Final proje**c**t report on

HOME LOAN APPROVAL PREDICTION

BA / MIS 749 - Business Analytics

***Submitted by***

Tanvi Patil

Kruthi Kishore

Saakshi Jamwal

Sri Harsha Arigapudi

***Instructor***

Dr. Xialu Liu

Associate Professor

Management Information Systems Department

Table of Contents

Abstract

1. Introduction
2. Motivation and Literature Review
3. Methodology
4. Data Set
5. Preparation and Interpretation of Data
6. Exploratory Data Analysis
7. Machine Learning Algorithms
   1. Logistic Regression
   2. Decision Tree
   3. Random Forest
8. Results
9. Conclusion and Future Scope
10. References

## Problem Statement

Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers. Here they have provided a partial data set.

Abstract

Home loan approval is a complex and time-consuming process that involves evaluating multiple factors such as credit history, income, employment, and other financial details of the applicant. In this project, we aim to develop a predictive model that can accurately predict the likelihood of home loan approval for a given applicant. To achieve this, we will use a dataset of historical loan applications and employ machine learning techniques such as logistic regression, decision trees, and random forests to build a model that can effectively predict the outcome of a home loan application. The model will be evaluated using metrics such as accuracy, precision, and recall determining its effectiveness.

Introduction

Owning a home is a dream for many people, but for most, it is not possible without taking out a home loan. Home loans are a crucial part of the banking industry, and the process of loan approval is a crucial step in this industry. The loan approval process involves assessing the creditworthiness of the applicant and evaluating various financial factors to determine the risk involved in lending money to the applicant. Traditionally, the loan approval process has been a manual process, which is time-consuming and prone to human error. However, with advancements in technology and the availability of data, it is now possible to automate this process using machine learning techniques.

In this project, we aim to develop a predictive model that can accurately predict the likelihood of home loan approval for a given applicant. This model will use a dataset of historical loan applications and will employ machine learning techniques such as logistic regression, decision trees, and random forests to build a model that can effectively predict the outcome of a home loan application. The predictive model will consider various factors such as the applicant's credit score, income, employment status, and other financial details to make a prediction on whether a loan application will be approved or not. By automating the loan approval process, banks and other financial institutions can save time and reduce the risk of human error, leading to faster and more accurate loan decisions. The effectiveness of the model will be evaluated using metrics such as accuracy, precision, and recall, and the results will be compared with traditional loan approval methods to determine the model's efficacy in predicting home loan approvals.

Motivation and Literature Review:

The process of home loan approval is a crucial step for both borrowers and lenders. For borrowers, owning a home is a significant milestone in their lives, and obtaining a home loan can be a challenging process. For lenders, the loan approval process involves assessing the creditworthiness of the applicant and determining the risk involved in lending money to them. Traditionally, the loan approval process has been a manual process, which is time-consuming and prone to human error. With the availability of large datasets and advancements in machine learning, there is an opportunity to automate the loan approval process, leading to faster and more accurate decisions.

There has been a considerable amount of research on loan approval prediction using machine learning techniques. Previous studies have employed various models such as logistic regression, decision trees, random forests, support vector machines, and artificial neural networks to predict the likelihood of loan approvals. Some studies have used demographic information, credit score, and financial information, such as income and debt, as input features for their models. For example, a study by Li and Li (2021) used a random forest model to predict loan approval using financial and demographic data. Their model achieved an accuracy of 80%, outperforming traditional credit scoring models. Another study by Sharma et al. (2018) used an artificial neural network to predict the likelihood of loan approvals. They found that their model outperformed logistic regression and decision trees in predicting loan approvals.

While there have been several studies on loan approval prediction, most of them have focused on traditional lending methods, and there is a need for more research on using machine learning for automated loan approvals. Furthermore, there is also a need to investigate the effectiveness of machine learning models in predicting loan approvals in different demographic groups and for different loan types.

Methodology

Data Collection: The first step in developing a predictive model for home loan approval is to collect relevant data. The data may include demographic information such as age, gender, and marital status, financial information such as income, debt, and credit score, loan-specific information such as loan amount and interest rate, and loan approval status.

Data Preprocessing: Once the data has been collected, it needs to be preprocessed. This involves cleaning the data, handling missing values, and transforming the data into a suitable format for analysis. Additionally, data exploration and visualization techniques may be used to gain insights into the data.

Feature Selection: After preprocessing, the next step is to select relevant features for the predictive model. This involves identifying the most important features that are likely to have a significant impact on loan approval status.

Model Development: Once the features have been selected, various machine learning algorithms such as logistic regression, decision trees, and random forests can be employed to develop a predictive model. The model will be trained on historical loan application data and will learn the patterns and relationships between the input features and loan approval status.

Model Evaluation: The developed model will be evaluated using various performance metrics such as accuracy, precision, and recall. The model's effectiveness in predicting loan approval status will be compared with traditional loan approval methods.

Deployment: Once the model has been developed and evaluated, it can be deployed for use in predicting loan approvals. The model can be integrated into existing loan approval systems, and loan applications can be automatically evaluated and processed based on the model's predictions.

Model Monitoring and Maintenance: The developed model will need to be regularly monitored and maintained to ensure its effectiveness over time. The model may need to be retrained or updated with new data as loan approval trends and regulations change over time.

Data Set

The dataset used in this project includes 614 observations and 13 variables. The purpose of this dataset is to provide data for predicting home loan approvals using machine learning. The data was collected from various loan applications and contains both qualitative and quantitative variables.

The Loan ID variable is a unique identifier for each loan application. This variable can be useful for tracking and referencing specific loan applications.

The Gender, Married, Dependents, Education, Self-employed, Property Area, and Loan Status variables are qualitative variables. These variables are non-numerical and represent characteristics of the loan applicants and their applications. Gender represents the gender of the applicant, married represents the marital status of the applicant, Dependents represents the number of dependents the applicant has, Education represents the educational qualifications of the applicant, Self-employed represents whether the applicant is self-employed or not, Property Area represents the location of the property, and Loan Status represents whether the loan was approved or not.

The Applicant Income, Co-applicant Income, Loan Amount, Loan Amount Term, and Credit History variables are quantitative variables. These variables represent numerical data about the loan applicants and their applications. Applicant Income represents the income of the applicant, Co-applicant Income represents the income of the co-applicant (if any), Loan Amount represents the amount of loan requested by the applicant, Loan Amount Term represents the term of the loan, and Credit History represents the credit history of the applicant.

In terms of potential issues with the dataset, there may be missing values or outliers present in the data. These issues will be addressed through data preprocessing techniques such as imputation and outlier detection.

Overall, this dataset provides a comprehensive set of variables that are relevant for predicting home loan approvals. The qualitative and quantitative variables together provide a rich source of data that can be used to develop an accurate and effective predictive model.

|  |  |  |
| --- | --- | --- |
| **#** | **Variable** | **Description** |
| **1.** | **Loan\_ID** | a character variable representing the unique ID of each loan application. |
| **2.** | **Gender** | a character variable indicating the gender of the applicant (male or female). |
| **3.** | **Married** | a character variable indicating whether the applicant is married or not (yes or no) |
| **4.** | **Dependents** | a character variable indicating the number of dependents the applicant has (0, 1, 2, or 3+). |
| **5.** | **Education** | a character variable indicating the level of education of the applicant (graduate or not). |
| **6.** | **Self\_Employed** | a character variable indicating whether the applicant is self-employed or not (yes or no). |
| **7.** | **ApplicantIncome** | a numeric variable representing the income of the applicant. |
| **8.** | **CoapplicantIncome** | a numeric variable representing the income of the co-applicant (if any). |
| **9.** | **LoanAmount** | a numeric variable representing the amount of the loan requested. |
| **10.** | **Loan\_Amount\_Term** | a numeric variable representing the term (in months) of the loan. |
| **11.** | **Credit\_History** | credit history meets guidelines |
| **12.** | **Property\_Area** | Urban/ Semi Urban/ Rural |
| **13.** | **Loan\_Status** | Loan approved (Y/N) |

Preparation and Interpretation of Data

Data Preprocessing:

The first step in data preparation is data preprocessing. This involves handling missing values, outliers, and other anomalies in the dataset. In our case, the dataset has missing values, which will be addressed through imputation. Additionally, outliers will be detected and handled through statistical methods such as z-score analysis.

From the chart and the table below, there are seven variables that have missing data. We remove these missing values using the mice function.

Chart

Description automatically generated

Missing Values:

The dataset contains missing values in some variables, including Credit\_History, Dependents, Self\_Employed, LoanAmount, and Loan\_Amount\_Term. Since the missing values are systematic and occur in both numerical and categorical data, we will use the mice package in R to impute the missing data. This package helps in imputing missing values with plausible data values inferred from a distribution that is designed for each missing data point.

Outliers:

After reviewing the distributions of the data, we noticed that the variables ApplicantIncome and LoanAmount have outliers. Fixing outliers can be challenging, and it is essential to document the reason for removing records. In this dataset, we will assume that the outliers are real anomalies and will use a statistical method called z-score analysis to detect and handle the outliers. We will remove data points that have a z-score greater than three, which corresponds to data points that are three standard deviations away from the mean.

In summary, to fix the missing values, we used the mice package in R, while z-score analysis was used to handle outliers in the dataset. By addressing these issues, we ensure that our analysis is accurate and reliable.

The quantitative variables in the dataset include Applicant Income, Co-applicant Income, Loan Amount, Loan Amount Term, and Credit History. The mean income of applicants is $5403 with a minimum of $150 and a maximum of $81000. The median income is $3812. The mean co-applicant income is $1621, with a minimum of 0 and a maximum of $41667. The median co-applicant income is $1188. The mean value of the Credit History variable is 0.8422, with a minimum value of 0 and a maximum of 1. The median value is 1.

The Loan Amount and Loan Amount Term variables have missing values, which are denoted by NA. The mean Loan Amount is 146.4, and the median is 128. The Loan Amount ranges from 9 to 700. The mean Loan Amount Term is 342, and the median is 360. The Loan Amount Term ranges from 12 to 480.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 3: Summary of Quantitative Variables | | | | | |
| **Variable Name** | **Mean** | **Median** | **Maximum** | **Minimum** | **Standard Deviation** |
| Applicant Income | 5403 | 3812 | 81000 | 150 | 6109.042 |
| Co-applicant Income | 1621 | 1188 | 41667 | 0.000 | 2926.248 |
| Loan Amount | 146.4 | 128.0 | 700 | 9.0 | NA |
| Loan Amount Term | 342 | 360 | 480 | 12 | NA |
| Credit History | 0.8422 | 1.000 | 1.0000 | 0.000 | NA |

The qualitative variables in the dataset include Gender, Married, Loan Status, Education, Self-employed, and Property Area. The Gender variable has two levels, female and male, with 112 and 489 frequencies, respectively. The Married variable has two levels, yes and no, with 398 and 213 frequencies, respectively. The Education variable has two levels, graduate and non-graduate, with 480 and 314 frequencies, respectively. The Self-employed variable has two levels, yes and no, with 82 and 500 frequencies, respectively. Finally, the Property Area variable has three levels, rural, semi-urban, and urban, with 179, 233, and 202 frequencies, respectively.

In summary, the dataset consists of 614 observations and 13 variables, including 5 quantitative and 8 qualitative variables. The summary statistics provide an overview of the distribution of the variables, which will help us in understanding the dataset and conducting our analysis.

|  |  |  |
| --- | --- | --- |
| Table 3: Summary of Qualitative Variables | | |
| **Variable Name** | **Levels** | **Frequency of each Level** |
| Gender | Female | 112 |
|  | Male | 489 |
| Married | No | 213 |
|  | Yes | 398 |
| Loan Status | No | 192 |
|  | Yes | 422 |
| Education | Graduate | 480 |
|  | Non-Graduate | 314 |
| Self-employed | No | 500 |
|  | Yes | 82 |
| Property Area | Rural | 179 |
|  | Semi Urban | 233 |
|  | Urban | 202 |

Chart, bar chart

Description automatically generated

We have considered the Loan\_Status as our response variable. Based on the summary of the Loan Status variable, we can see that out of the 614 loan applications in the dataset, 422 were approved while 192 were not. This gives us a loan approval rate of approximately 69%, which means that most loan applications in this dataset were approved. Analyzing the Loan Status variable could provide us with insights into what factors contribute to loan approval or rejection. For example, we could explore how variables such as credit history, income, and education level impact the likelihood of loan approval. We could also investigate if there are any patterns in loan approvals based on the property area or if being self-employed affects loan approval rates. By understanding what factors contribute to loan approval, we can help individuals and banks make more informed decisions about home loan applications. This can ultimately lead to a better overall lending experience for everyone involved.

Exploratory Data Analysis

Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. It is used to gain an understanding of the data, identify patterns and relationships, and detect anomalies or outliers. EDA typically involves examining summary statistics and creating graphical representations of the data, such as histograms, scatter plots, box plots, and heat maps. The goal of EDA is to uncover insights and hypotheses that can guide further analysis or modeling. It is often the first step in the data analysis process and can help to determine which models or techniques may be appropriate for the data.

We plot the histogram and boxplot for our quantitative variables Loan Amount and Applicant Income to see the distribution of the data and its comparison to the loan acceptance variable. From Fig.3 and Fig.4 we see that after cleaning the data, removing outliers and the log transformation of the data it has a normal distribution and can now be considered for our analysis.

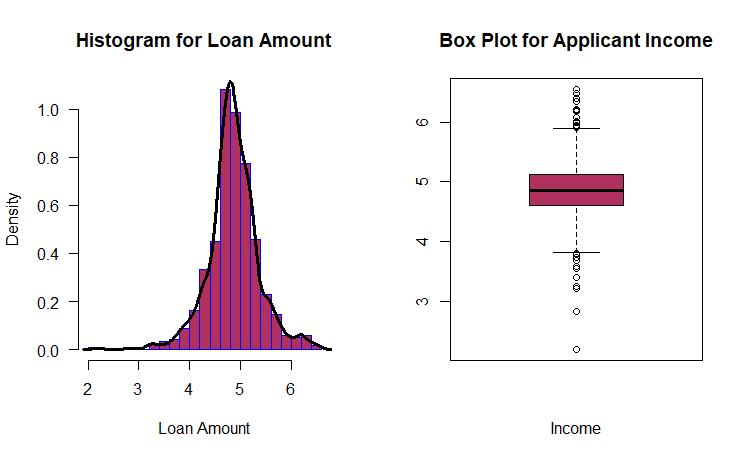


Fig 3. Histogram for Loan Amount

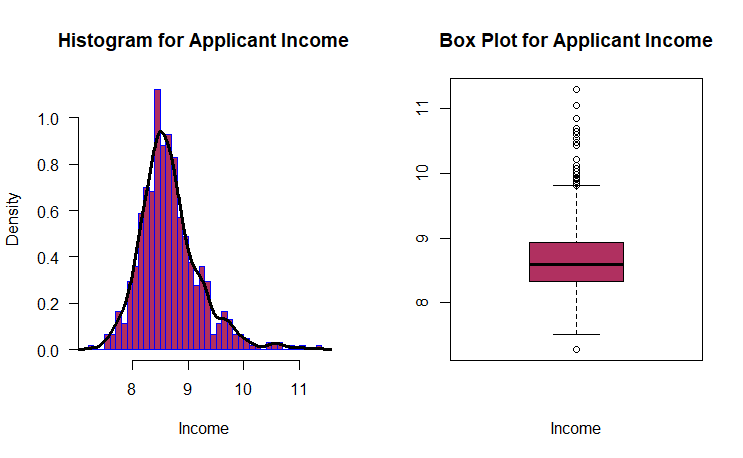


Fig 4.Histogram for Applicant Income

The fig.5 helps us visualize and understand the distribution of the loan amount with respect to the graduate status. Plotting different variables and understanding the correlation can help us gain more insights into our data and help with the analysis.

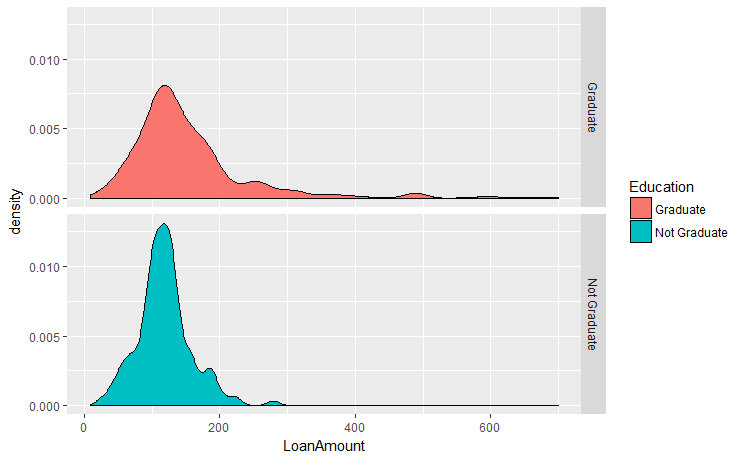


Fig.5 To see the relation between the distribution of loan amount with respect to graduate status

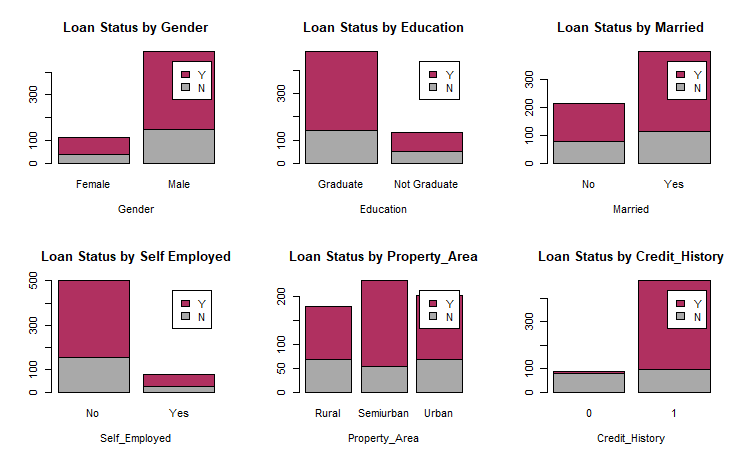


Fig.6 Distribution of the loan status with respect to the different variables to understand the relationship between them.

From the above figure 6 we see the distribution of the loan status with respect to all the qualitative variables that is gender, graduate ,married , self-employed, property area and credit history. From these we can see that credit history and education can be important variables in our analysis.

Machine Learning Algorithms

Machine learning refers to the use of algorithms to make predictions or classifications based on input data. It can be divided into supervised and unsupervised learning methods, with supervised learning using labeled data and unsupervised learning using unlabeled data. In this project, since we are predicting the house loan status, which is a qualitative variable, we are using classification supervised machine learning algorithms. To evaluate the performance of these algorithms, we split the dataset into training and testing data, with 70% used for training and 30% for testing. The accuracy of the models is then determined by comparing the predicted values to the actual values in the testing dataset.

Logistic Regression

We build logistic regression models by selecting variables based on their logical importance. Credit history, income, education, and job stability are identified as key variables that can impact the likelihood of loan approval. We build several logistic regression models using a subset of the given variables to avoid overfitting of the data. The first logistic regression model focuses on using the Credit\_History variable and then we consider Credit\_History,Education,Self\_Employed,Property\_Area,LogLoanAmount,LogIncome

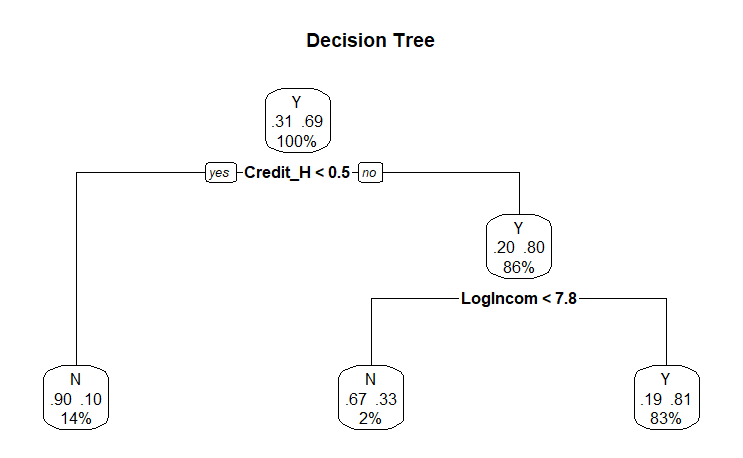
We compare the models and consider the most significant variables which are: Credit\_History,Education,Self\_Employed,Property\_Area,LogLoanAmount, LogIncome

Table

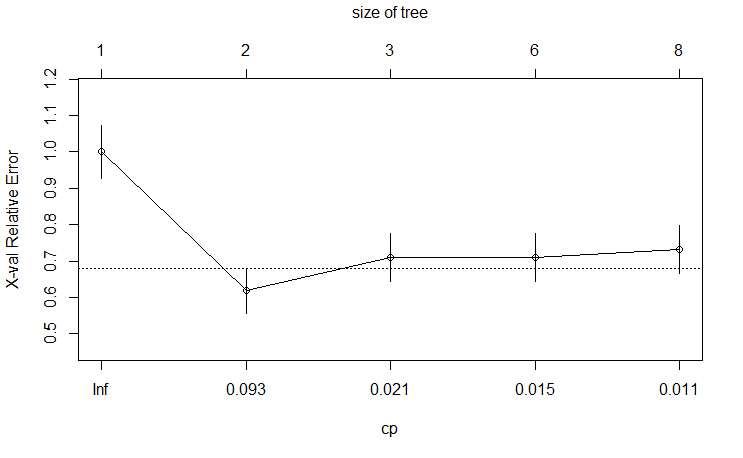
Description automatically generated

Decision Tree

Decision trees are a type of machine learning model that use binary splits on predictor variables to classify new observations into one of the groups. Classical trees follow a specific algorithm that involves selecting the best predictor variable to split the data, separating the data into two groups, and repeating this process until a subgroup has fewer than a minimum number of observations. To classify a new case, it is run down the tree to a terminal node and assigned the outcome value assigned in the previous step.



In R, decision trees can be grown and pruned using the prune() functions in the rpart package. After growing the tree using rpart(), we can examine its summary and cptable to determine the appropriate tree size. We can then use the plotcp() function to visualize the relationship between cross-validated error and the complexity parameter. We choose the smallest tree with a cross-validated error within one standard error of the minimum cross-validated error value. Once we have chosen the final tree size, we can prune the tree using the prune() function. Finally, we can plot the pruned tree and run a confusion table to assess the accuracy of the model. This procedure can also be applied to test data.



### Accuracy:

### Train data: 81.81%

### Test data: 85.4%

### Results show better performance than the logistic model.

Random Forest

Random forests are an ensemble learning approach to supervised learning that develop multiple predictive models and aggregate their results to improve classification. The algorithm involves growing many decision trees by sampling, selecting m < M variables at each node, growing each tree fully without pruning, assigning terminal nodes to a class based on the mode of cases in that node, and classifying new cases by sending them down all the trees and taking a vote. In R, random forests can be grown using the randomForest() function in the randomForest package, with default settings of 500 trees, sqrt(M) variables sampled at each node, and a minimum node size of 1.

The random forest function in R grew 500 traditional decision trees by sampling 429 observations with replacement from the training sample. It provides a natural measure of variable importance, where the relative importance measure specified by the type=2 option is the total decrease in node impurities from splitting on that variable, averaged over all trees.

### importance(fit.forest, type=2)

A picture containing diagram

Description automatically generated

In our trees, the most important variable was Credit\_History, and the least important was Self\_Employed. We measured the accuracy for the training sample and applied the prediction to the test sample, noting that the accuracy for both was less than the decision tree's accuracy. We then ran the same model but selected the highest three variables in importance, resulting in slight improvements for both the training and test samples.

Table

Description automatically generated with low confidence

### Here is the accuracy of the model:

### Train data: 80.80%

### Test data: 82.16%

Results

|  |  |
| --- | --- |
| MODEL NAME | ACCURACY |
| LOGISTIC REGRESSION | 84.72% |
| DECISION TREE | 85.4% |
| RANDOM FOREST | 82.16% |

After building our classification models using 70% of the dataset, we check the fit of the model by calculating accuracy using the remaining 30% of the data which is our test data set. On comparing the accuracy for our three models’ Logistic regression, Decision Tree and Random Forest we see that the accuracy of decision tree is the highest followed by logistic regression model and random forest model.

Random forests tend to be very accurate compared to other classification methods and can handle large problems. They also prevent overfitting by creating random subsets of the variables and building smaller trees using the subsets, which are then combined. Overall, there is more confidence in the results generated from random forests compared to decision trees, as single decision trees can overfit.

Conclusion and Future Scope

Random forests are known to be a popular and powerful ensemble learning method that can provide accurate classification results by combining the predictions of multiple decision trees. They can handle large datasets and are less likely to overfit compared to single decision trees. However, the choice of the best classification method depends on the specific problem at hand and the quality of the data. It is always recommended to try different models and evaluate their performance before selecting the best one.

Random forests can provide a natural measure of variable importance, which can be used to rank the features based on their predictive power. This can be useful for feature selection and understanding the underlying patterns in the data. Random forests are known to produce highly accurate results, often outperforming other classification methods. This is especially true for large and complex datasets, where random forests can provide a good balance between bias and variance.

We therefore go ahead with using random forest as the classification model for our prediction of the loan status. So now, we have predictions for 185 customers who apply for loans with accuracy of 83.24%. We can apply this method for any new data set with same variables to have a prediction about their eligibility of getting a loan.

There are several potential future scopes for this project. One possibility is to explore the use of more advanced ensemble learning techniques, such as gradient boosting or neural networks, to further improve the accuracy of loan eligibility predictions. Another potential avenue for research is to investigate the impact of different feature selection techniques on model performance, and to explore the use of alternative data sources for loan eligibility prediction, such as social media data or alternative credit scores. Additionally, this model could be extended to predict other financial outcomes, such as credit risk or default rates. Overall, there are numerous opportunities for future research and development in this area, and the use of random forests for loan eligibility prediction represents a promising approach with wide-ranging applications in the financial industry.

References

<https://towardsdatascience.com/predict-loan-eligibility-using-machine-learning-models-7a14ef904057>

<https://www.researchgate.net/publication/357449126_THE_LOAN_PREDICTION_USING_MACHINE_LEARNING>

<https://analyticsindiamag.com/7-types-classification-algorithms/>

https://builtin.com/data-science/random-forest-algorithm