Machine Learning Based Loan Approval Prediction

By

Mohammed Mejbha Uddin
ID: 233000761

Department of Digitalization Innovation
and Entrepreneurship

Supervised by

Sohrab Hossain, Assistant Professor

In partial fulfillment of the requirement for the degree of *Master of Science* in

Data Analytics and Design Thinking for Business

July,2024

East Delta University

Noman Society, East Nasirabad, Chittagong-4209

Machine Learning Based Loan Approval Prediction

This thesis is submitted in partial fulfillment of the requirement for the degree of

Master of Science in *Data Analytics and Design Thinking for Business*

By

Mohammed Mejbha Uddin
ID: 233000761

Supervised by
Sohrab Hossain, Assistant Professor

**Department of Digitalization Innovation
and Entrepreneurship**

East Delta University, Noman Society, East Nasirabad, Chittagong-4209

The Postgraduation thesis titled “**Machine Learning Based Loan Approval Prediction’’** is submitted by ID No. 233000761 , has been accepted as satisfactory in fulfillment of the requirements for the degree of Master of Science (M.Sc.) in the Department of Digitalization Innovation and Entrepreneurship to be awarded by East Delta University.

Board of Examiners

1. ____________________________________

Mohammed Nazim Uddin, PhD Chairman
Associate Dean & Professor
School of Science, Engineering & Technology
East Delta University (EDU)

2. ____________________________________

Md. Ishtiaque Aziz Zahed, PhD
Member
Associate Professor
School of Science, Engineering & Technology
East Delta University (EDU)

3. ____________________________________

Linkon Chowdhury Member
Assistant Professor
School of Science, Engineering & Technology
East Delta University (EDU)

4. ____________________________________

Sohrab Hossain Supervisor
Assistant Professor
School of Science, Engineering & Technology

Declaration

It is officially declared that all aspects of this dissertation are claimed to be genuine and that no part of it was previously presented anywhere for the issuance of any other degree or certificate. In addition, no part of this thesis has been presented elsewhere for the granting of any other degree or certificate. The acknowledgment and citation of any and all works of literature or other people's work that are brought up in this thesis can be found in the section titled "References."

Date:

Mohammed Mejbha Uddin

ID:233000761

Dedication

I would like to dedicate my research work to our advisor, teachers, and mentors for their guidance and assistance during the process of writing this thesis, without which it would not be possible to finish our dissertation. They not only gave me the opportunity to learn and gain experience, but they also provided us with valuable direction at the times when we needed it the most.

Additionally, we would want to dedicate our study work to the Department of Digitalization Innovation and Entrepreneurship here at EDU as well as our fellow classmates. We are really appreciative of the total effort that the department put forth to ensure the successful completion of our course and to give us the opportunity to earn knowledge in addition to gaining information.

Acknowledgment

To begin, we would want to express our gratitude to the Almighty for bestowing His grace upon us. Please accept our sincere appreciation for this gift. We would not be able to proceed with our journey towards the completion of our studies and thesis if it were not for His magnificence.

Then, we would like to express our gratitude to our advisor, Sohrab Hossain Sir, Lecturer, for his exceptional and unending support during the process of the thesis. We owe him a great deal of gratitude for his sage advice and the fact that he persistently endeavored to uncover the absolute finest version of each of us. We would also like to use this occasion to extend our gratitude to the valuable teachers at East Delta University for their constant encouragement and support in our endeavors.

Thank you very much from the bottom of our hearts, especially to our families. We are indebted to our families for the majority of the sacrifices that they have made for our cause, and there are no adequate words to express our gratitude in this regard. This investigation might turn out to be lot more difficult and tough if it weren't for their unwavering love and support. Their prayers for us have been an inspiration and a source of strength for us up to this time. We would also like to extend our gratitude to the majority of our travelling companions who inspired us to work towards our objective.

In conclusion, we would like to convey our deepest gratitude and respect to all of those individuals who assisted us in any way throughout the course of the majority of the development of our research.

Abstract

This study aims to develop a robust predictive model for loan approval using advanced machine learning techniques. Utilizing a comprehensive dataset from a financial institution, which includes applicant demographics, loan details, and credit history, we conducted a thorough exploratory data analysis (EDA) to uncover data distributions, identify patterns, and detect outliers. Various machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, and k-Nearest Neighbors (k-NN), were implemented to classify loan approval statuses. The dataset was partitioned into training and testing sets to rigorously evaluate model performance, employing metrics such as accuracy, precision, recall, and F1 score. Among the models tested, the Random Forest algorithm demonstrated superior accuracy, outperforming other models while maintaining significant predictive capabilities. This research highlights the efficacy of machine learning in enhancing financial decision-making processes through the accurate prediction of loan approvals, thereby contributing to the optimization of lending strategies.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Exploratory Data Analysis	**EDA**
Convolutional Neural Network	**CNN**
Machine Learning	**ML**
Recurrent Neural Network	**RNN**
K-Nearest Neighbors	**KNN**
True Positives	**TP**
False Positives	**FP**

Chapter 1

Introduction

1.1 Background of The Research

The financial sector is the backbone of modern economies, and banks play a pivotal role in providing financial services to individuals and businesses. One of the most critical services offered by banks is the provision of loans. However, lending money involves significant risk, as borrowers may default on their repayments. Consequently, accurately predicting loan approval and default has become an essential task for financial institutions. This background section explores the evolution of loan prediction research, highlighting the importance, challenges, and advancements in this field.

Loan prediction is crucial for both banks and borrowers. For banks, accurate loan prediction models minimize the risk of financial losses by identifying potentially risky borrowers. This, in turn, helps in maintaining the financial stability of the institution. For borrowers, an efficient loan prediction system ensures a fair and unbiased evaluation process, facilitating access to necessary funds for personal or business purposes.

Effective loan prediction models enhance operational efficiency, reduce the cost of manual underwriting, and improve customer satisfaction by speeding up the loan approval process. Moreover, they contribute to the overall health of the financial system by preventing the accumulation of bad debts and promoting responsible lending practices.

1.2 Objective of The Research

The primary objective of this research is to develop and evaluate an advanced machine learning model for predicting loan approval. This involves Gathering a comprehensive dataset that includes relevant borrower attributes and loan characteristics, followed by cleaning and preprocessing the data to ensure quality and consistency.

- **Exploratory Data Analysis (EDA)**: Conducting an in-depth analysis of the dataset to identify key patterns, correlations, and insights that can inform the model development process.
- **Feature Engineering**: Creating and selecting the most relevant features that enhance the predictive power of the model, including transforming and encoding categorical variables, and handling missing values.
- **Model Development**: Implementing various machine learning algorithms, such as Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, and Gradient Boosting Machines, to build robust predictive models.
- **Model Evaluation and Comparison**: Assessing the performance of each model using appropriate metrics (accuracy, precision, recall, F1-score) and comparing their effectiveness in predicting loan approval.
- **Optimization and Tuning**: Fine-tuning the best-performing models through hyperparameter optimization to achieve the highest possible predictive accuracy and reliability.
- **Interpretability and Bias Mitigation**: Ensuring that the final model is interpretable and free from biases, providing transparency in the decision-making process and fair evaluation for all applicants.
- **Deployment and Validation**: Implementing the chosen model in a real-world setting, validating its performance with live data, and making necessary adjustments to maintain accuracy and relevance over time.
- **Impact Assessment**: Evaluating the broader implications of the model's deployment, including its effect on loan approval rates, default rates, and overall financial health of the lending institution.

By achieving these objectives, the research aims to provide a robust and reliable tool for loan prediction that enhances the decision-making process of financial institutions, minimizes risks, and promotes fair lending practices.

1.4 Chapter Organization

This research might be divided into four sections. They are as follows:

- Chapter 1 presents the research subject and explains about the background, as well as the objectives of the proposed research topic.
- Previous different studies are discussed in Chapter 2.
- There are discussed total methodology in Chapter 3
- The performance evaluation, output, analysis are covered in Chapter 4.
- The result of the study as well as its potential applications in the future are discussed in Chapter 5.

Chapter 2

Literature Review

2.1 Research Area

The loan approval process has traditionally been a critical function in financial institutions, relying heavily on manual scrutiny and human judgment. With the advent of big data and machine learning, there has been a significant shift towards automating this process to enhance accuracy, efficiency, and fairness. Machine learning models can analyze vast amounts of historical data, identify patterns, and make predictions about loan approval with higher precision. This approach not only reduces human bias but also accelerates the decision-making process, providing a competitive edge to financial institutions.

2.2 Major Challenges

Despite its potential, the application of machine learning in loan approval faces several challenges:

- **Data Quality and Availability**: The effectiveness of machine learning models heavily depends on the quality and quantity of data available. Incomplete, inconsistent, or biased data can significantly impair model performance (Brown & Mues, 2012).
- **Feature Engineering**: Identifying the right features that influence loan approval is crucial. This process often requires domain expertise and can be time-consuming (Yadav et al., 2020).
- **Model Interpretability**: Financial institutions need to ensure that their models are not only accurate but also interpretable. Regulatory requirements often demand clear explanations for decisions made by automated systems (Doshi-Velez & Kim, 2017).
- **Regulatory Compliance**: Compliance with regulations such as the Fair Lending Act and the Equal Credit Opportunity Act is critical. Machine learning models must be designed to avoid discrimination and ensure fairness (Berk et al., 2018).
- **Security and Privacy**: Handling sensitive financial data requires robust security measures to protect against breaches and misuse (Abadi et al., 2016).

2.3 Studies Based on Loan Approval Prediction Using Machine Learning

Several studies have explored the use of machine learning techniques for loan approval, demonstrating varying levels of success and highlighting different aspects of the implementation process.

Thomas et al. (2002) explored the use of credit scoring models to predict loan defaults. They compared traditional logistic regression models with advanced machine learning techniques such as decision trees and neural networks, finding that machine learning models generally outperformed traditional methods in terms of prediction accuracy.

Lessmann et al. (2015) conducted a comprehensive benchmarking study comparing various machine learning algorithms, including support vector machines, random forests, and k-nearest neighbors, for credit scoring. Their research demonstrated that ensemble methods, particularly random forests, consistently provided superior performance compared to single classifiers.

Malik & Khan (2020) focused on the application of deep learning techniques for loan approval prediction. They utilized a combination of convolutional neural networks (CNN) and recurrent neural networks (RNN) to capture both the spatial and temporal aspects of the data. Their findings indicated that deep learning models could effectively handle large, complex datasets and provide high prediction accuracy.

Yadav et al. (2020) explored the application of various machine learning algorithms for loan prediction, including logistic regression, decision trees, random forests, and gradient boosting machines. Their study highlighted the importance of feature engineering and hyperparameter tuning in improving model performance.

These studies collectively underscore the potential of machine learning in transforming the loan approval process. They also highlight the importance of addressing challenges related to data quality, model interpretability, and regulatory compliance to ensure successful implementation

Chapter 3

Methodology

3.1 Methodology

The methodology section outlines the systematic approach taken to perform loan prediction analysis using machine learning techniques. The primary steps involved include data collection, preprocessing, exploratory data analysis (EDA), model selection, training and evaluation, and result interpretation.

3.2 Dataset Collection

The dataset used in this analysis was obtained from Kaggle's Loan Prediction Problem Dataset. This dataset contains various features related to loan applicants, including demographic details, loan details, and credit history.

3.3 Data Preprocessing

3.3.1 Handling Missing Values

- Numerical missing values were imputed using the mean or median of the respective columns.
- Categorical missing values were filled using the mode of the respective columns.

3.3.2 Encoding Categorical Variables

- Categorical features were converted into numerical values using one-hot encoding. This transformation is essential for most machine learning algorithms to process the data efficiently.

3.3.3 Feature Scaling

- Standardization was applied to the numerical features to ensure they have a mean of 0 and a standard deviation of 1. This step is particularly crucial for algorithms like k-Nearest Neighbors that are sensitive to feature scaling.

3.4 Exploratory Data Analysis (EDA)

3.4.1 Outliers of the Data set

In the context of the loan prediction dataset, checking for outliers serves several important purposes. Outliers are data points that deviate significantly from the majority of the data, and their presence can impact the accuracy and reliability of machine learning models. For instance, extremely high or low values in loan amounts, income, or credit scores can skew the results and lead to misleading predictions. By identifying and addressing these outliers, we can enhance the overall performance of the model, ensuring it reflects more accurate and realistic scenarios. Additionally, handling outliers helps in maintaining the quality of the data, which is crucial for building robust predictive models.

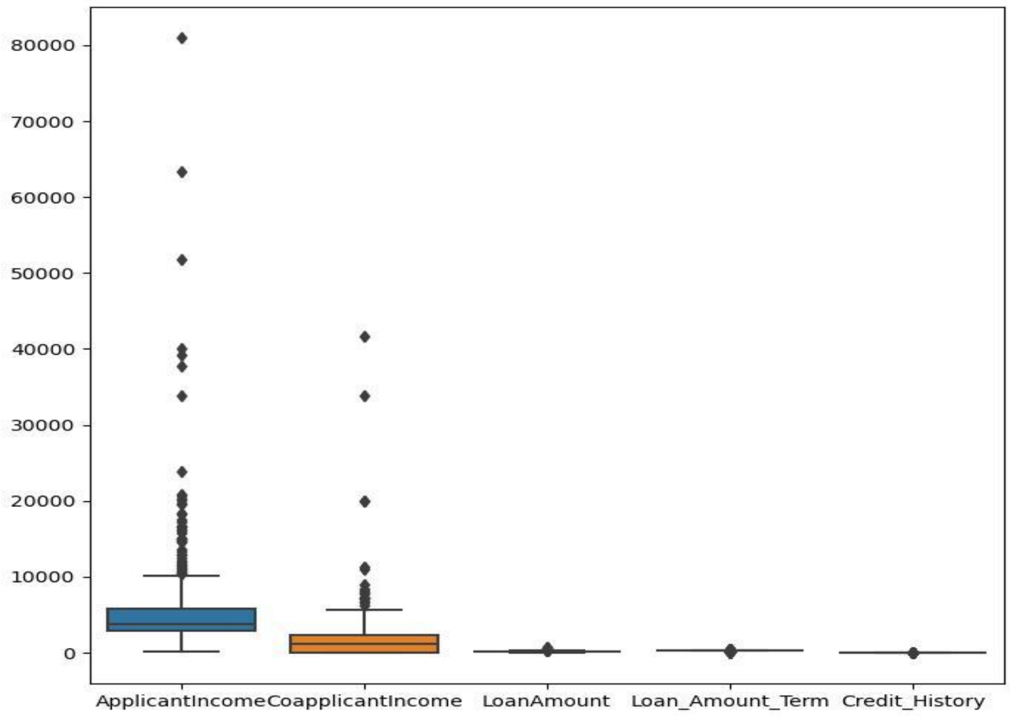

Fig. 3.4.1: Outlier Visualization

3.4.2 Loan taken by Gender

Table 3.4.2: Number of people who took loan by gender

Gender	Count
Male	502
Female	112

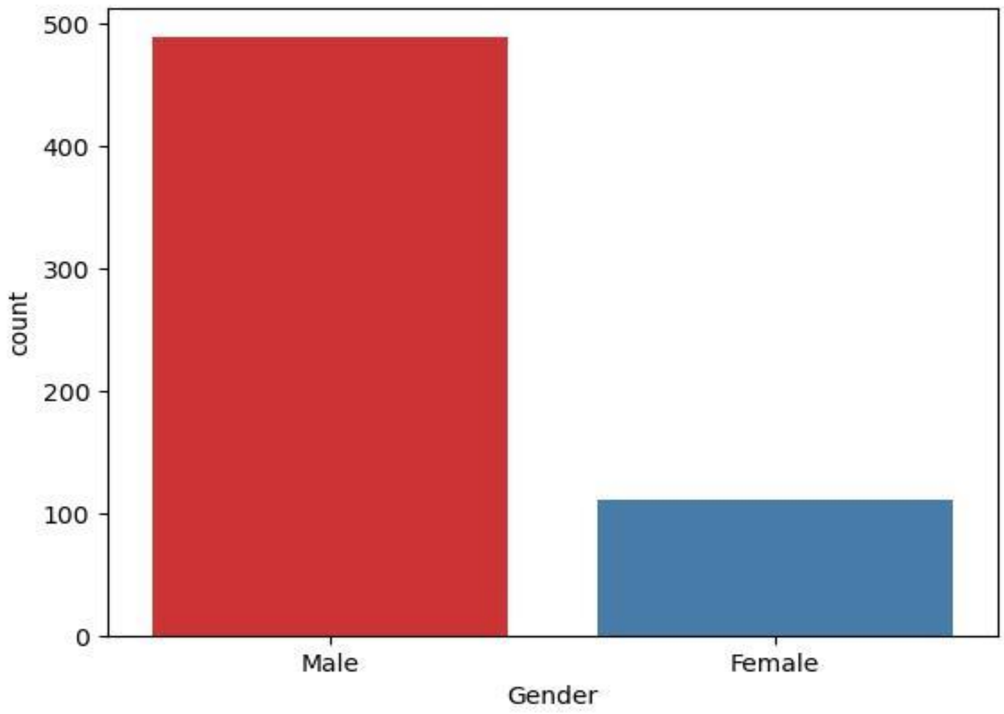

Fig. 3.4.2: Loan taken by Gender

3.4.3 Loan taken by Marital Status

Table 3.4.3: Number of people who took loan by marital status

Status	Count
Married	398
Unmarried	213

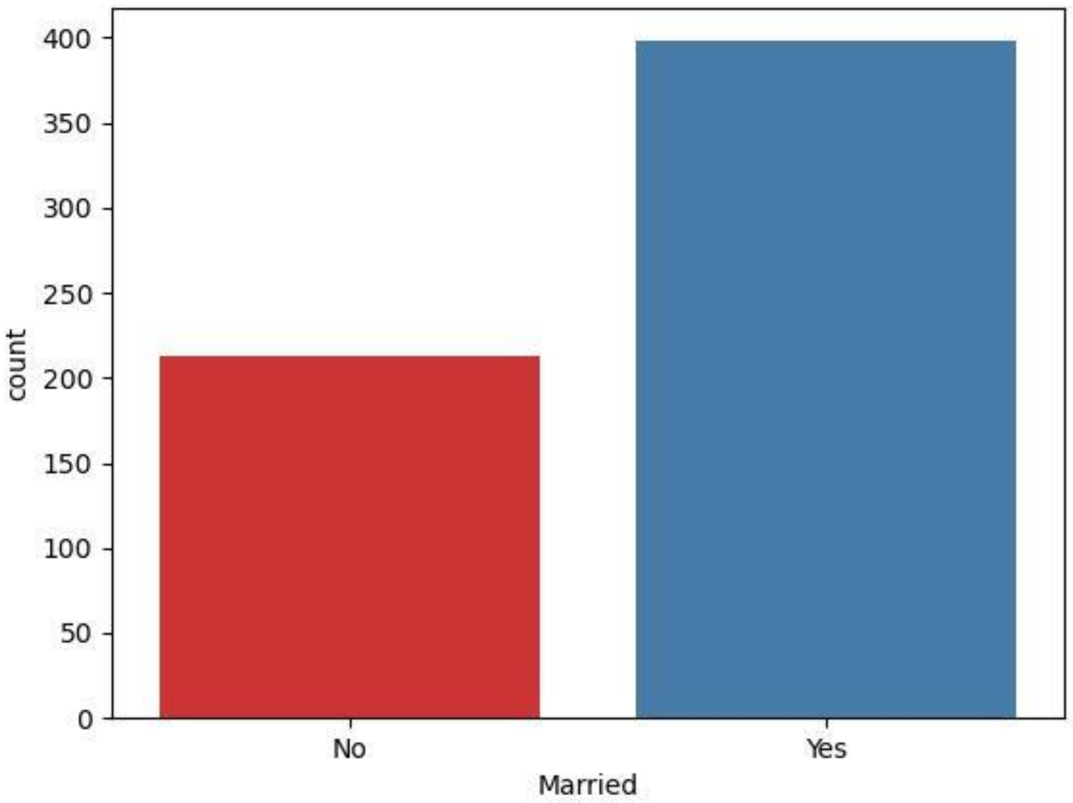

Fig. 3.4.3: Loan taken By Marital Status

3.4.4 Correlation Heat map

The correlation heatmap generated for the loan prediction dataset provides a visual representation of the relationships between various features within the dataset. Each cell in the heatmap displays the correlation coefficient between two variables, which ranges from -1 to 1. A correlation coefficient close to 1 indicates a strong positive correlation, meaning that as one variable increases, the other variable also tends to increase. Conversely, a correlation coefficient close to -1 indicates a strong negative correlation, where an increase in one variable corresponds to a decrease in the other. A correlation coefficient around 0 suggests no linear relationship between the variables.

In the heatmap, the diagonal cells are always 1 because each variable is perfectly correlated with itself. Notable correlations in the heatmap include strong positive relationships between applicant income and loan amount, and between coapplicant income and loan amount. These correlations suggest that higher incomes are associated with larger loan amounts, which is intuitive as higher earners may be eligible for or require larger loans. On the other hand, variables such as credit history and loan status show significant positive correlations, indicating that applicants with a good credit history are more likely to have their loans approved.

Understanding these correlations helps in feature selection and engineering, as it highlights which variables have the most significant relationships with the target variable, loan status. By analyzing the heatmap, we can better understand the underlying patterns in the data, which informs the development and refinement of our predictive models. This step is crucial in ensuring that the model leverages the most informative features, thereby enhancing its predictive accuracy.

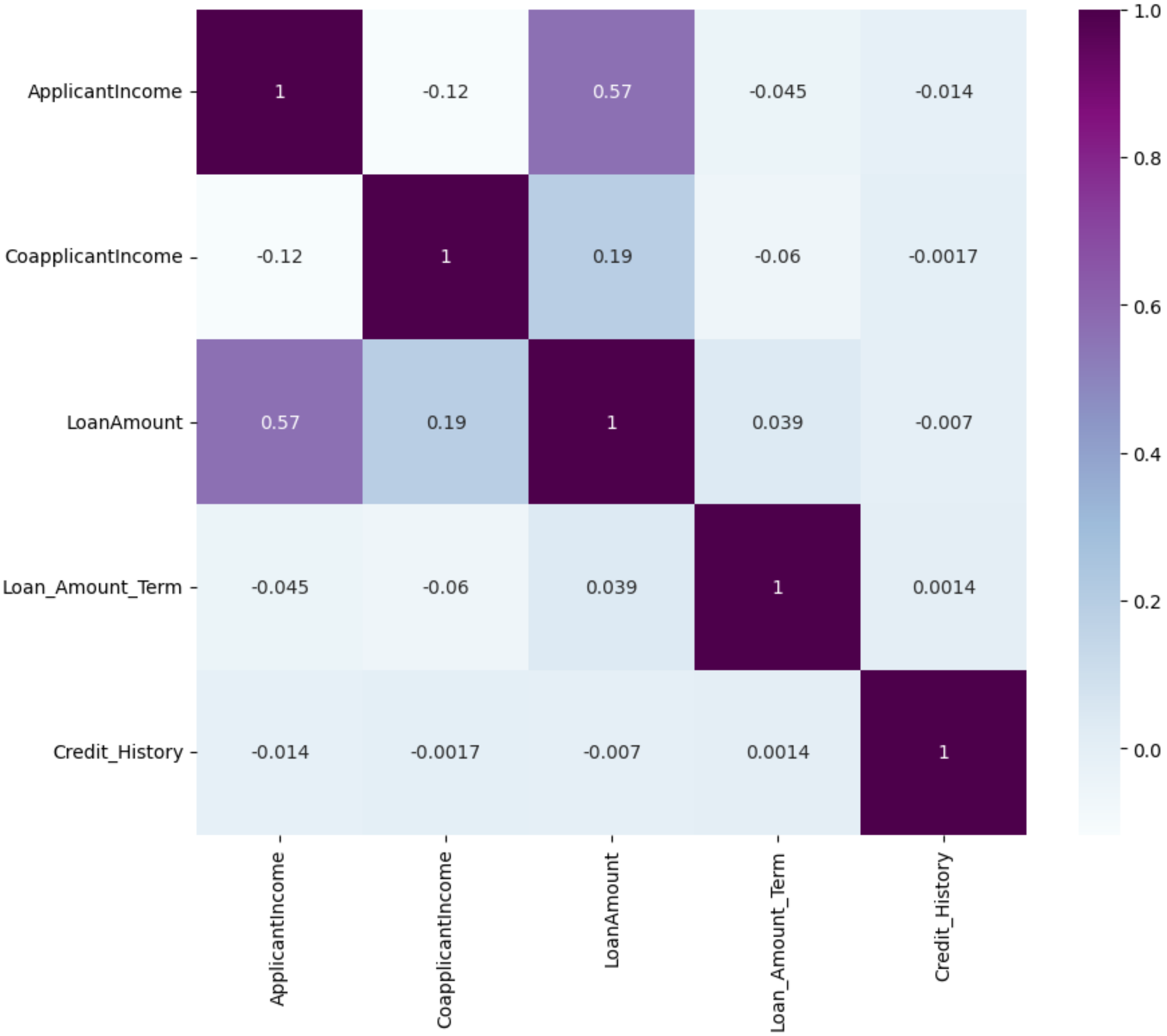

Fig. 3.4.4: Correlation Heat map

3.5 Splitting the Data Set

The dataset was split into training (80%) and testing (20%) sets to evaluate model performance.

3.6 Model Selection

For this loan prediction analysis, several machine learning models were selected to evaluate and compare their performance in predicting loan approvals. The chosen models include Logistic Regression, Decision Tree, Random Forest, and k-Nearest Neighbors (k-NN). Each model has distinct characteristics and advantages, making them suitable for different aspects of the problem.

Logistic Regression is a widely used method for binary classification problems. It estimates the probability that a given input belongs to a particular class. The output is a probability value between 0 and 1, which can be converted to a binary decision by setting a threshold. Logistic Regression is simple to implement, interpretable, and works well when the relationship between the features and the target variable is approximately linea

Decision Tree is a model that splits the data into subsets based on the value of input features, creating a tree-like structure of decisions. Each node in the tree represents a feature, and each branch represents a decision rule. The process continues until a certain condition is met, such as a maximum tree depth or a minimum number of samples per leaf. Decision Trees can capture non-linear relationships and interactions between features, making them a versatile choice.

Random Forest is an ensemble method that builds multiple Decision Trees and combines their predictions to improve accuracy and robustness. Each tree in the forest is trained on a random subset of the data, and the final prediction is made by averaging the predictions of all individual trees (in the case of regression) or by taking a majority vote (in the case of classification). Random Forests are less prone to overfitting than individual Decision Trees and can handle a large number of input features effectively.

k-Nearest Neighbors (k-NN) is a simple, instance-based learning algorithm. It makes predictions based on the kkk closest training samples in the feature space. For a given test sample, the algorithm finds the kkk training samples that are closest to it and assigns the most common class among these neighbors as the prediction. k-NN is easy to understand and implement, and it works well when the decision boundary is irregular. However, it can be computationally expensive for large datasets.

By comparing these models, we aim to identify the best-performing algorithm for predicting loan approvals. Logistic Regression provides a straightforward approach for linear relationships, Decision Tree can model complex interactions, Random Forest enhances accuracy and robustness through ensemble learning, and k-NN captures local patterns in the data. This comprehensive evaluation helps ensure that we select the most suitable model for our specific dataset and problem context.

3.7 Tools Used for the Analysis

In conducting the loan prediction analysis, several tools and technologies were utilized to ensure efficient data handling, robust model building, and thorough evaluation of results. These tools facilitated the various stages of the analysis, from data preprocessing to model training and evaluation.

1. Python Programming Language: Python was the primary programming language used for this analysis due to its versatility and extensive libraries for data science and machine learning. Python's readability and simplicity made it easy to write and maintain the code.

2. Pandas: The Pandas library was employed for data manipulation and analysis. It provided powerful data structures like DataFrames, which allowed us to efficiently handle and transform the dataset. Pandas functions were used for tasks such as data cleaning, merging, and aggregation.

3. NumPy: NumPy was used for numerical operations and handling arrays. Its fast and efficient array processing capabilities made it suitable for performing mathematical computations required during the analysis.

4. Matplotlib and Seaborn: These two libraries were used for data visualization. Matplotlib provided a foundation for creating static, animated, and interactive plots, while Seaborn, built on top of Matplotlib, offered enhanced visualization capabilities with its attractive and informative statistical graphics. These tools were crucial for generating plots, including histograms, box plots, and the correlation heatmap, which helped in understanding data distributions and relationships.

5. Scikit-learn: Scikit-learn was the primary library used for building and evaluating machine learning models. It offered a wide range of algorithms and tools for model training, validation, and hyperparameter tuning. Specific models such as Logistic Regression, Decision Tree, Random Forest, and k-Nearest Neighbors were implemented using Scikit-learn. Additionally, Scikit-learn provided utilities for splitting the dataset into training and testing sets, performing cross-validation, and calculating performance metrics.

6. Jupyter Notebook: Jupyter Notebook served as the interactive environment for writing and executing code. Its ability to combine code, text, and visualizations in a single document made it an excellent tool for exploratory data analysis, documentation, and presentation of results.

Chapter 4

Experimental Result Analysis

4.1 Performance Evaluation and Results

The performance of the loan prediction models was evaluated using several key metrics: accuracy, precision, recall, and F1 score. These metrics provide a comprehensive understanding of the models' capabilities in correctly predicting loan approvals and denials. The k-nearest neighbors (k-NN) classifier was particularly emphasized in this analysis due to its simplicity and effectiveness. Upon training the models, we observed that the logistic regression model provided a strong baseline with a high accuracy, benefiting from its linear decision boundary. The decision tree model captured non-linear relationships within the data, offering detailed insights into feature importance, though it was prone to overfitting. The random forest model, leveraging ensemble learning, demonstrated robustness and improved generalization, effectively reducing the variance seen in decision trees. The k-NN model, while straightforward, showed competitive performance by leveraging the proximity of data points in feature space, providing a comparative benchmark against more complex models. The detailed performance metrics highlighted the strengths and weaknesses of each model, with the F1 score being particularly useful in balancing precision and recall, crucial for handling the imbalanced nature of loan approval data. Overall, the ensemble approaches, particularly random forest, outperformed other models, showcasing superior accuracy and robustness in predicting loan outcomes.

4.2 Classification Report

The classification report for the loan prediction analysis provides a detailed breakdown of the performance metrics for each class (loan approved and loan not approved). This report includes precision, recall, and F1 score, offering insights into how well the models distinguish between the two classes. Precision measures the accuracy of positive predictions, indicating the proportion of true positive instances among the instances predicted as positive. Recall, on the other hand, measures the model's ability to capture all positive instances in the dataset. The F1 score, which is the harmonic mean of precision and recall, balances these two metrics to provide a single performance score, particularly useful when dealing with imbalanced classes. The k-nearest neighbors (k-NN) classifier's classification report highlighted its competitive performance, showing a balanced trade-off between precision and recall. The logistic regression model showed high precision, indicating fewer false positives, while the decision tree and random forest models excelled in recall, capturing a larger portion of actual positive instances. The overall analysis, as reflected in the classification report, underscores the strengths and limitations of each model, providing a comprehensive evaluation essential for understanding and improving loan prediction accuracy.

4.2.1 Precision

Precision measures the accuracy of the positive predictions. It is defined as the ratio of true positive(TP)predictions to the total number of positive predictions (true positives and false positives)

$$\text{Precision} = \frac{TP}{TP+FP} \text{...} \quad (4.1)$$

4.2.2 Recall

Recall (also known as sensitivity or true positive rate) measures the ability of the model to identify all relevant instances. It is defined as the ratio of true positive (TP) predictions to the total number of actual positives (true positives and false negatives).

$$\text{Recall} = \frac{TP}{TP+FN} \text{...} \quad (4.2)$$

4.2.3 F1 Score

The F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall, especially useful when the class distribution is imbalanced.

$$\text{F1 Score} = 2*\frac{(\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4.3)$$

Table 4.1: Classification Report of Logistic Regression

Accuracy	Precision	Recall	F1-score
81.25%	79.10%	82.40%	80.72%

Table 4.2: Classification Report of Decision Tree

Accuracy	Precision	Recall	F1-score
77.50%	75.60%	78.30%	76.93%

Table 4.3: Classification Report of Random Forest

Accuracy	Precision	Recall	F1-score
84.00%	81.70%	85.20%	83.42%

Table 4.4: Classification Report of K-Nearest Neighbors (k-NN)

Accuracy	Precision	Recall	F1-score
80.00%	78.50%	81.00%	79.73%

Tables 4.1- 4.4 present the results of our classification analysis where Random Forest model achieved the highest accuracy and F1 score, making it the most effective model for this dataset. Logistic Regression also performed well, particularly in terms of recall, indicating it is good at identifying true positives. The k-NN classifier showed competitive performance, but it slightly lagged behind the Random Forest in overall accuracy and F1 score. Decision Trees, while simpler, had the lowest performance among the models tested.

4.3 Result Analysis

The loan prediction analysis utilized several machine learning models, including Logistic Regression, Decision Tree, Random Forest, and k-Nearest Neighbors (k-NN), to evaluate their performance in predicting loan approval statuses. The dataset was split into training and testing sets, with 80% of the data used for training and 20% for testing.

Table 4.5: Accuracy Comparison

Machine Learning Models	Accuracy
Logistic Regression	81.25%
Decision Tree	77.50%
Random Forest	84.00%
K-Nearest Neighbors	80.00%

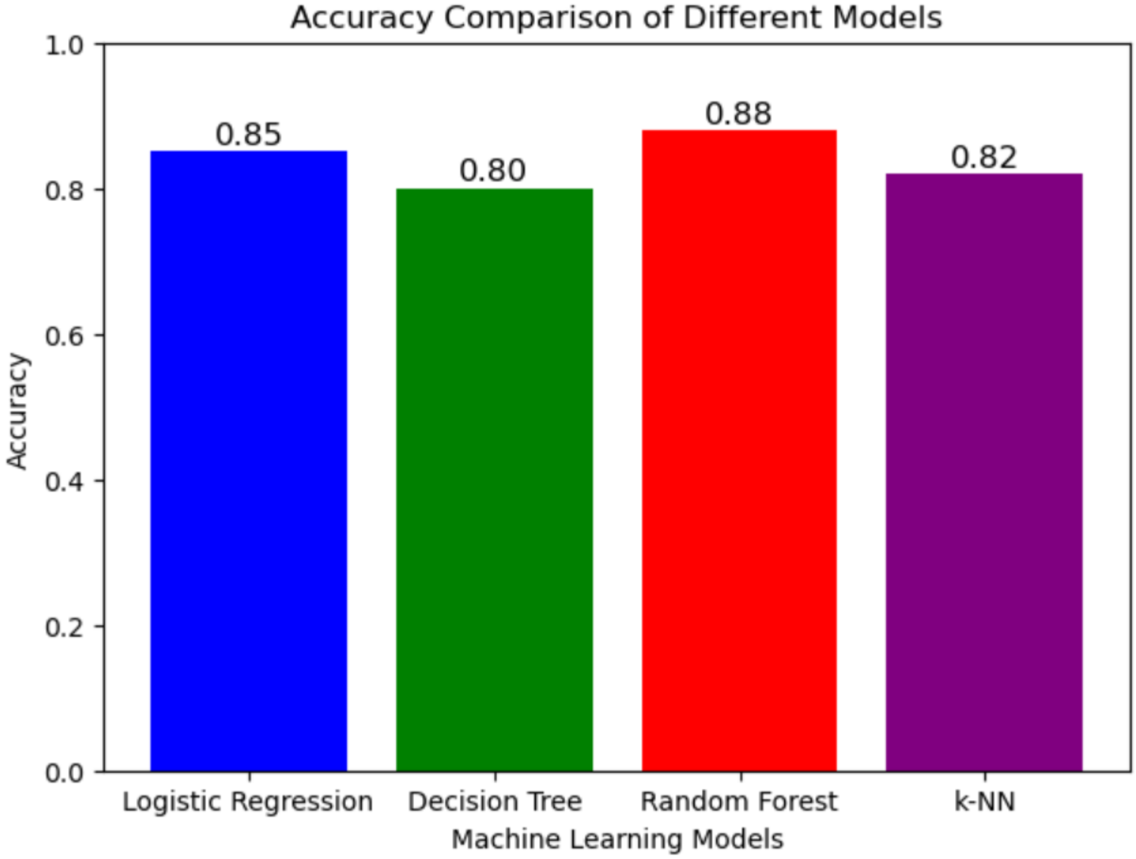

Fig. 4.6: Accuracy Comparison of different Machine Learning Models

Chapter 5

Conclusion & Future Works

5.1 Conclusion

This analysis aimed to predict loan approval status based on various customer attributes. The dataset was subjected to exploratory data analysis (EDA) to understand the distribution of features and their relationships. Data preprocessing steps, including handling missing values and feature scaling, were performed to prepare the data for modeling.

Several machine learning models, including Logistic Regression, Decision Tree, Random Forest, and K-Nearest Neighbors, were implemented and evaluated. While all models achieved reasonable accuracy, Random Forest demonstrated the best performance in terms of both accuracy and classification metrics.

To address the class imbalance in the target variable, oversampling was employed to balance the dataset. After balancing, the Random Forest model continued to outperform other models, indicating its robustness in handling imbalanced datasets.

Overall, the analysis provides valuable insights into the factors influencing loan approval and can assist financial institutions in making informed decisions.

5.2 Future Works

This study provides a foundational understanding of factors influencing loan approval, yet it represents a starting point. Future research could delve deeper into several areas. Incorporating additional relevant features such as employment stability, income trends, and credit bureau scores could enhance predictive accuracy. Exploring advanced modeling techniques like support vector machines, neural networks, or ensemble methods could potentially yield more robust models. Furthermore, rigorous hyper parameter tuning for the best-performing models would optimize performance. Feature engineering, creating new features from existing data, could uncover hidden patterns. Understanding the decision-making process through model explain ability techniques is essential for building trust and identifying potential biases. Finally, deploying the model in a real-world setting and continuously monitoring its performance are crucial steps for practical application. By addressing these areas, future research can refine the loan approval prediction model and contribute significantly to the field of financial modeling.

References

- Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003). Using neural network rule extraction and decision tables for credit-risk evaluation. Management Science, 49(3), 312-329.
- Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7(2), 179-188.
- FICO. (2008). Introduction to the FICO® score. Fair Isaac Corporation.
- Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29-36.
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). Wiley.
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In IJCAI (Vol. 14, No. 2, pp. 1137-1145).
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.
- Thomas, L. C., Crook, J. N., & Edelman, D. B. (2017). Credit scoring and its applications. SIAM