TAYLOR'S UNIVERSITY
Wisdom • Integrity • Excellence

SCHOOL OF COMPUTER SCIENCE

GROUP ASSIGNMENT (Weightage 30%)
APRIL 2024 SEMESTER

MODULE NAME	**: Statistical Inference and Modelling**
MODULE CODE	**: ITS66804**
DUE DATE	**: Week 9**
PLATFORM	**: MyTIMES**

This paper consists of TEN (10) pages, inclusive of this page.

Group No: "6"

Project Title: Loan Approval Prediction

STUDENT DECLARATION

1. I confirm that I am aware of the University's Regulation Governing Cheating in a University Test and Assignmentand of the guidance issued by the School of Computing and IT concerning plagiarism and proper academic practice, and that the assessed work now submitted is in accordance with this regulation and guidance.
2. I understand that, unless already agreed with the School of Computing and IT, assessed work may not besubmitted that has previously been submitted, either in whole or in part, at this or any other institution.
3. I recognise that should evidence emerge that my work fails to comply with either of the above declarations, then Imay be liable to proceedings under Regulation

No.	Student Name	Student ID	Date	Score
1.	Prabin Joshi	0358667	17th June	
2.	Anuj Bhandari	0358445	17th June	
3.	Sandesh Maharjan	0358810	17th June	
4.	Dawa Tamang	0358448	17th June	
5.	Pooja Thapa	0358808	17th June	

Marking Rubrics

Group Assignment Marking Rubrics					
Abstract (5 marks)	5 marks A clear and concise abstract that gives the reader a clear idea of what the project is about and why it is interesting. The following components need to be included i. Purpose and motivation of this research ii. Problem you are addressing iii. Methods and materials iv. Results v. Conclusion	4 marks A clear abstract that gives the reader a clear idea of what the project is about. Four of the following components are included i. Purpose and motivation of this research ii. Problem you are addressing iii. Methods and materials iv. Results v. Conclusion	The abstract is difficult to read and/or is very vague and/or doesn't sell the project as well as it might have. Three of the following components are included i. Purpose and motivation of this research ii. Problem you are addressing iii. Methods and materials iv. Results v. Conclusion	2 marks Unable to read the abstract and/or is very vague and/or doesn't sell the project as well as it might have. Only two of the following components are included i. Purpose and motivation of this research ii. Problem you are addressing iii. Methods and materials iv. Results v. Conclusion	1 mark Unable to read the abstract. Only one of the following components is included i. Purpose and motivation of this research ii. Problem you are addressing iii. Methods and materials iv. Results v. Conclusion
Introduction (10 marks)	9-10 marks A readable write-up that explains what the problem is and why it is of interest. The following components need to be included i. Problem ii. Negative	7-8 marks A readable write-up that explains what the problem is. Three of the following components are included. i. Problem ii. Negative impact of the problem iii. Parties	5-6 marks The write-up is difficult to read, somewhat vague, or doesn't make a really good case for why the problem is of interest. Two of thefollowing components	3-4 marks Unable to read the write-up and/or is very vague. Only one of the following components are included. i. Problem ii. Negative impact of the problem	1-2 marks Unable to read the write-up. None of the following components are included. i. Problem ii. Negative impact of the problem iii. Parties affected

	impact of the problem iii. Parties affected iv. Benefit of solving the problem	affected iv. Benefit of solving the problem	are included. i. Problem ii. Negative impact of the problem iii. Parties affected iv. Benefit of solving the problem	iii. Parties affected iv. Benefit of solving the problem	iv. Benefit of solving the
Literature Review (20marks)	18-20 marks An outstanding overview, with an insightful analysis of prior work and a clear connection between prior work and the proposed method. The following components are given. i. Introduction of the topic ii. Taxonomy Mapping iii. Paragraphs for each branch of the taxonomy tree iv. Conclusion v. Critical Review	15- 17 marks A comprehensive overview of prior work that gives the reader a clear idea of what's out there and how the proposed method is different. Four of the following components are given. i. Introduction of the topic ii. Taxonomy Mapping iii. Paragraphs for each branch of the taxonomy tree iv. Conclusion v. Critical Review	10-14 marks A fairly good overview of prior work, and some connection is made to the proposed method. Three of the following components are given. i. Introduction of the tropic ii. Taxonomy Mapping iii. Paragraphs for each branch of the taxonomy tree iv. Conclusion v. Critical Review	5-9 marks An overview of several papers related to the proposed method, and some attempt is made to connect the prior work to the current method. Two of the following components are given. i. Introduction of the topic ii. Taxonomy Mapping iii. Paragraphs for each branch of the taxonomy tree iv. Conclusion v. Critical Review	1-4 marks An overview of several related papers, but not within a coherent conceptual frame- work. One of the following components are given. i. Introduction of the topic ii. Taxonomy Mapping iii. Paragraphs for each branch of the taxonomy tree iv. Conclusion v. Critical Review
Data (5marks)	5 marks The data are comprehensive and clearly described. At least 6 of the following	4 marks The data are fairly explained. At least 5 of the following components	3 marks The data are not comprehensive and/or there is a flaw in the	2 marks The explanations are significantly flawed. At least 3 of the	1 mark The explanations are flawed. At least 2 of the following

	components are given. i. Source of the data ii. Description of the data and its context iii. Statistics of the data iv. Presentation, visualization and quantification of the data and images v. Conclusion	are given. i. Source of the data ii. Description of the data and its context iii. Statistics of the data iv. Presentation, visualization and quantification of the data and images v. Conclusion	explanation. At least 4 of the following components are given. i. Source of the data ii. Description of the data and its context iii. Statistics of the data iv. Presentation, visualization and quantification of the data and images v. Conclusion	following components are given. i. Source of the data ii. Description of the data and its context iii. Statistics of the data iv. Presentation, visualization and quantification of the data and images v. Conclusion	components are given. i. Source of the data ii. Description of the data and its context iii. Statistics of the data iv. Presentation, visualization and quantification of the data and images v. Conclusion
Method (20 marks)	17-20 marks The methods of analysis Are comprehensive and clearly described. At least 6 of the following components are given. i. Explanatory data analysis ii. Statistical data analysis methods iii. Appropriate data analysis iv. Statistical methods address the research objective v. Information on data analysis	13-16 marks The methods of analysis are fairly explained. At least 5 of the following components are given. i. Explanatory data analysis ii. Statistical data analysis methods iii. Appropriate data analysis iv. Statistical methods address the research objective v. Information on data analysis process vi. Clear	9-12 marks The methods of analysis are not comprehensive and/or there is a flaw in the explanation. At least 4 of the following components are given. i. Explanatory data analysis ii. Statistical data analysis methods iii. Appropriate data analysis iv. Statistical methods address the research objective v. Information	5-8 marks The methods of analysis are significantly flawed. At least 3 of the following components are given. i. Explanatory data analysis ii. Statistical data analysis methods iii. Appropriate data analysis iv. Statistical methods address the research objective v.	1-4 marks The methods of analysis are flawed. At least 2 of the following components are given. i. Explanatory data analysis ii. Statistical data analysis methods iii. Appropriate data analysis iv. Statistical methods address the research objective v. Information on data analysis

	process vi. Clear relationship between methods	relationship between methods	on data analysis process vi. Clear relationship between methods	Information on data analysis process vi. Clear relationship between methods	process vi. Clear relationship between methods
Result & Discussion (20 marks)	17-20 marks The results are comprehensive and clearly described. At least 6 of the following components are given. i. Subheadings are included and are clear and informative ii. Figures and tables are supported by text iii. Correct interpretation of the results iv. Results with tables and diagrams v. Additional insight to the content vi. Critical analysis of the results vii. Clearly addresses the research question	13-16 marks The results are fairly explained. At least 5 of the following components are given. i. Subheadings are included and are clear and informative ii. Figures and tables are supported by text iii. Correct interpretation of the results iv. Results with tables and diagrams v. Additional insight to the content vi. Critical analysis of the results vii. Clearly addresses the research question	9-12 marks The results are not comprehensive and/or there is a flaw in the explanation. At least 4 of the following components are given. i. Subheadings are included and are clear and informative ii. Figures and tables are supported by text iii. Correct interpretation of the results iv. Results with tables and diagrams v. Additional insight to the content vi. Critical analysis of the results vii. Clearly addresses the research question	5-8 marks The results are significantly flawed. At least 3 of the following components are given. i. Subheadings are included and are clear and informative ii. Figures and tables are supported by text iii. Correct interpretation of the results iv. Results with tables and diagrams v. Additional insight to the content vi. Critical analysis of the results vii. Clearly addresses the research question	1-4 marks The results are flawed. At least 2 of the following components are given. i. Subheadings are included and are clear and informative ii. Figures and tables are supported by text iii. Correct interpretation of the results iv. Results with tables and diagrams v. Additional insight to the content vi. Critical analysis of the results vii. Clearly addresses the research question
Limitation and future	9-10 marks An insightful and correct	7-8 marks A correct analysis that	5-6 marks An incomplete or	3-4 marks An incorrect analysis. One	1-2 marks No analysis. None of the

Study (10 marks)	analysis. The following components are given. i. Discussion addresses the major finding of the study ii. Results are interpreted with respect to outside sources iii. Identify the limitation or limitations iv. Explain these limitations in detail v. Propose a future direction for future studies	could be more complete and is not very insightful. One of the following components is missing. i. Discussion addresses the major finding of the study ii. Results are interpreted with respect to outside sources iii. Identify the limitation or limitations iv. Explain these limitations in detail v. Propose a future direction for future studies	somewhat incorrect analysis. Two of the following components are missing. i. Discussion addresses the major finding of the study ii. Results are interpreted with respect to outside sources iii. Identify the limitation or limitations iv. Explain these limitations in detail v. Propose a future direction for future studies	of the following components are given. i. Discussion addresses the major finding of the study ii. Results are interpreted with respect to outside sources iii. Identify the limitation or limitations iv. Explain these limitations in detail v. Propose a future direction for future studies	following components are given. i. Discussion addresses the major finding of the study ii. Results are interpreted with respect to outside sources iii. Identify the limitation or limitations iv. Explain these limitations in detail v. Propose a future direction for future studies
Conclusion (5 marks)	5 marks A clear and insightful summary of the paper, perhaps with interesting ideas for future work. The following components are given. i. Restate your research topic ii. Restate the objective iii. Summarize the main topics iv.	4 marks A summary of the experiments is given, but the conclusion is a mere summary. The ideas for future work are not interesting. One of the following components is missing. i. Restate your research topic ii. Restate the	3 marks A flawed conclusion. Two of the following components are missing. i. Restate your research topic ii. Restate the objective iii. Summarize the main topics iv. Significance of results v. Conclude	2 marks An incorrect conclusion. Three of the following components are missing. i. Restate your research topic ii. Restate the objective iii. Summarize the main topics iv. Significance of results	1 marks No conclusion. One of the following components is given. i. Restate your research topic ii. Restate the objective iii. Summarize the main topics iv. Significance of results

	Significance of results v. Conclude the thoughts	objective iii. Summarize the main topics iv. Significance of results v. Conclude the thoughts	the thoughts	v. Conclude the thoughts	v. Conclude the thoughts
Format (5 marks)	5 marks A clear and correct formatting. The following components are given. i. Number of pages 10 -15 ii. Use the correct template iii. Similarity index less than 20% iv. All the sections given in proper order v. Readable pdf file	4 marks A clear and correct formatting. One of the following components is missing. i. Number of pages 10 -15 ii. Use the correct template iii. Similarity index less than 20% iv. All the sections given in proper order v. Readable pdf file	3 marks Two of the following components are missing. i. Number of pages 10 -5 ii. Use the correct template iii. Similarity index less than 20% iv. All the sections given in proper order v. Readable pdf file	2 marks Three of the following components are missing. i. Number of pages 10 -15 ii. Use the correct template iii. Similarity index less than 20% iv. All the sections given in proper order v. Readable pdf file	1 marks One of the following components is given. i. Number of pages 10 -15 ii. Use the correct template iii. Similarity index less than 20% iv. All the sections given in proper order v. Readable pdf file

Table of Content

List of Figures

Abstract

The proposed research examines a dataset with different applicant characteristics to examine how statistical methods may be applied to predict loan approval. By using a variety of statistical and machine learning models, the study seeks to increase the accuracy of loan approval predictions, which is crucial for financial institutions to lower risks and advance fair access to credit. Identifying the critical elements that influence loan acceptance, examining socioeconomic trends, providing assistance to applicants, and developing the best prediction models are the goals of the study. During the data preprocessing stage, imbalances in the data were addressed and missing values were handled. To assess their prediction effectiveness, two main models were used: decision trees and logistic regression. The decision tree model shows higher accuracy as compared to the logistic regression model - training set (***81.81%***) and testing set (***85.4%***). This implies that for loan approval outcomes, this tree algorithm possesses superior generalization as well as prediction abilities. Model accuracy was further increased by augmenting the dataset with characteristics including debt-to-income ratios and applicant's job stability. Important drawbacks were noted, such as possible biases in linear models and overfitting in smaller datasets. In order to improve forecast accuracy and resilience, future research areas recommend investigating alternative models such as neural networks and ensemble approaches. The results highlight the need of thorough data preparation and the promise of cutting-edge machine learning algorithms to enable more trustworthy loan approval choice.

1.0 Introduction

In the financial sector, accurately predicting loan approval results is essential for lenders as well as borrowers. Evaluating a number of factors like income, employment status, credit history, and other personal information is part of the process to find out how likely it is that a loan would be approved. As data becomes more widely available, statistical models have become important tools for boosting the accuracy of these predictions. So, financial organizations can increase customer satisfaction, reduce risks, and enhance their decision-making procedures by using such models.

The significance of resolving this issue goes beyond operational effectiveness. Loan approvals for indexed accounts may result in financial losses for lending institutions because of loan defaults on account that shouldn't have been approved. Furthermore, those who receive approval for loans they cannot afford may experience financial hardship and loss due to their credit score. However, excessively strict regulations could hinder individuals with disabilities from receiving the necessary financial support, which may impede their economic and personal development. Hence, development of trustworthy and accurate loan prediction models is very crucial. It also helps to reduce financial risks and guarantees fairs' access to credit, which promotes stability and economic growth.

2.0 Problem Statement

Developing a predictive model that can reliably forecast the likelihood of loan approval based on many application attributes is the main issue this project aims to address. Features like *Marital Status, Gender, Dependents, Employment Status, Education, Income levels, Loan Amount, Loan Term, Credit History, Property Area, and Loan Status*. These characteristics have a cumulative effect on the decision making process for loan approvals.

There are numerous harmful effects of this problem. Inaccurate loan applications can result in significant financial losses for financial organizations since they may contain errors about loans that should not have been approved. Rejecting a loan application can be a big setback for borrowers, especially if it's because of incomplete or inaccurate assessment criteria. This issue affects multiple parties: lending institutions that bear the financial risks, potential borrowers whose financial needs might not be met, and the broader economy that relies on efficient credit distribution for growth. Solving this problem benefits all these parties by improving the accuracy and fairness of loan approvals.

3.0 Objectives of the Study

1) To identify the factors influencing loan approval processes and make appropriate adjustments to loan processing procedures.
2) To examine trends and correlations among socioeconomic, demographic, demographic variables, and loan approval outcomes in order to facilitate personalized recommendations for applicants.
3) To provide target audience support strategies through loan approval analysis, assisting applicants in becoming more employable and improving their eligibility.
4) To develop optimal statistical models based on given features or variables, and compare the performance of multiple machine learning models.

4.0 Literature Reviews

First Research Paper: ***Machine Learning-Based Prediction for Bank Loan Approval***

A machine learning approach for forecasting bank loan endorsement is presented by the authors of this research work. The research covers collecting data, cleaning, targeting, feature engineering, EDAs, and tailoring to various machine learning methodologies. The authors recommend using algorithms such as K-Neighbors (KNN), Random Forest, Naïve Bayes, and decision trees for this purpose.

The strategy for forecasting loan acceptance shows promising results. The Naïve Bayes method yielded the best accuracy rate of 83.73% out of all the models that were assessed. The other models yielded competitive results but fell short of the Naïve Bayes model's performance, including Random Forest with 77.23% accuracy, Decision Tree with the lowest accuracy of 63.41%, and KNN with likewise 77.23% accuracy.

The technology provides the staff of banks with an effective and dependable tool to make well-informed judgments regarding loan approvals by integrating the loan eligibility evaluation based on client facts supplied throughout the application process using strong loan prediction model.

Second Research Paper: ***An innovative application for a combined machine learning-based system that forecasts bank loan approval***

The present study's authors developed a machine learning model for predicting bank loan approvals. The authors employ a diverse set of algorithms, including Logistic Regression, Decision

Tree, Random Forest, Extra Trees, Support Vector Machine, K-Nearest Neighbors, Gaussian Naive Bayes, AdaBoost, Gradient Boosting, and deep learning models like Dense Neural Networks, Recurrent Neural Networks, and Long Short-Term Memory networks.

Out of all the models, the Extra Trees model performed better. The standalone Extra Trees model was beaten by 0.62% by an ensemble voting model that merged the best 3 ML algorithms, exhibiting an even greater accuracy. Among the models assessed in the study, this ensemble model had the highest accuracy of 87.26%, making it the most trustworthy model for predicting loan approvals.

Additionally, the study describes the creation of a straightforward desktop program for actual time loan approval prediction, which makes it possible for banks and people to evaluate loan statuses efficiently. The system's ability to greatly improve and quicken bank loan approval procedures is highlighted by its seamless integration of machine learning models into a useful application.

Third Research Paper: ***Using a comparison of the Random Forest and Decision Tree algorithms, a machine learning model for loan approval prediction***

The authors of this article, distinguished the performance of decision tree and RandomForest algorithms to provide a machine learning method for approved loan estimations. They analyzed the success rate of conventional Decision Tree algorithms in loan prediction in conjunction with a new Random Forest classifier.

A sample size of 20, evenly divided between two selected models, were used for the analysis by the authors. Considering a precision rate of 67.28% and an error rate of 32.71%, the random forest algorithm performs better than the decision tree approach, exhibiting superior accuracy. Its precision rate was 79.44% and its loss rate was 21.03%.

IBM SPSS is utilized here for statistical implementation. Although the findings were not statistically significant, the p-value of 0.33 suggests that the Random Forest model is still trending in favor. As a result, the algorithm of random forest surpasses the decision tree approach in terms of forecasting probable accuracy, demonstrating its potential for use in financial institutions.

Fourth Research Paper: ***Loan Approval Estimation Employing Machine Learning Methodology***

The authors of this study propose a model via statistical modeling for estimating loan authorization. The research involves collecting data, cleaning it, preprocessing it, and feature

engineering with EDAs. For this reason, the authors suggest using algorithms like Gradient Boosting, decision tree, naive bayes' , logistic regression and SVMs.

The methodology for loan approval prediction produces encouraging outcomes. Among the models examined in the study, the Naïve Bayes algorithm performed better than the others, particularly in terms of loan predicting accuracy. Through improved loan defaulter prediction, as demonstrated by the experimental testing, banks were able to lower their non-performing assets (NPAs) because of the Naïve Bayes model.

The paper also addresses the usefulness of using the Naïve Bayes model for real-time loan app roval estimations. The study's conclusion is that financial institutions may greatly reduce the probability of loan defaults and increase overall profitability by utilizing a strong loan app roval prediction model.

Fifth Research Paper: ***Loan Approval Prediction: A Comparative Analysis***

In this article, the authors provided a thorough examination of loan approval prediction models in their research work. For forecasting loan approvals, the study evaluates a number of different models, like random forest, decision tree, and logistic regression. To achieve a fair comparison, the authors thoroughly assessed each model using a consistent dataset, highlighting the advantages and disadvantages of each strategy.

The outcomes demonstrate that, in terms of accuracy, the Random Forest algorithm scores better than both the logistic regression and decision tree models when it comes to forecasting loan approvals. Particularly, the decision tree and logistic regression models had accuracies of 83.388% and 80.945%, respectively, and 93.648% for the Random Forest model. The authors also performed cross-validation, and although Random Forest's cross-validation score was similar to that of Logistic Regression, it still showed better generalization skills. The research highlights how the Random Forest model may greatly expedite the loan approval procedure, saving financial institutions time and risk.

5.0 Data Information

Our dataset (i.e. Loan Approval Prediction Dataset) is extracted from Kaggle and was published by 'Debdatta Chatterjee'. The purpose of this dataset, which anonymizes real-life loan applications to safeguard individual privacy, is to assist in developing predictive models that can improve the

effectiveness of loan approval procedures for financial organizations and improve the experience for loan applicants. The dataset is designed to help statistical and ML enthusiasts and practitioners develop models that predict loan approval based on several application information. This is a classical classification issue, suitable for applying different statistical and ML algorithms and techniques.

This dataset has 614 rows and 13 columns:

Features	**Descriptions**
Loan_ID	A unique identifier for each applicant
Gender	Gender of applicant (Male or Female)
Married	Marital status of applicant (Yes or No)
Dependents	No. of dependents of applicants (0, 1, 2, or 3+)
Education	Education level of applicant (Graduate or Not Graduate)
Self-Employed	If the applicant is self-employed or not (Yes or No)
ApplicationIncome	Income of applicant
CoapplicantIncome	Income of co-applicant
LoanAmount	The loan amount requested (in thousands)
Loan_Amount_Term	Term of loan in months
Credit_History	Shows if applicant has credit history
Property_Area	Area of property is located (Urban, Semiurban, and Rural)
Loan_Status	whether the loan was approved or not (Y for approved, N for not approved)

Figure 1: Features and Descriptions of Dataset

6.0 Methodology

❖ 6.1 Visualization

```
   Loan_ID Gender Married Dependents     Education Self_Employed ApplicantIncome
1 LP001002   Male      No          0      Graduate            No             5849
2 LP001003   Male     Yes          1      Graduate            No             4583
3 LP001005   Male     Yes          0      Graduate           Yes             3000
4 LP001006   Male     Yes          0 Not Graduate            No             2583
5 LP001008   Male      No          0      Graduate            No             6000
6 LP001011   Male     Yes          2      Graduate           Yes             5417
  CoapplicantIncome LoanAmount Loan_Amount_Term Credit_History Property_Area
1                 0          NA               360              1         Urban
2              1508         128               360              1         Rural
3                 0          66               360              1         Urban
4              2358         120               360              1         Urban
5                 0         141               360              1         Urban
6              4196         267               360              1         Urban
  Loan_Status
1           Y
2           N
3           Y
4           Y
5           Y
6           Y
```

Figure 2: Load the Dataset

We may determine any clear mistakes or inconsistencies by looking through the first part of the dataset. It is also important to evaluate the data quality in-depth. In order to accomplish this, we will examine data distributions and summary statistics in order to determine any anomalies, missing values, or unusual trends.

```
> summary(dataset)
     Loan_ID        Gender    Married    Dependents          Education   Self_Employed
 LP001002:  1          : 13      :  3     : 15      Graduate    :480      : 32
 LP001003:  1   Female:112   No :213   0 :345      Not Graduate:134   No :500
 LP001005:  1   Male  :489   Yes:398   1 :102                         Yes: 82
 LP001006:  1                         2 :101
 LP001008:  1                         3+: 51
 LP001011:  1
 (Other) :608
 ApplicantIncome CoapplicantIncome   LoanAmount    Loan_Amount_Term
 Min.    :  150   Min.    :    0     Min.    :  9.0   Min.    : 12
 1st Qu.: 2878   1st Qu.:    0     1st Qu.:100.0   1st Qu.:360
 Median : 3812   Median : 1188     Median :128.0   Median :360
 Mean   : 5403   Mean   : 1621     Mean   :146.4   Mean   :342
 3rd Qu.: 5795   3rd Qu.: 2297     3rd Qu.:168.0   3rd Qu.:360
 Max.   :81000   Max.   :41667     Max.   :700.0   Max.   :480
                                   NA's   :22      NA's   :14
 Credit_History     Property_Area Loan_Status
 Min.   :0.0000   Rural    :179   N:192
 1st Qu.:1.0000   Semiurban:233   Y:422
 Median :1.0000   Urban    :202
 Mean   :0.8422
 3rd Qu.:1.0000
 Max.   :1.0000
 NA's   :50
```

Figure 3: Summary of Dataset

Here we can see that, 51 references are marked with a plus sign (+), which could suggest a data formatting problem or special cases. The credit_history variable's average is 0.8422, which marks as strange because it should be a binary value (Customers having a past credit

history score as one, while those without do not) which mean there is an imbalance or some incorrect values in the data. There are absent or empty values in a number of categorized variables such as gender, dependents, married, and self_employed. Either appropriate values should be added to these values that are not present or the sources containing the missing data should be eliminated. Likewise, Nas are present in the loan_amount, credit_history, loan_amount_terms variables, and should be handled similarly.

To make sure the dataset is accurate and reliable for analysis, we need to carefully examine and preprocess the data by addressing these issues, including fixing inconsistencies and understanding any unusual patterns or values.

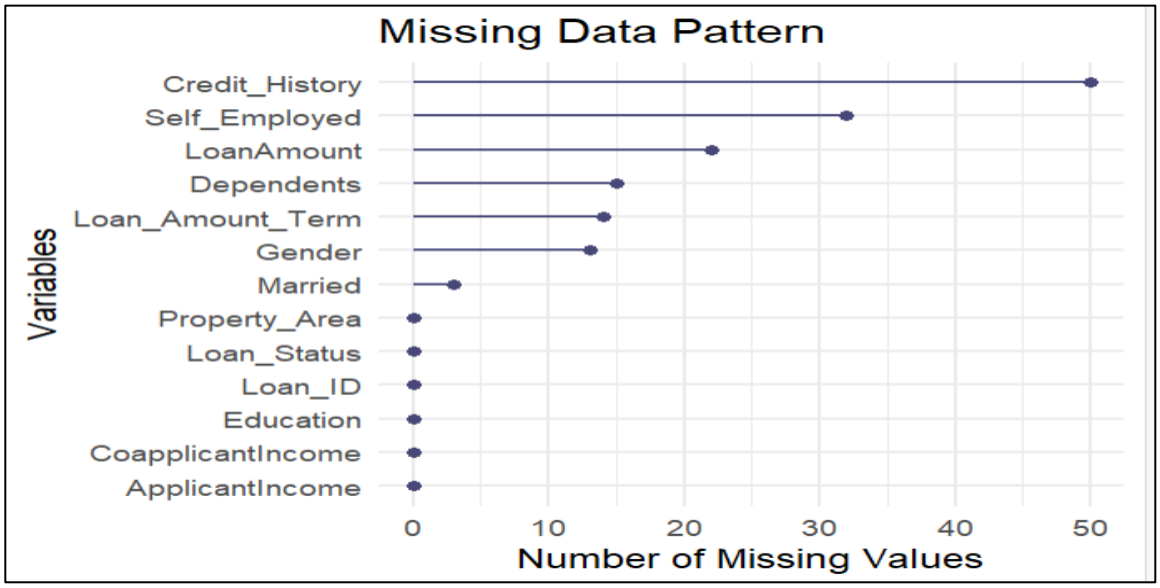

Figure 4: Checking Missing Values

Here in our dataset, there are seven variables with missing values. We should examine the pattern of distribution of numerical variables such as LoanAmount and ApplicationIncome to get a more thorough unpacking of the dataset. To interpret these variables, we use histrograms and boxplots that will provide crucial insights on their distributions and draw attention to any anamolies.

- Let's start by drawing histograms for LoanAmount and Application. These will display the distribution of the data and indicate any old patterns or skews.
- Let's then create boxplots for the same variables. We can more clearly observe the mean, spread, and any outliers in our data with the use of boxplots.

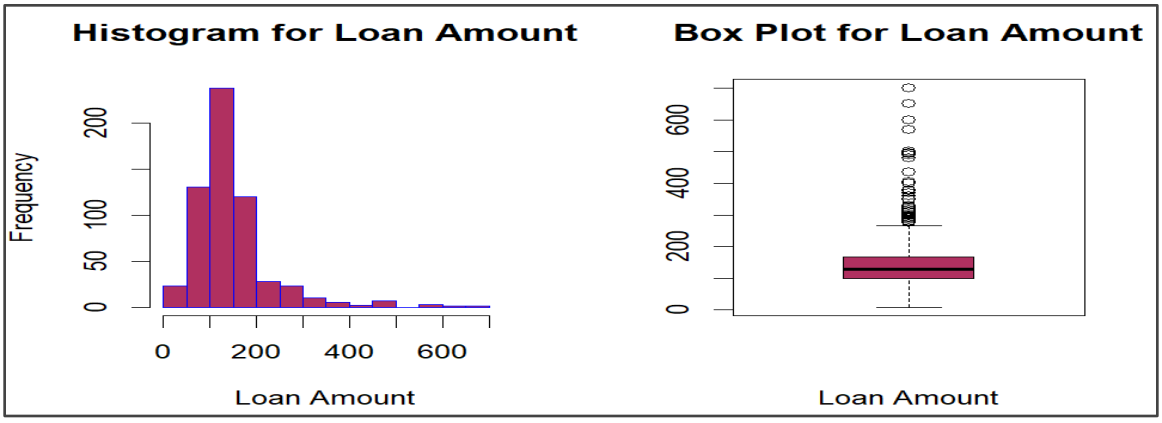

Figure 5: Histogram and Boxplot for Loan Amount

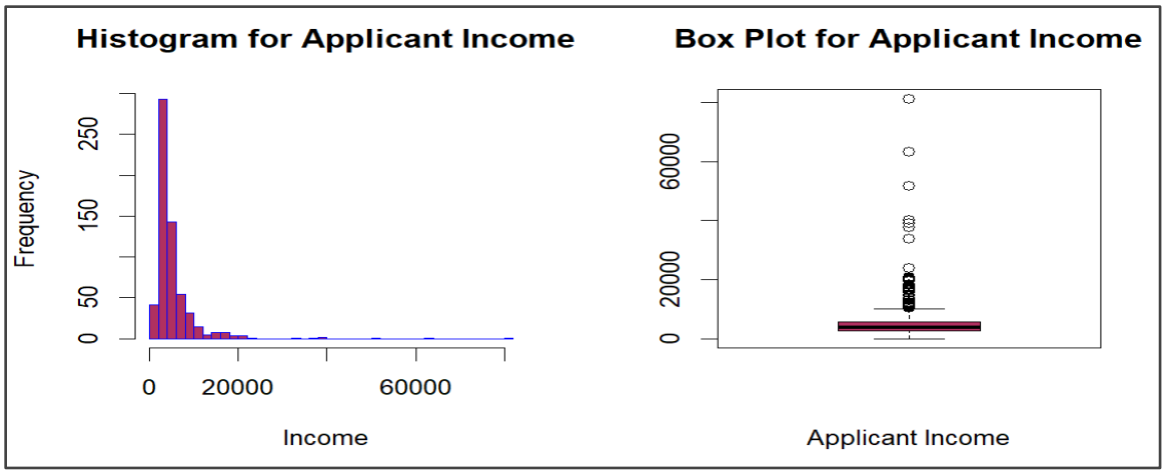

Figure 6: Histogram and Boxplot for Applicant Income

As we can see, there are some extreme values in LoanAmount as well as in ApplicationIncome. In order to delve deeper into the dataset, let us investigate whether the applicants' educational level has an impact on the distribution of loan amounts:

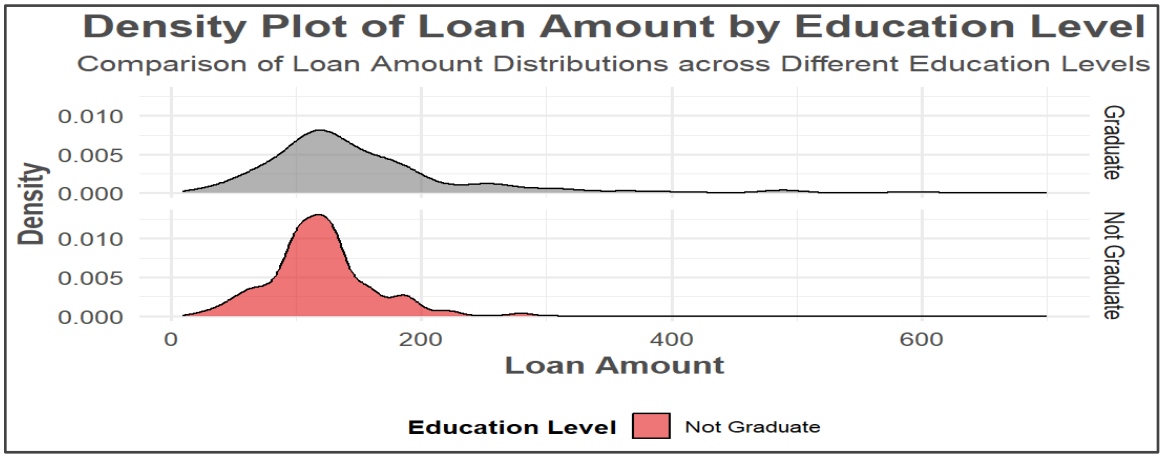

Figure 7: Density Plot of Loan Amount by Education Level

We see that graduates shows more anomalies and a wider distribution of loan amount than non-graduates. In order to gain a deeper comprehension of the the information set, let's analyze the categorized variables, such as

Figure 8: Barplot of Categorical Variables

As we can see in the Gender graph, males account for a higher percentage of the set of data, with over half of their requests for loans being approved. Even though it's less in number, a sizable portion of female applicants' applications are also accepted.

Likewise, the Education graph illustrates that graduates have more loan approvals than non-graduates. In the Married graph, married applicants have a higher number of loan approval as compared to unmarried ones. The Self Employed graph demonstrates that applicants who are not self-employed are more likely to have their loan applications approved. Urban and rural areas have lower approval rates than semi-urban areas, according to the Property Area graph. Finally, the Credit History graph clearly illustrates that clients with a credit history have a significantly better loan approval rate than those without one.

❖ 6.2 Data Cleaning

Data Cleaning, we need to deal with the issues we have discovered within the dataset. Let's summarize the major issues:

1) **Missing Values:** Taking into account the significance of each variable, we must select an appropriate handling strategy for the missing data.

2) **Loan Amount and ApplicantIncome Outliers:** We need to decide how to handle these outliers, these are the result of measurement errors, recording problems, or real anomalies.

❖ 6.3 Handling Missing Values

Given that the missing values in this dataset exist in several variables in an apparently random manner, we make the assumption the missing values are systematic. Both category and numerical data can have missing values. In order to solve this, we calculated the missing values in each column and first showed the missing data pattern which is shown in above.

And then, using the complete.cases method, we filtered out any rows with incomplete data in order to address the missing values. The loan_clean dataframe, which now only has rows with no missing values in any of the fields, is the product of this operation. After cleaning the dataset, another check verified that there are no longer any missing values for any of the variables, meaning the dataset is ready for further analysis.

```
+   filter(complete.cases(.))
> colSums(is.na(loan_clean))
          Gender          Married       Dependents        Education     Self_Employed
               0                0                0                0                0
 ApplicantIncome CoapplicantIncome       LoanAmount Loan_Amount_Term    Credit_History
               0                0                0                0                0
   Property_Area      Loan_Status
               0                0
>
```

Figure 9: Handling Missing Values

Now, let's talk about the dataset's extreme values. It makes sense that certain clients might ask for greater loans for a variety of reasons when looking at the LoanAmount variable. We may use a log transformation to normalize the data and lessen the effect of these extreme values.

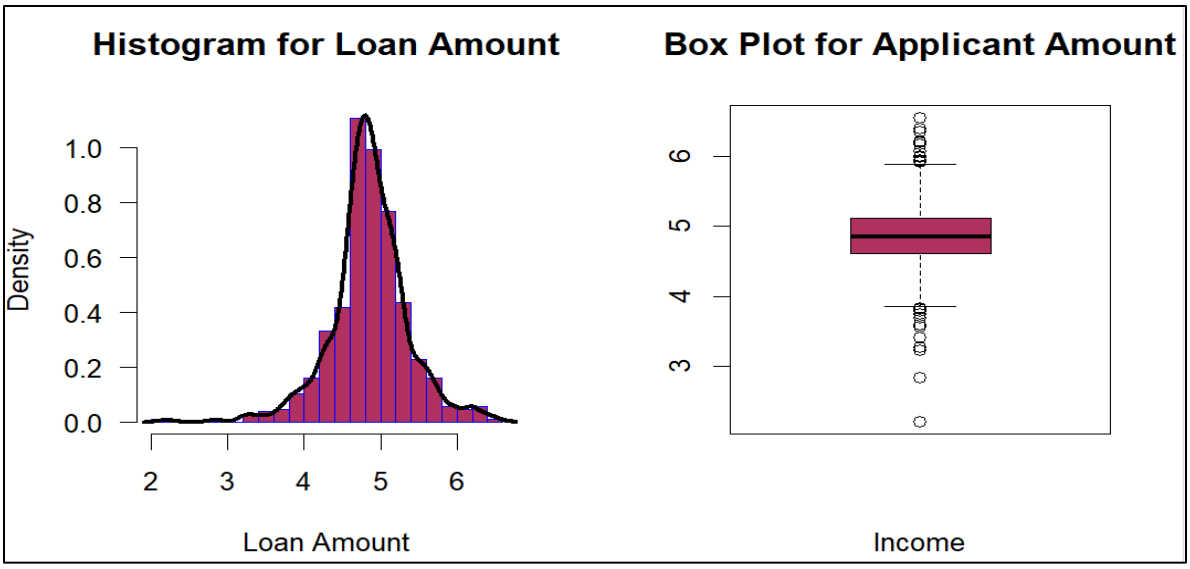

Figure 10: Histogram and Boxplot for Loan Amount and Applicant Income

With this modification, the distribution now seems more in line with normal, and the impact of extreme values has diminished dramatically.

It is preferable to add **applicant income** and **Co-applicant income** to total income for ApplicantIncome, and then log convert this combined variable. The CART imputation teachnique will be applied. The average of a measurement can be used to fill in missing values if we are aware that the measurement's values should fall within a specific range. As a result, the distribution is improved and becomes closer to being normal.

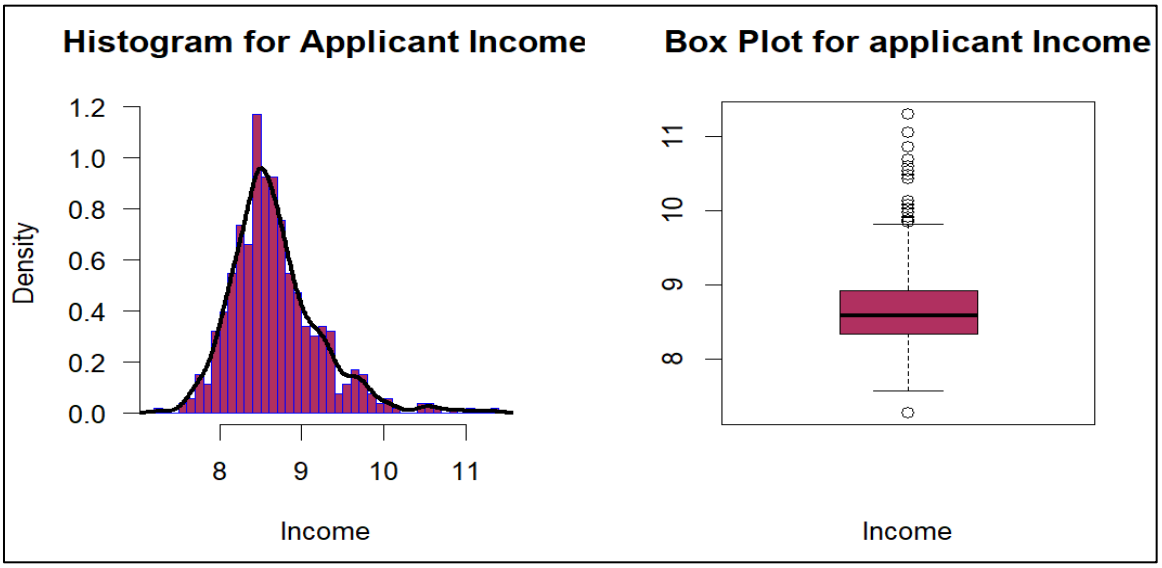

Figure 11: Histogram and Boxplot for Applicant Income

7.0 Modeling

Splitting the data could play a crucial role while training a machine learning model. It helps to avoid overfitting. We have split the data into two for each algorithm training and testing dataset.

A selected portion of data obtained from the set used for training would be the test collection, whereas the data used for training would be a portion of the data we use to train the models. The test set acts as a substitute for the actual data, which is useful in assessing how well the model performs.

```
168  # Train-test split
169  set.seed(42)
170  sample <- sample.int(n = nrow(tr), size = floor(.70*nrow(tr)), replace = FALSE)
171  trainnew <- tr[sample, ]
172  testnew  <- tr[-sample, ]
173
```

Figure 12: Train-Test Split

Model Chosen:

7.1 Logistic Regression

One supervised machine learning technique termed logistic regression is used which performs binary classifications. To avoid overfitting the variables should be selected cautiously. To determine important variables. Let's consider some rational significance in forecasting approved loans. An applicant's chances of getting approved for a loan might be raised if

- The candidate has a greater amount of loan-taking experience.
- The applicant has a greater salary
- The applicant has a higher education
- The applicant has solid employment.

For logistic regression, the Credit_history variable is focused more as it has appeared as one of the key factors for determining loan approval.

The reason we chose logistic regression for the loan approval prediction task is due to several key reasons.

- Loan approval is a binary classification problem where the dependent variable (loan_status) has two values either yes or no. Logistic regression is well suited for this type of prediction.
- Logistic regression coefficients are easily interpretable which allows us to easily understand the predictor variable.

7.2 Decision Tree

Binary splits on variables that predict are created by decision trees to categorize fresh data. The classical decision trees we are utilizing here adhere to the following algorithm:

- Selecting the Optimal Predictor Variable Decide which variable divides the data between two groups the most effectively.
- Divide the Data: Using the chosen variable as a guide, separate the data into two separate categories.
- Replicate the actions: This procedure should be repeated until there are fewer observations in each subgroup than the required minimum.
- Categorize Fresh Observations: A new case is classified by passing it to a final node along the tree and assigning it that node's model outcome value.

Applying the rpart package inside R, decision trees may be developed and trimmed. The procedure operates as follows:

- Grow the Tree: The first tree may be grown by using the rpart() function.
- Print the tree and examine it: Examine the fitted model by printing the tree as well as its summary. The tree could often be too big.
- Check out the table of complexity parameters: Examine the Cptable found in the Rpart output. This table has the following items:
    - Relative Error: It is the error rate for a tree of a specific size in the training sample.
    - Cross-Validated Error (xerror): Defined by applying the training sample to tenfold cross-validation.
    - Standard Error (xstd): The cross-validated error standard error.
- Plot Cross-Validated Error: For plotting the cross-validated error versus the complexity parameter, use the plotcp() method.
- Select the Tree's Final Size: Choose the smallest possible tree whose validated error is less than the minimal cross-validated error by one standard error. As an illustration:
    - Cross-Validated Error Minimum: 0.618
    - Error standard: 0.0618
    - Range: 0.56 to 0.68, or 0.618 ± 0.0618
- Choose the Correct Tree: A tree that has only one split (cross-validated error = 0.618) from the table satisfies this condition. Trim the Tree: Make use of the proper complexity

parameter when using the prune() function. Pruning a tree that has a single split, for example, yields a tree with the necessary size since the tree's complexity parameter is 0.02290076.

- The Pruned Tree's Plot: See the trimmed tree to forecast the status of the loan. Starting at the top, move left if the scenario is true and right otherwise to understand the tree. An observation is categorized upon arrival at a terminal node. Every node displays the percentage of each class of the sample that it represents in addition to the possibility of the classes that are present in that node.

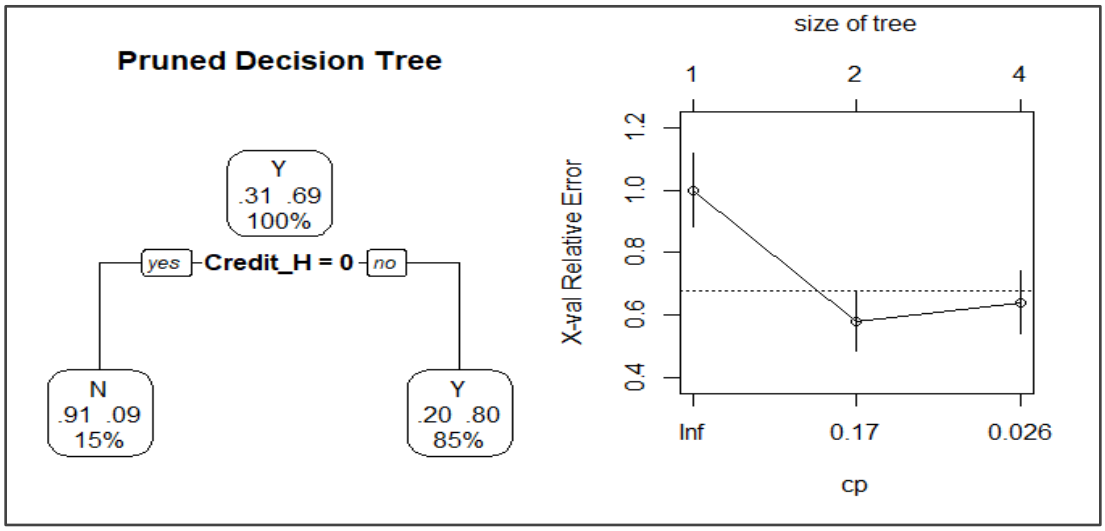

Figure 13: Pruned DT and Size of Tree

In R, while building decision trees using the rpart() function which is derived from the rpart package, we first grow a tree and evaluate the fit by printing the tree structure. Frequently, the tree becomes extremely complicated and may require pruning to improve generality. To identify the optimal mix of complexity and efficiency, we examine the "cptable" from the rpart() result. This table contains information about prediction errors for trees of various sizes. The complexity parameter (cp) discourages bigger trees and encourages simpler models. The number of splits, each of which leads to a new terminal node, is used to calculate tree size. The table contains columns such as relative error (rel error), which displays the error rate in training data for trees of various sizes. Cross-validated error (xerror), calculated using 10-fold cross-validation on training data, predicts how well the tree will generalize to new data. The standard error of cross-validation (xstd) measures the variability of these estimations.

Using the plotcp() method, we can see the cross-validated error vs the complexity parameter. To determine the final tree size, we want to find the smallest tree whose cross-validated error is within one standard error of the minimal cross-validated error value. In our example, the minimal cross-validated error is 0.618, whereas the standard error is 0.0618. To find the shortest tree, we look for a cross-validated error of 0.618 ± 0.0618, or 0.56 to 0.68. According to the table, a tree with one split (cross-validated error = 0.618) fits the condition. The cptable shows that a tree with one split has a complexity value of 0.02290076. Using the expression prune (dtree, cp=0.2290076), we may acquire the required tree size. Next, we plot the trimmed tree to anticipate loan status, starting at the top and progressing left if a condition is true and right otherwise. When an observation reaches the terminal node, it is categorized. Each node provides the probability of the classes in that node, as well as the proportion of the whole sample.

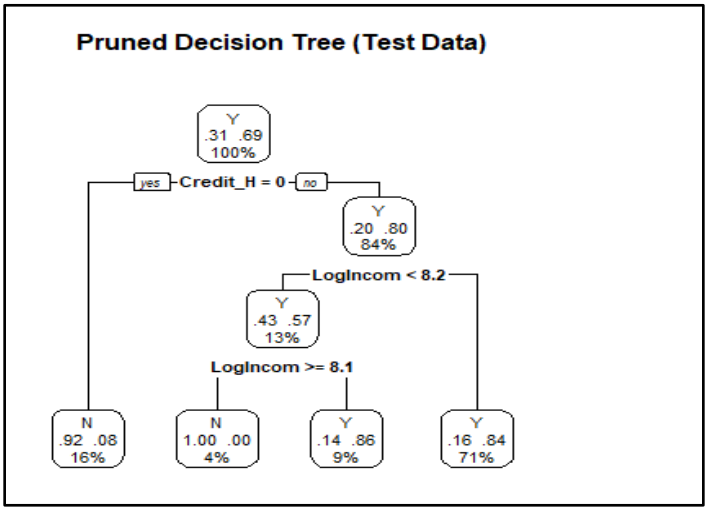

Figure 14: Pruned DT (Test Data)

7.3 Model Performance

Decision Tree

```
>
> # Calculate accuracy for the training data
> train_accuracy <- sum(diag(dtree.perf)) / sum(dtree.perf)
> train_accuracy
[1] 0.8135135
>
> # Calculate accuracy for the test data
> test_accuracy <- sum(diag(dtree_test.perf)) / sum(dtree_test.perf)
> test_accuracy
[1] 0.8616352
>
```

Figure 15: Accuracy of DT for Training & Test Data

We can observe that the training accuracy of the decision tree is 81%. However, there is a slight increase in the accuracy on test data by 5%.

Logistic Regression

```
> cat("Logistic Regression Training Accuracy:", train_accuracy_log1, "\n")
Logistic Regression Training Accuracy: 0.8135135
> cat("Logistic Regression Test Accuracy:", test_accuracy_log1, "\n")
Logistic Regression Test Accuracy: 0.8176101
>
```

Figure 16: Accuracy of Logistic Regression for Training $ Test Data

The logistic regression accuracy is about 81% on both training and testing. However, test accuracy is slightly more accurate.

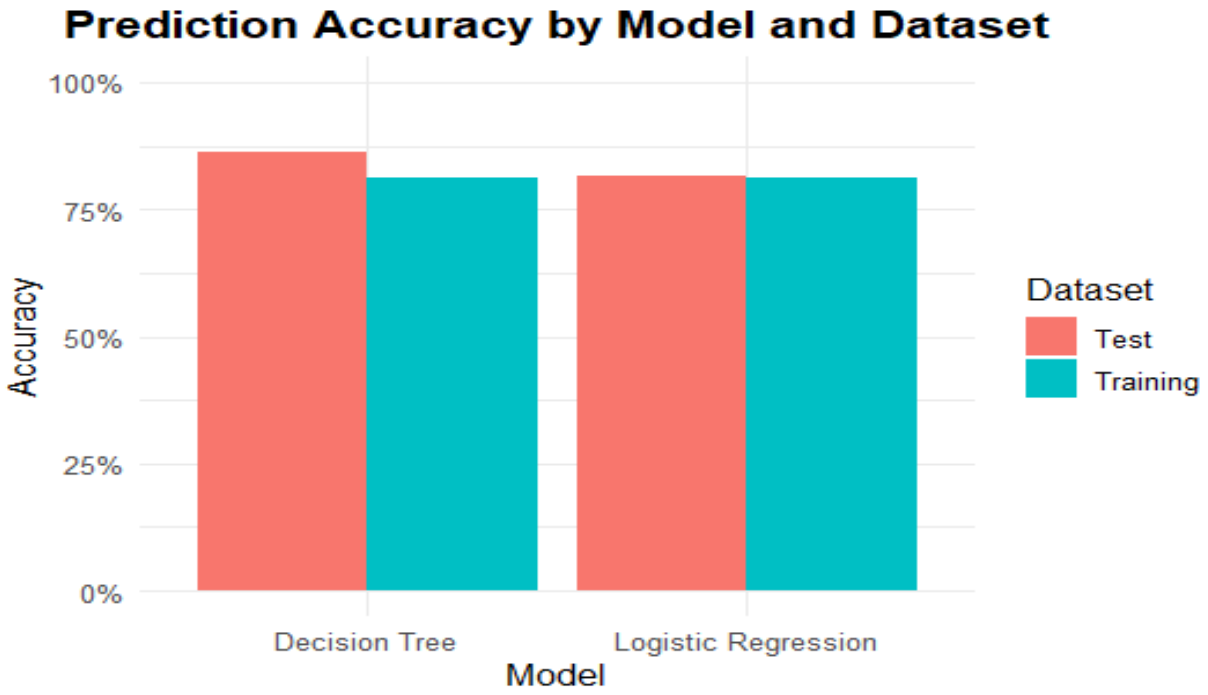

Figure 17: Prediction Accuracy by Model & Dataset

The above bar chart illustrates the comparison of two algorithm's accuracy of predicting loan approval. Here we have used accuracy metrics to evaluate the models. Accuracy is the ratio of correctly predicted instances out of total instances.

Both models Decision Tree and Logistic Regression have similar training accuracy However, with test data the model decision tree is marginally better performing than logistic regression.

We can also say that, both logistic regression and decision tree show robust performance with almost similar accuracy and they are effective for the prediction tasks and are well-calibrated. Which indicates that there were no significant overfitting.

8.0 Limitations and Future Study

Limitations

- Small datasets can trigger decision trees to overfit, performing well during training but poorly on fresh data. If the true relationship is non-linear, bias may result from the logistic regression which assumes linearity.
- Small sample sizes, which can result in inaccurate forecasts, and also the missing variables like debt-to-income ratios, applicant's job stability, and economic conditions that could influence model performance are some instances of data limitations.
- Unreliable imputation of missing data might introduce biases and inaccuracies into the data, resulting in skewed conclusions and poor decision-making.
- The decision tree algorithm and logistic algorithm algorithm might not be the most appropriate choices for the data. To improve accuracy, more statistical models or machine learning models might be investigated.

Future Study:

- In the future, we can concentrate on investigating alternate algorithms like as deep learning neural network, RF, and SVMs in order to increase accuracy using larger datasets.
- By lowering bias and variance and using ensemble techniques like boosting and bagging can improve the model performance even further.
- In order to improve prediction accuracy, we might enrich the data with key variables like debt-to-income ratio and applicant's job stability to improve model robustness and adaptability.
- More sophisticated methods, including feature engineering and hyperparameter tuning can improve the accuracy and fit of the model to the data, hence increasing its predictive power.
- To foster trust and guarantee moral decision-making, fairness-aware machine learning approaches may be used to eliminate bias and to encourage transparency in model development.

9.0 Impact

The study conducted for our project indicates that the imputed techniques employed have the potential to greatly improve loan application decision-making. Reducing the likelihood of lending money to applicants who pose a high risk and streamlining the loan approval procedure as a whole might be possible with appropriate imputation of missing data. Because institutions of finance would be better able to recognize and mitigate lending risks, adopting effective imputed techniques might result in to lower expenses related to loan defaults.

Financial institutions could reduce any unfair advantage in their decision-making procedure with more precise and comprehensive data. By doing this, it would be guaranteed that all applicants receive equal treatment and that financing choices are made using impartial standards. The use of these techniques may benefit loan applicants, especially those who might have previously suffered because of inadequate or missing information in their applications. It's possible that these candidates will now have a higher probability of getting loans, giving them access to necessary funds.

Our analysis has the potential to change the relationship of power between creditors and their customers by enabling institutions of finance along with those seeking loans to make more informed choices through the provision of more precise and thorough data. It is imperative to take into account any possible disadvantages linked to our techniques. Our imputation models have the potential to greatly enhance the loan application procedure if they prove to be correct. Inaccurate models, on the other hand, run the risk of adding to the biases and inaccuracies already present in the data, which could result in poor decisions and even worsen already-existing inequities.

Furthermore, there is a chance that an excessive dependence on imputed data would cause other crucial elements, such qualitative data and unique situations, to be disregarded throughout the decision-making process.

10.0 Conclusion

With the use of a variety of application features, we designed and evaluated statistical prediction model for approved loans in this study. In order to reduce financial risk for lending institutions and promote equitable access to credit, the main goal was to increase the accuracy of loan approval predictions. We aimed to develop strong models which could expedite the loan approval process by examining different variables included.

Based on our findings, the decision tree model surpassed the model of logistic regression approach on the training and test datasets, demonstrating its improved generalization and predictive power for loan approval outcomes. This discovery highlights the applicability of decision tree models in complicated financial decision-making processes involving several factors.

Regardless of these optimistic results, our analysis also pointed us a number of areas where the models needed improvement. It was observed that there were problems with small sample sizes and missing important variables, as well as overfitting in decision trees and possible bias in logistic regression. We explored about a number of approaches to overcome these constraints, such as investigating alternative models, ensemble techniques, and data enrichment. In order to guarantee fair and impartial loan approval procedures, it was also emphasized that fairness-aware techniques were necessary.

To sum up, the creation of precise and trustworthy loan prediction models is essential for lowering financial risks and fostering economic expansion. Our results lay the groundwork for more studies and advancements in loan approval prediction modeling. Lenders and borrowers will gain from these models' increased efficacy and fairness when the constraints that have been found are addressed and innovative techniques are investigated.

11.0 References

Bhargav, P., & Sashirekha, K. (2023). A machine learning method for predicting loan approval by comparing the random forest and decision tree algorithms. *Journal of Survey in Fisheries Sciences, 10*(1S), Special Issue 1. https://doi.org/10.17762/sfs.v10i1S.414

Kadam, A. S., Nikam, S. R., Aher, A. A., Shelke, G. V., & Chandgude, A. S. (2021). Prediction for loan approval using machine learning algorithm. *International Research Journal of Engineering and Technology (IRJET), 8*(4), 4089. https://www.irjet.net/archives/V8/i4/IRJET-V8I4708.pdf

Khan, A., Bhadola, E., Kumar, A., & Singh, N. (2021). Loan approval prediction model: A comparative analysis. *Advances and Applications in Mathematical Sciences, 20*(3), 427-435. Retrieved from https://www.mililink.com/upload/article/1759044670aams_vol_203_january_2020_a10_p427-435_afrah_khan_and_nidhi_singh.pdf

Uddin, N., Ahamed, M. K. U., Uddin, M. A., Islam, M. M., Talukder, M. A., & Aryal, S. (2023). An ensemble machine learning based bank loan approval predictions system with a smart application. *International Journal of Cognitive Computing in Engineering, 4,* 327-339. https://doi.org/10.1016/j.ijcce.2023.100029

Viswanatha, V., & Ramachandra, A. C. (2023). Prediction of loan approval in banks using a machine learning approach. *International Journal of Engineering and Management Research, 13*(4), 7-19. https://doi.org/10.31033/ijemr.13.4.2

Appendix

GitHub Link: https://github.com/anuz505/Loan-approval-prediction.git