|  |  |
| --- | --- |
| THE LOAN ELIGIBILITY PREDICTION  **BUAN 6356.004, Fall 2022, Prof. Zhe Zhang** | Team 11  Martin Navarro  Mitaali Dayal  Shukla Shanthakumara  Tanmayee Ashok Dharam  Tzu- Ying Lin |

*Table of Contents*

[Executive Summary 3](#_Toc120910991)

[Research Objective 4](#_Toc120910992)

[Method of Data Collection 5](#_Toc120910993)

[Data Collection and Exploration 5](#_Toc120910994)

[Data Cleaning 9](#_Toc120910995)

[After Treatment 11](#_Toc120910996)

[Correlation between the variables 11](#_Toc120910997)

[Findings 13](#_Toc120910998)

[Random Forest Tree and Boosted Tree 14](#_Toc120910999)

[Importance of the Attributes for Decision Making 15](#_Toc120911000)

[Inference 15](#_Toc120911001)

[Logistic Regression Model 16](#_Toc120911002)

[Inferences 18](#_Toc120911003)

[Decision Tree Model 19](#_Toc120911004)

[Inference 21](#_Toc120911005)

[Neural Network 22](#_Toc120911006)

[Inference 22](#_Toc120911007)

[Eligible Loan Amount 23](#_Toc120911008)

[Inference 23](#_Toc120911009)

[Conclusion 24](#_Toc120911010)

[Recommendations 24](#_Toc120911011)

[Appendices 25](#_Toc120911012)

# 

# Executive Summary

In today's rapid world of finance, financial institutions such as banks want to automate the loan eligibility process (in real-time) to increase efficiency and reduce time. They can perform such automation based on customer information like income, age, experience, previous loans, loan amount, and loan period, which can be obtained in an online application form. Since the number of transactions in the banking sector is rapidly increasing and massive amounts of data are available, customer behavior can be easily analyzed, and bad loan risks can be reduced. As a result, it is critical to predict loan application approval in real-time.

In this project, we have built and compared four classifier models. Then the best fit model for the given data set will be analyzed and used for predictions. The four models considered in this study are Random Forest Tree, Logistic Regression, Decision Tree, and Neural Network classifiers in R.

# 

# Research Objective

This project's aim is to check the customer's eligibility for receiving a loan using the inputs they have filled in on the loan application form online. If the customer is not eligible for the loan, check the amount of loan they might qualify for after rejection.

# 

# Method of Data Collection

The primary objective of this project is to derive patterns from the datasets used for the loan sanctioning process and create a model based on the patterns derived from previous steps. Models implemented for loan predictions are Random Forest Tree, Logistic Regression, Decision/Classification Tree, and Neural Network. The following section deals with a brief description of loan prediction models used in the modeling and analysis. We used the training set to build the model.

Determine the accuracy

Applying suitable model for Prediction

Data Cleaning and Preprocessing

Split the data into Training & Testing

Loading the loan Dataset

Figure – Flow of Process

## Data Collection and Exploration

The dataset we are using for analysis of loan eligibility is second-hand data. Our dataset has 13 variables. For this project, we will analyze and implement 12 out of 13 variables in the data mining process. The variables are listed below:

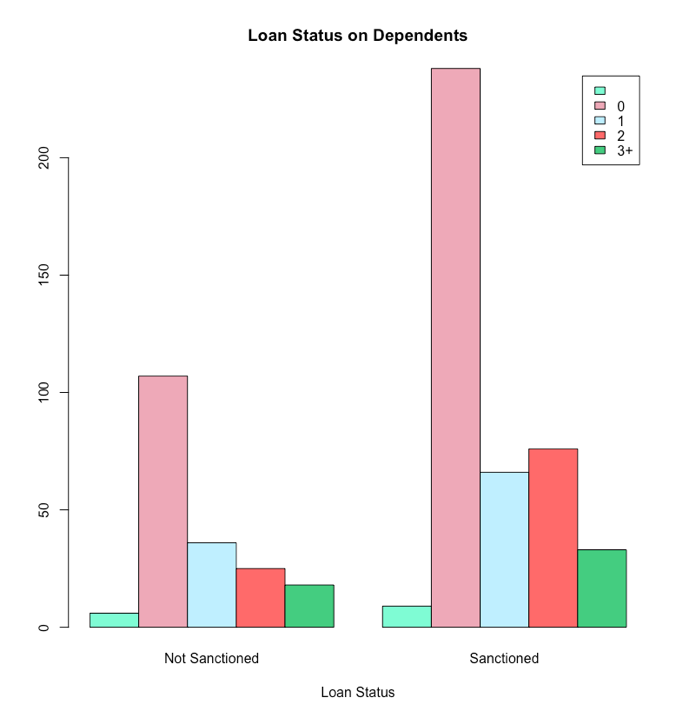
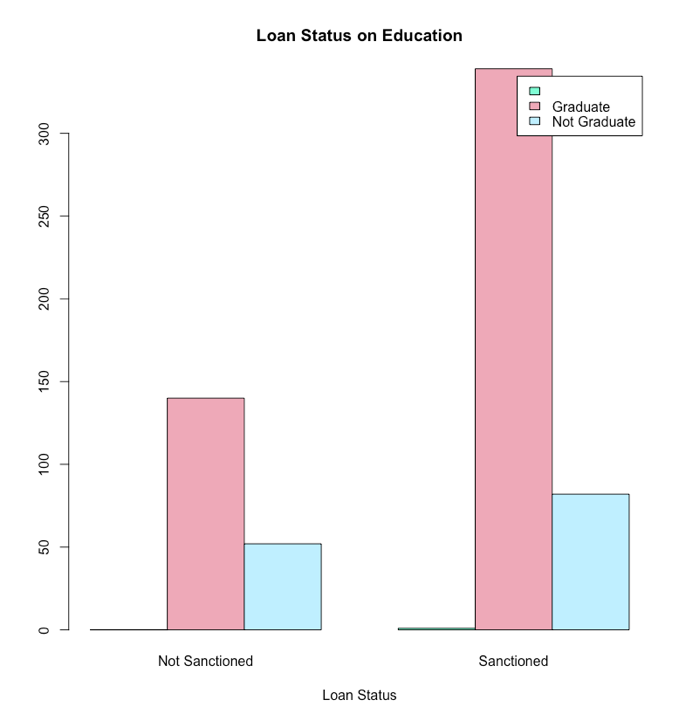
* Gender
* Married
* Dependents
* Education
* Self-Employed
* Applicant income
* Co-applicant income
* Loan amount
* Term of loan amount
* Credit history
* Property area
* Loan Status
* Loan ID (Indentation)

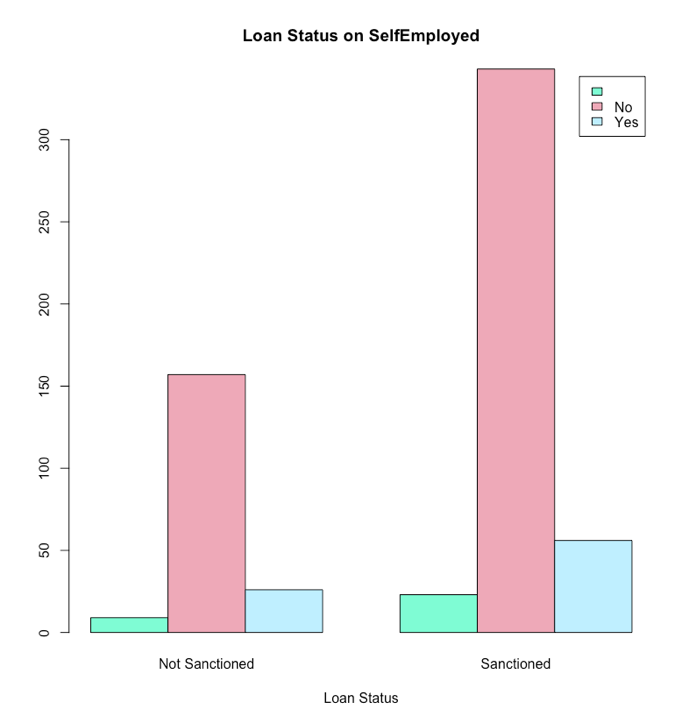
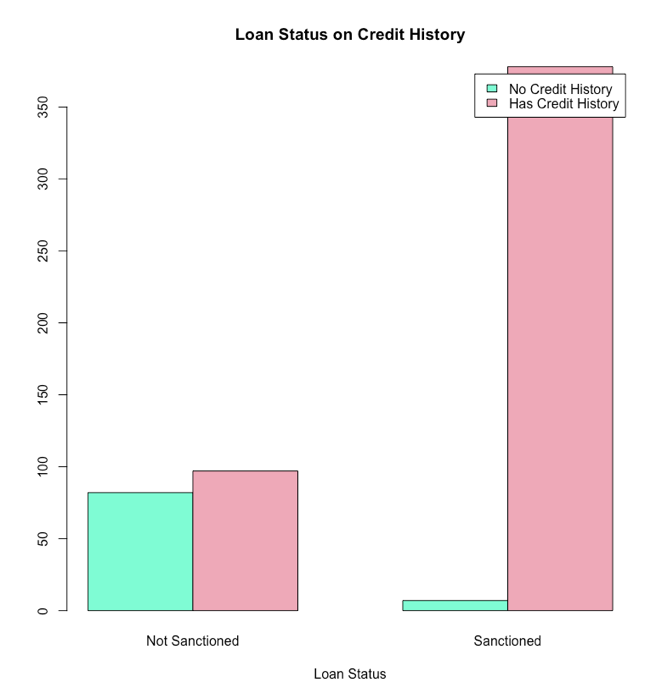
Source: Kaggle

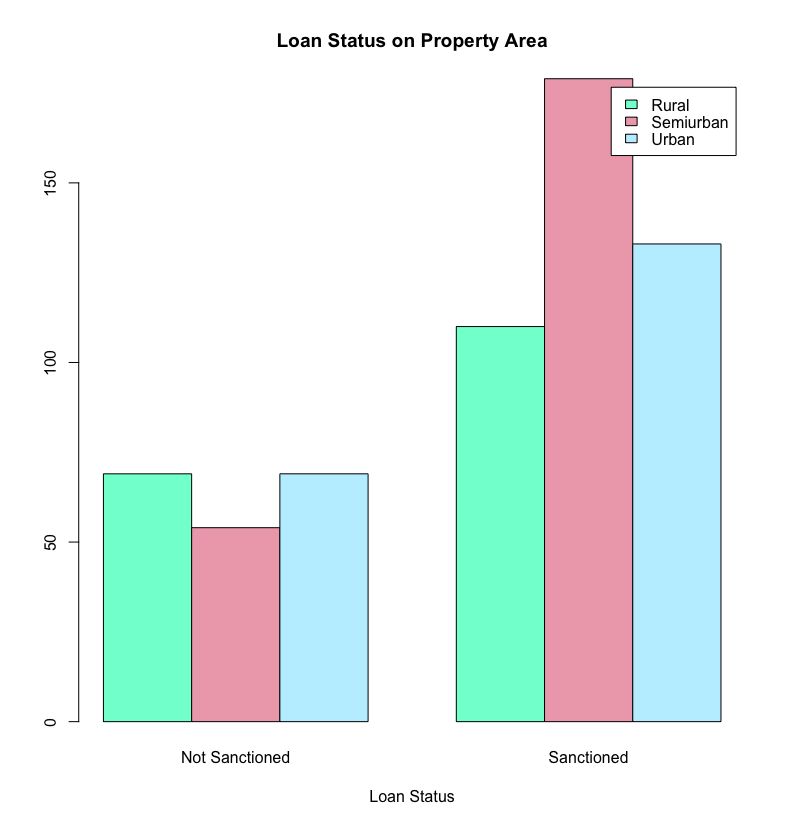
Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

 Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

It can be inferred from the above bar plots that is observed in our data in terms of sanction and non-sanction.

* In training dataset, 80% of loan applicants are male, whose loans are approved.
* Nearly 70% applicants are married
* Almost 50% of the applicants have no dependents.
* Around 80% of applicants are graduates.
* Nearly 85-90% of Loan applicants are self-employed.
* Loan has been approved for more than 98% of those who have credit history.
* The highest number of applicants are from semi-Urban areas, followed by urban areas.

From above graphs, we can notice that high count of loan approval happened for:

* who are male
* who are married
* who has no dependents
* who are graduated
* who are not self employed

## Data Cleaning

Summary of the dataset is shown in the following image. As you can observe, there are quite a few missing values in the dataset.

A screenshot of a computer

Description automatically generated with medium confidence

Since our dataset is small and each missing value is present in separate rows. Instead of deleting the entries, we are replacing the missing values in the dataset. Depending on the Predictors and its characteristics, we are replacing the missing values with mode, mean, or min value. The decision and assumptions are present in the following table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Attribute** | **Datatype** | **Using in Analysis** | **Missing Substitute?** | **Reason/Assumptions** |
| **Loan ID** | Character | No |  |  |
| **Gender** | Character | Yes | Maximum Value | We are assuming if the gender is missing, we are substituting with the maximum of applicant gender. |
| **Married** | Character | Yes | No | If marriage status is not mentioned, we are assuming the customer is not married. |
| **Dependents** | Character | Yes | 0 | If more dependents, less money. If dependents are not mentioned. We are assuming applicants have no dependents. |
| **Education** | Character | Yes | If Education is not mentioned, we are assuming the person is not a graduate. | The higher the education, the higher the chances that they are paid more money hence the higher applicant income. |
| **Self-employed** | Character | Yes | Yes |  |
| **Income** | Numeric | Yes | Mean |  |
| **Co-applicant Income** | Numeric | Yes | Minimum - 0 | If co-applicant income is not mentioned, we are considering that the co-applicant has no income. |
| **Loan Amount** | Numeric | Yes | Median |  |
| **Loan Term** | Numeric | Yes | Median |  |
| **Credit History** | Numeric | Yes | 0 | If credit history is not mentioned, we assume the person has no credit history. It means it is not available |
| **Property Area** | Character | Yes |  | It has no missing values |

## After Treatment

Text, letter

Description automatically generated

## Correlation between the variables

In datasets with many variables, there is usually much overlap in the information covered by the set of variables. A straightforward way to find redundancies is to look at a correlation matrix. This shows all the pairwise correlations between variables. Pairs with a solid (positive or negative) correlation contain a lot of overlap in information and are good candidates for data reduction by removing one of the variables. We wanted to check if there are any attributes for our datasets. We especially wanted to check the correlation between applicant income and co-applicant income to see if we could combine them into one column.

**Table

Description automatically generated**

**Chart

Description automatically generated**

Looking at the table and the graph, we can see that all the attributes are independent and can be used individually in our data mining models. The highest correlation is between the loan amount and applicant income, which is expected.

# Findings

Datamining techniques that are used for this project are:

1. Random Forest Tree
2. Logistic Regression
3. Decision Tree
4. Neural Network

All the R codes used in the model creation and analysis are present in the BI.004.11.Final.TheLoanEligibilityProject.Codes.R file.

## Random Forest Tree and Boosted Tree

Random forest is a commonly used machine learning algorithm which combines the output of multiple decision trees to reach a single result. The accuracy of this model is usually higher.

|  |  |
| --- | --- |
| The accuracy of the model using confusion matrix in parallel mode is 73.51%. | Table  Description automatically generated |
| The accuracy of the boosted model is 71.35%. | A picture containing text  Description automatically generated |

### Importance of the Attributes for Decision Making

To know the importance of variables in making loan status decisions, we ran a random forest analysis with all 12 variables. Created variable importance plot. The plot is as shown below:

Graphical user interface, application, table, Excel

Description automatically generated

### Inference

The accuracy of the random forest tree model is 73.51%. When we draw the variable importance plot for the model, we can see that credit history, co-applicant income, applicant income are the major variables. We also realized that property area also plays a major role in loan sanction decision.

## Logistic Regression Model

It is a type of predictive analysis model used to describe data and understand the relationship between binary dependent variable and independent variable and predict probability of “Y”.

The following image is the table summary of the logistic regression. There are three variables that show significant differences in the p-value: Married, Credit History and Property Area. Therefore, we concluded that from the logistic regression analysis, Married, Credit History and Property Area are significant predictors of the eligibility of the loan status (p< .05).

|  |  |
| --- | --- |
| **Our full model summary** |  |
| **Our full model accuracy** |  |
| **Model selection 1.**  **Loan status depends on only Credit History + Property Area + Married** |  |
| **Selected model 1 accuracy.** |  |
| **Checking relationship between the variables.** |  |

Inferences

Loan sanction status depends on credit history, property area, marriage status. When we ran model to consider the interaction between the variables, though the realtionhisp is not strong under 95% CI. Around 90% CI, we can see following relations playing major role, in loan sanction status

* If the person is not married and living in a semiurban or urban area.
* If the person has credit history and not married

## Decision Tree Model

To determine the customer's eligibility to receive a loan, we have created a decision tree using all the attributes but the loan id. The decision tree is a technique for classification/prediction based on nested tests.

The decision tree below was performed by replacing any missing values with the data cleaning process mentioned above. The reason for cleaning the data is that the data in the decision tree without replacing any missing values are noisy and overfitting. The decision tree is expressed in the image below.

Diagram

Description automatically generated

|  |  |
| --- | --- |
| Training model using the training dataset has accuracy of 80.65%. | A picture containing table  Description automatically generated |
| Training model behavior using the unknown dataset has accuracy of 71.89%. |  |
| Deeper tree accuracy for unknown dataset is 62.16% |  |
| After pruning the tree, the accuracy is 72.43% |  |

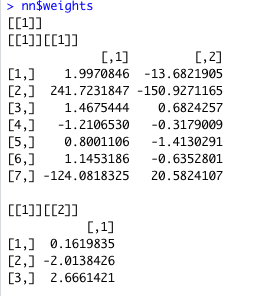
### Inference

From the decision tree, credit history is the root node, and loan amount and applicant income are the most relevant attributes to determine the customer's eligibility to receive a loan. The model developed using these inputs has accuracy level of 72.43% which is par with the forest tree and boosting tree.

## **Neural Network**

The following diagram shows the neural network model. We used only the numeral variables, and 2 hidden layers to construct the neural network model.

Chart, radar chart

Description automatically generated Table

Description automatically generated with medium confidence

### Inference

The model has an accuracy of **69.19%** and a P-value = **0.5357.**

## Eligible Loan Amount

To calculate the loan amount that can be sanctioned for the loan applicant, we created a subset out our dataset where loan status is yes. This dataset information will give a general idea about how much loan got sanctioned for customers with diverse backgrounds. Using this information, we created a regression model for loan amounts.

Table

Description automatically generated with medium confidence

### Inference

From the summary we can recognize that the ApplicantIncome, CoapplicantIncome, LoanAmountTerm has good linear relationship with loan amount.

We can use this information to predict the loan amount a customer is eligible.

LoanAmount = 85.9 + 0.0091 ApplicantIncome + 0.0087CoapplicantIncome + 0.15LoanAmountTerm.

# Conclusion

After analyzing the customer's eligibility for receiving a loan from the models created, it is observed that Random Forest and Decision Tree predict the eligibility range highly and correctly. In contrast, other models, such as Neural Network and Logistic Regression models, predict less eligibility. The decision Tree model gave the highest accuracy with 72.43%. Since we had a deficient number of observations and this is the highest accuracy, we could get out of all the models.

# Recommendations

There are two ways we can implement this model.

We can implement this model on the bank’s website and allow applicants to check their eligibility and the loan amount they might qualify for after applying. Doing so will enable applicants to improve their profile if the loan status is negative. They can also apply for a lesser loan amount or improve their credit scores before applying. The more the applicant is aware of their situation and improves their profile, the more business for banks.

Banks can also make use of this model. When the loan application is sent, they can run the model and check to see their eligibility. If the loan status is No, it shows that lending a loan to the customer can be risky. Depending on the customer profile, the bank can do more background research into the applicant’s profile to make a further decision or reject the applicant. It saves the bank a lot of money invested in background verification. The model also helps the employee save time getting invested in a bad profile, thus indirectly driving more revenue and profit.

# Appendices

https://www.kaggle.com/datasets/yashpaloswal/loan-prediction-with-3-problem-statement?datasetId=2453831&sortBy=dateRun&tab=profile&select=testing\_set.csv