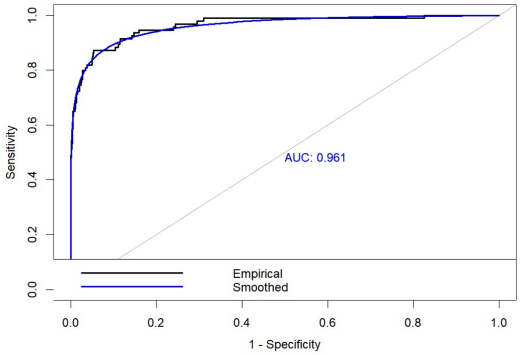

Fig. 2. ROC plot for 80/20 split

Overall Result

	80%	20%	50%	50%
Metric	Training	Testing	Training	Testing
Accuracy	0.96	0.96	0.9648	0.9576
Sensitivity	0.9889319	0.9922737	0.9903169	0.9893238
Specificity	0.6891192	0.6489362	0.7105263	0.6746032
PPV	0.9675149	0.9645923	0.9715026	0.9644406
NPV	0.869281	0.8970588	0.8804348	0.8762887
AUC	0.9636	0.9631	0.9676	0.956

Fig. 3. Summary Table

2. Decision Tree-Based Classification: A data mining method called decision tree-based classification builds a tree-like structure to repeatedly divide the data into subsets according to the values of predictor variables. While each leaf node of the tree represents a class label (accept or refuse) or a probability distribution, each interior node of the tree reflects a judgment based on a particular attribute. Decision trees are useful because they can handle both numerical and categorical data and are easy to interpret. We utilize the Classification and Regression Trees (CART) algorithm in our analysis to create decision tree models that forecast client behavior with relation to approval of personal loans.

The decision tree algorithm minimizes impurity inside each of the resultant nodes by choosing the best feature and split point. Measures of impurity, such entropy or Gini impurity, quantify the disorder or ambiguity of class labels inside a node. At each stage, the split that minimizes impurity is selected.

50/50 Split:

Decision tree-based classification is used to build a decision tree model in the provided case of a 50/50 split, where the dataset is split into training and testing subsets. Classification error metrics, such as accuracy, are calculated for both the training and testing datasets in order to assess the decision tree model's performance. These metrics show how well the model classifies loan acceptance results properly and how well it generalizes to new data.

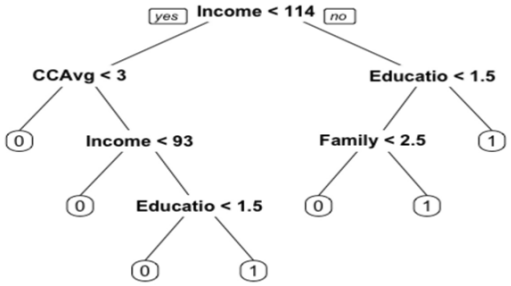

Fig. 4. Decision Tree

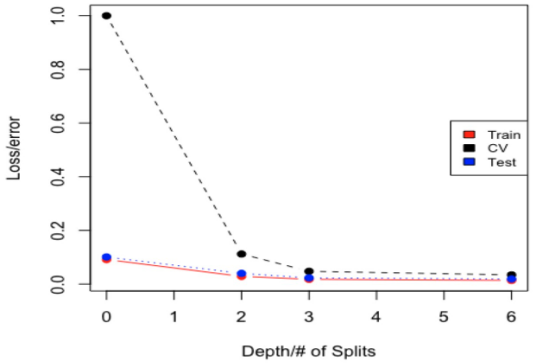

Fig. 5. Validation Plot

In order to balance model complexity and prediction accuracy, the ideal degree of tree pruning is also determined by analyzing the complexity parameter (CP) table. The best appropriate model may be chosen thanks to the information about the tree's structure that the CP table provides at various complexity levels.

A training set (train.data) and a testing set (test.data) comprise the two sections of the dataset. In order to assist determine the model's capacity for generalization, this is done to train the model on one subset of the data and test its performance on another independent subset. In the end, the percentage of correctly identified examples relative to the total number of instances is used to determine the accuracy of the model. This offers a solitary metric for evaluating the model's overall performance.

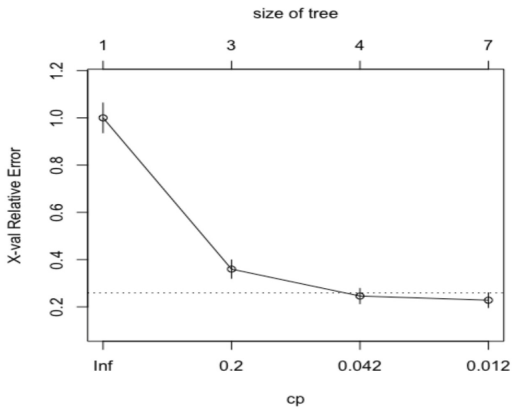

Fig. 6. Complexity Pruning Plot

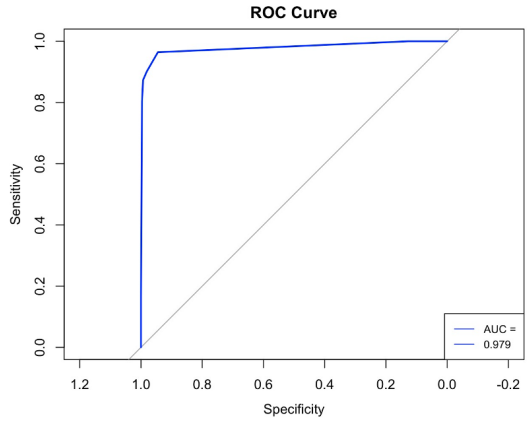

Fig. 7. ROC-AUC Curve

To summarise, the utilisation of the CART algorithm in decision tree-based categorization offers a clear and comprehensible structure for forecasting loan approval. Through a process of recursive feature space partitioning and tree pruning, the model is able to accurately forecast whether a loan will be approved or rejected and effectively capture the underlying patterns in the data and it is shown on the performance metrics.

80/20 Split:

An 80/20 split of the dataset is used to separate the training and testing sets in order to evaluate the predictive performance of the CART model. Eighty percent of the data is utilized as the training set, which is used to train the model, and the remaining twenty percent is used as an independent test set for model validation. This division aids in assessing the model's capacity to generalize to new data.

The CART model construct a decision tree by recursively partitioning the feature space based on the value of predictor variables. The formula in this instance designates "Personal.Loan" as the response variable and incorporates a number of predictors, including credit card usage, age, income, and education. The 'rpart()' function is used to fit the CART

```
> cart_model$cptable
          CP nsplit rel error    xerror       xstd
1 0.34429825      0 1.0000000 1.0000000 0.06313450
2 0.11842105      2 0.3114035 0.3596491 0.03905982
3 0.01461988      3 0.1929825 0.2456140 0.03245187
4 0.01000000      6 0.1491228 0.2280702 0.03129699
```

Fig. 8. Performance Metrics

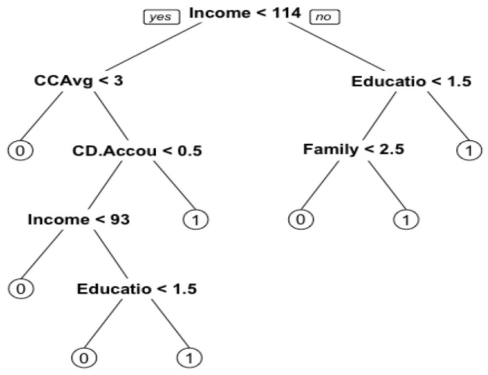

Fig. 9. Decision Tree

model to the training set, with the method set to "class" for classification tasks.

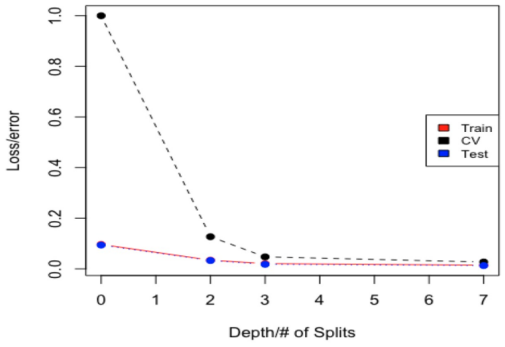

Fig. 10. Validation Plot

Pruning is done to improve generalization and avoid overfitting after building the first CART model. Choosing the ideal complexity parameter to reduce classification errors is the process of pruning.

The absence of a significant increase in xeror despite increasing complexity and high accuracy suggests that the decision tree model is performing well and generalizing effectively to unseen data. This indicates the effectiveness of the model and its potential for real-world application.

A confusion matrix produced from the predictions made on the testing dataset is used to evaluate the predictive performance of the CART model. From the confusion matrix, performance metrics including specificity, sensitivity, and accuracy are computed. While sensitivity and specificity offer

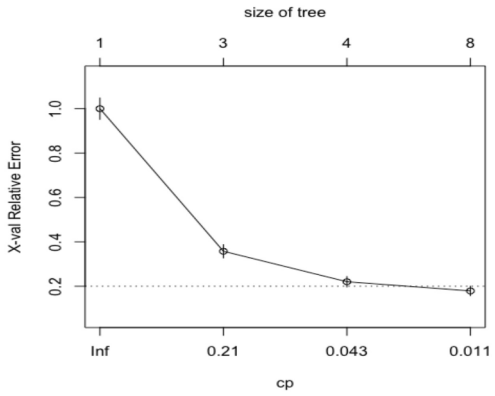

Fig. 11. Complexity Pruning Plot

```
> cart_model$cptable
          CP nsplit rel error    xerror       xstd
1 0.32253886      0 1.0000000 1.0000000 0.04838051
2 0.13989637      2 0.3549223 0.3575130 0.02990394
3 0.01295337      3 0.2150259 0.2202073 0.02362969
4 0.01000000      7 0.1502591 0.1787565 0.02133334
```

Fig. 12. Performance Metrics

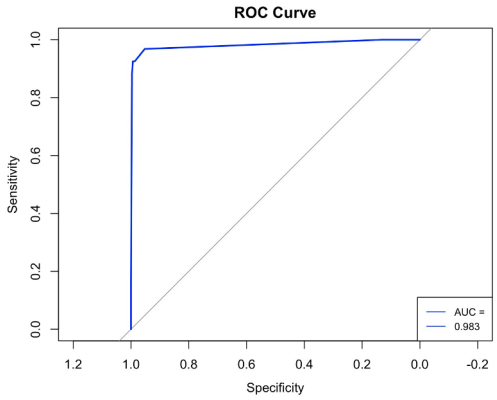

Fig. 13. ROC-AUC Curve

Accuracy	Unpruned (50/50) Split	Unpruned (80/20) Split
Tree Size	6	7
Train	0.9864	0.9855
Test	0.9812	0.987
All (Cross-Validated)	0.9659895	0.9731402

Fig. 14. Summary Table

information about the model's capacity to accurately detect positive and negative examples, respectively, accuracy measures the predictions' overall correctness.

To sum up, the utilization of the CART algorithm to forecast the approval of personal loans yields encouraging outcomes. By utilizing financial and demographic characteristics, the model strikes a compromise between interpretability and prediction accuracy. To improve model performance, more study might look into adding more features or using ensemble methodologies. All things considered, financial organizations looking to streamline their loan approval procedures and better meet the demands of their clients will benefit greatly from the use of the CART algorithm.

III. SIGNIFICANCE

Our project report's contribution to the financial services industry and data-driven decision-making is what makes it significant. Here are some salient points emphasizing its importance.

1. Enhanced Decision Making: Our research intends to give financial organizations a reliable framework for predicting personal loan acceptance by utilizing cutting-edge machine learning techniques such as the CART algorithm. This reduces the risk of default and streamlines the loan approval procedures for banks and lenders by empowering them to make better informed judgments when assessing loan applications.

2. Improved Customer Experience: Accurately predicting the acceptance of personal loans can improve the client experience by expediting the loan application procedure. Banks can provide individualized loan solutions that are suited to each customer's needs by utilizing demographic and financial data, which increases customer happiness and retention.

3. Operational Efficiency: Operational efficiency can be greatly increased by incorporating predictive models into loan approval procedures. Banks can improve overall organizational efficiency by decreasing manual processing times, increasing throughput, and lowering operating costs by automating decision-making based on data-driven insights.

4. Competitive Advantage: Through the application of state-of-the-art machine learning and data mining, our project puts financial institutions at the forefront of innovation. Accurately predicting loan acceptance and providing tailored lending solutions provide banks a competitive edge in the market, drawing in new business and keeping hold of their current clientele.

IV. CHALLENGES

Even though your proposal has the potential to completely transform the approval process for personal loans, there could be a number of difficulties when it is put into practice. Here are a few possible difficulties to think about:

1. Data Quality and Quantity: Both the quantity and quality of training data are critical factors in the efficacy of machine learning models such as CART and Logistic Regression. If the dataset has errors, outliers, or missing numbers, problems could occur. Furthermore, the model's accuracy in capturing intricate patterns and correlations may be restricted by a lack of data.

2. Feature Selection and Engineering: It can be difficult to pinpoint the most important traits (predictors) and efficiently develop them to enhance model performance. To choose

characteristics that actually affect loan approval decisions and to convert unstructured data into useful predictors, domain expertise is needed.

3. Overfitting and Generalisation: When a model learns the training set too well, it becomes overfit and starts to capture noise rather than underlying patterns. Poor generalization performance on untested data may result from this. To reduce overfitting and enhance generalization, regularization strategies and model hyperparameter tuning are crucial.

V. FUTURE WORK

Examine the creation of a hybrid predictive modeling strategy that integrates CART and logistic regression's advantages. The complementary qualities of these models could be utilized by investigating ensemble approaches like stacking or blending. The hybrid model may better capture complicated interactions by fusing the non-linear partitioning of CART with the linear decision bounds of logistic regression.

Use cost-sensitive learning algorithms, ensemble methods built to handle imbalanced data, or sophisticated sampling strategies to address the class imbalance problems present in loan approval datasets. Methods like ADASYN (Adaptive Synthetic Sampling) and SMOTE (Synthetic Minority Oversampling Technique) might assist reduce the effects of class imbalance and increase both models' predictive accuracy, especially for the minority class (approved loans, for example)[4].

VI. VISION

By utilizing cutting-edge machine learning techniques, particularly logistic regression and CART (Classification and Regression Trees), to improve predictive accuracy, transparency, and fairness, your project aims to transform the personal loan approval process. The initiative intends to address major issues facing the financial sector and open the door for a more effective and fair lending ecosystem by utilizing algorithmic modeling and data-driven insights.

Fundamentally, the project imagines a time where advanced prediction models, utilizing the collective intelligence of past data, direct decisions about loan acceptance, thereby evaluating creditworthiness and efficiently reducing risks. Through adoption of a hybrid modeling method that combines the advantages of non-linear partitioning in CART with the linear decision boundaries of logistic regression, the study aims to open up new avenues for precise loan result prediction while preserving interpretability and model transparency.

REFERENCES

[1] Arun, K., G. Ishan, and K. Sanmeet, Loan approval prediction based on machine learning approach. IOSR J. Comput. Eng, 2016. 18(3): p. 18-21.

[2] Fati, S.M., Machine learning-based prediction model for loan status approval. Journal of Hunan University Natural Sciences, 2021. 48(10): p. 1-8.

[3] Chitra, K. and Subashini, B. "Data mining techniques and its applications in banking sector". International Journal of Emerging Technology and Advanced Engineering, 3(8), 219-226, 2013. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.413.9171&rep=rep1&type=pdf [Accessed February 1, 2022]

[4] Pandey, N., R. Gupta, S. Uniyal, and V. Kumar, Loan approval prediction using machine learning algorithms approach. International Journal of Innovative Research in Technology, 2021. 8(1): p. 898-902.