**Loan Prediction using Machine Learning**

**A Dissertation**

****

*Submitted to*

**Amity University, Kolkata**

*For the partial fulfillment of the award of the degree*

**BACHELOR OF SCIENCE  
DATA SCIENCE HONOURS**

By

**Rohit Agarwal**

**Enrollment Number: A914138122009**

Guide: Dr. Ajanta Das

**Amity Institute of Information Technology  
Amity University, Kolkata**

August 2024

DECLARATION

I hereby declare that the dissertation entitled “Loan Prediction Using Machine Learning” submitted by me in partial fulfillment of the requirements for the award of the Degree of B.Sc.[H] Data Science to the AMITY UNIVERSITY, KOLKATA is based on the experiments and studies carried out by me. This work is original and has not been submitted in part or full for any other degree or diploma of any university or institution.

Date: 12th November 2024

Place: EduFabrica, IIT Hyderabad

Rohit Agarwal

A914138122009

**Amity Institute of Information Technology  
Amity University, Kolkata**

CERTIFICATE

The research work embodied in this dissertation entitled “Loan Prediction Using Machine Learning” submitted by Rohit Agarwal, A914138122009 in partial fulfillment of the requirements for the award of the Degree of B.Sc.[H] Data Science to the AMITY UNIVERSITY, KOLKATA is based on the experiments and studies carried out by him. This work is original and has not been submitted in part or full for any other degree or diploma of any university or institution.

Date: 12th November 2024

Place: EduFabrica, IIT Hyderabad

Dr. Indraneel Mukhopadhyay

(Signature)

**Acknowledgement**

I would like to express my deepest gratitude to the EduFabrica for providing me with the opportunity to undertake an internship in the esteemed organization. Specifically, I appreciate the Data Science Using Python team for their guidance and support throughout my internship.

I extend my sincere thanks to Dr.Ajanta D, who patiently mentored me and shared valuable insights, enriching my learning experience.

I also thank Other Team Members or Colleagues for their encouragement and assistance during my tenure.

This internship has been an invaluable experience, allowing me to apply theoretical concepts to real-world problems and enhance my skills in machine learning using python.

Lastly, I thank Dr Indraneel Mukhopadhyay and other AIITK faculty members for facilitating this internship opportunity, contributing significantly to my academic and professional growth.

DATE:

NAME: Rohit Agarwal

ENROLLMENT NO: A914138122009

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| Subject | Chapter | Page Number |
| 1. Introduction | 1. Background | 1 |
|  | 1. Problem Statement | 1-2 |
|  | 1. Objectives | 2-3 |
|  | 1. Scope | 3-4 |
| 1. Dataset Information | 1. Dataset Description | 4-5 |
|  | 1. Data Visualization | 6 |
| 1. Methodology | 1. **Data Collection and Preparation** | 7 |
|  | 1. **Exploratory Data Analysis (EDA):** | 7 |
|  | 1. **Model Selection and Training** | 8 |
|  | 1. Hyperparameter Tuning | 8-9 |
|  | 1. **Model Evaluation** | 10 |
| 1. Result and Analysis | 1. Model Performance | 10 |
|  | 1. **Model Deployment** | 11 |
|  | 1. **Conclusion** | 11 |
| 1. Conclusion | - | 11 |

**List of tables**

|  |  |  |
| --- | --- | --- |
| **Table Number** | **Table Description** | **Page no.** |
| 2.1 | Dataset features | 4-5 |
| 4.1 | Model Performance Comparison | 10 |

**List of Figures**

|  |  |  |
| --- | --- | --- |
| Figure number | Figure description | Page number |
| 2.1 | Bar charts representing the dominating values as per the dataset | 6 |
| 3.1 | Confusion Matrix for Logistic Regression before hyperparametric tuning | 9 |
| 3.2 | Confusion matrix for logistic regression after hyperparametric tuing | 9 |

**Abstract:**

This project focuses on predicting loan approval decisions using machine learning techniques. The primary objective is to develop a predictive model that can accurately determine whether a loan should be approved or not based on applicant information. A variety of machine learning models, including Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forest, and Gradient Boosting, were applied to a dataset containing historical loan application records.

After extensive data preprocessing and feature engineering, these models were trained and evaluated using metrics such as F1 score, Recall, Accuracy, and ROC AUC. Among the models, the Random Forest classifier demonstrated the best performance, achieving the highest accuracy and robustness in prediction. The final model was saved for future deployment. This study highlights the potential of machine learning in automating and improving the loan approval process, offering financial institutions a reliable tool for decision-making.

In order to maximize its predictive power, it was then retrained on the entire dataset. The performance of each model is analysed in detail in this report, as well as an argument on how significant different features really are and some suggestions for improvements of machine learning approaches to do loan prediction.

**1. Introduction:**

**1.1. Background:**

Loan prediction is a critical function in the financial services industry, particularly for banks and other lending institutions. The ability to accurately predict whether a loan applicant will repay their loan is vital for maintaining the financial health of the institution and minimizing the risk of default. Traditional methods of loan assessment often rely on manual evaluations, which can be time-consuming and subject to human bias. However, with the advent of big data and machine learning, it is now possible to leverage historical data to make informed, data-driven decisions about loan approvals.

This project aims to build a machine learning model that can predict the likelihood of a loan being approved based on a variety of applicant features, such as credit history, income, loan amount, and other relevant factors. By automating the loan approval process with a predictive model, financial institutions can significantly enhance their decision-making process, ensuring that loans are granted to applicants who are most likely to fulfill their repayment obligations.

The study involves the use of several machine learning algorithms, including Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forest, and Gradient Boosting. The goal is to identify the most effective model for predicting loan approval by comparing their performance across several key metrics. This project demonstrates the potential of machine learning to revolutionize traditional banking processes, offering a more efficient and objective approach to loan evaluation.

### ****1.2. Problem Statement:****

In the financial sector, the process of evaluating loan applications is crucial for managing risk and ensuring the stability of lending institutions. However, traditional methods of loan evaluation often rely heavily on manual assessments, which can be prone to errors, inefficiencies, and biases. This not only slows down the approval process but also increases the risk of approving loans to applicants who may default, thereby affecting the institution's financial health.

The challenge is to develop a robust and reliable system that can predict the likelihood of a loan being approved or rejected based on the applicant's data. This system must be able to analyze various factors such as credit history, income levels, loan amount, employment status, and other pertinent features to make an informed decision.

The problem this project addresses is the need for an automated, data-driven approach to loan approval prediction. The goal is to create a predictive model that can accurately assess the risk associated with each loan application, thereby enabling financial institutions to make faster and more objective decisions. By leveraging historical loan data and advanced machine learning techniques, this project aims to reduce the risk of loan defaults and improve the efficiency of the loan approval process.

**1.3. Objectives:**

In this project, the aim is to come up with different machine learning models that can help predict whether or not a person is eligible for a loan . Specifically, the project will focus on:

1. Accurately Predict Loan Eligibility: Develop a machine learning model to reliably determine whether an applicant qualifies for a loan based on their profile data.
2. Identify Key Predictive Factors: Analyze the dataset to uncover the main factors that influence loan eligibility, such as income, credit score, employment status, etc.
3. Improve Decision-Making Efficiency: Create a tool to help financial institutions assess loan applications faster and more consistently, reducing manual review time.
4. Minimize Default Risk: Aid in reducing loan defaults by identifying high-risk applicants, potentially saving financial institutions from granting loans to unqualified applicants.
5. Enhance Model Accuracy and Reliability: Ensure the model's predictions are highly accurate, with minimized false positives (incorrectly approving a loan) and false negatives (incorrectly rejecting a loan).
6. Ensure Model Transparency: Provide insights into the model's decision-making process to ensure that it is interpretable and that applicants understand the factors affecting their eligibility.
7. Optimize Operational Efficiency: Use predictive modeling to streamline the loan approval process, allowing more resources to be dedicated to complex cases or other priorities.
8. Support Regulatory Compliance: Ensure that the model adheres to legal and ethical guidelines for fair lending, avoiding biases in the loan eligibility process.
9. Facilitate Deployment of a Scalable Solution: Build a model that can be deployed as a scalable solution, handling varying volumes of applications with consistent accuracy.

**1.4. Scope:**

The scope of this project encompasses the development and evaluation of a machine learning model for predicting loan approvals. The project is structured into several key phases, each focusing on different aspects of the predictive modeling process.

1. Data Exploration: Analyze and clean the dataset, identifying and handling any missing or outlier values, and understanding key patterns that might affect loan eligibility.
2. Feature Engineering: Extract and select relevant features from the dataset (e.g., income, credit history, debt-to-income ratio) that may impact eligibility, transforming them if necessary to improve model performance.
3. Model Selection: Evaluate various machine learning algorithms, such as logistic regression, decision trees, random forests, and gradient boosting, to determine the most effective model for predicting loan eligibility.
4. Model Training and Testing: Train and validate models using cross-validation or train-test splits, and evaluate their accuracy and efficiency in predicting loan eligibility.
5. Hyperparameter Tuning: Optimize the selected model’s hyperparameters to enhance its predictive accuracy and robustness.
6. Performance Metrics: Define and use appropriate metrics (e.g., accuracy, precision, recall, F1-score) to evaluate model performance, ensuring reliable results.

**2. Dataset Information**

**2.1. Data Description**

The dataset for predicting loan eligibility is gathered from the Kaggle public repository. Thirteen (13) characteristics and 614 records build up the dataset. Loan id, Gender of the applicant, marital status of the applicant, dependents in the family, qualification, employment status, applicant’s income, co-applicant income,

loan amount and term, credit worthiness, property area, and status of loan are the features present in the dataset. The table below shows all the features with their description.

**Table 2.1: Dataset features**

|  |  |  |
| --- | --- | --- |
| **Subject** | **Description** | **Impact on Loan Prediction** |
| Loan\_ID | Unique identifier for each loan application | No direct impact, used as identifier only |
| Gender | Applicant's gender (Male/Female) | Moderate impact, potential for minor bias |
| Married | Applicant's marital status (Yes/No) | Moderate impact, often associated with stability |
| Dependents | Number of dependents the applicant has | Moderate impact, affects financial capacity |
| Education | Applicant's education level (Graduate/Not Graduate) | Moderate impact, higher education may favor eligibility |
| Self\_Employed | Whether the applicant is self-employed (Yes/No) | Moderate impact, income stability varies |
| **Subject** | **Description** | **Impact on Loan Prediction** |
| ApplicantIncome | Applicant's income | High impact, determines repayment capability |
| CoapplicantIncome | Co-applicant's income | High impact, increases repayment capacity |
| LoanAmount | Amount of loan requested | High impact, affects loan approval based on capacity |
| Loan\_Amount\_Term | Term of the loan in months | Moderate impact, longer terms may affect eligibility |
| Credit\_History | Whether applicant has credit history (1.0/0.0) | High impact, strong indicator of eligibility |
| Property\_Area | Location of the property (Urban/Rural/Semiurban) | Moderate impact, may influence risk assessment |
| Loan\_Status | Target variable indicating if loan was approved (Y/N) | - Target variable for prediction |

**2.2. Data Visualisation**

Data visualization plays an important role in understanding the patterns and relationships between the features in the dataset. Through visual tools like histograms, box plots, scatter plots, and pair plots, we can identify trends, outliers, and correlations among the variables. To Visualize all the unique values in columns using barplot, this will simply show which value is dominating as per our dataset.

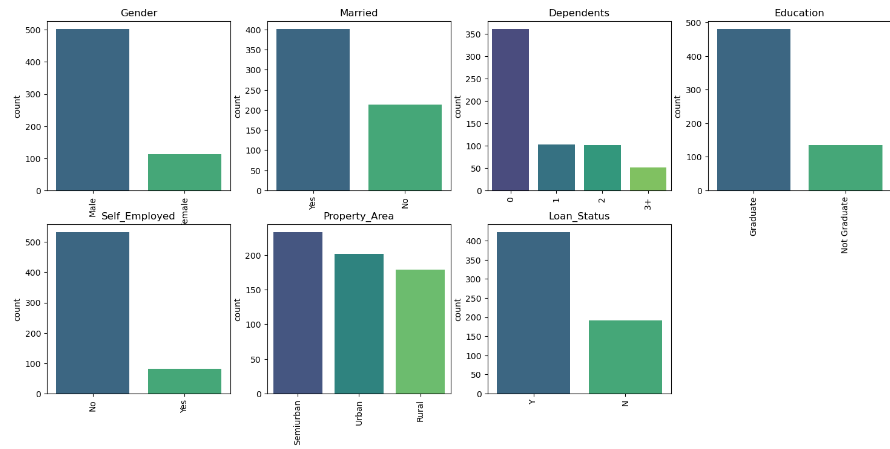


Fig 2.1. Bar charts representing the dominating values as per the dataset

### ****3. Methodology:****

The methodology of this project is structured to systematically approach the development of a machine learning model for predicting loan approvals. The process involves several key steps, from data collection and preprocessing to model selection, training, evaluation, and deployment.

#### 3.1. ****Data Collection and Preparation:****

* **Data Source:** The dataset used in this project was sourced from a publicly available loan prediction dataset. It contains information about previous loan applicants, including features like Credit\_History, ApplicantIncome, LoanAmount, Gender, and others.
* **Data Cleaning:** The dataset underwent cleaning to address missing values, which were imputed using techniques such as filling with the median, mean, or mode depending on the feature. The Loan\_ID column was dropped as it was not relevant for prediction.
* **Feature Encoding:** Categorical variables were encoded using techniques like label encoding to convert them into numerical format, making them suitable for model input.
* **Feature Scaling:** Numerical features were scaled using standardization or normalization techniques to ensure that the models could perform optimally without being biased by the scale of the features.

#### 3.2. ****Exploratory Data Analysis (EDA):****

* **Data Visualization:** Various visualization techniques were employed to understand the distribution of features and their relationships. Bar plots, pair plots, and heatmaps were used to identify correlations and patterns in the data.
* **Insights:** EDA helped in identifying key features that are strongly correlated with the loan approval status, such as Credit\_History and ApplicantIncome. These insights guided the selection of features for model training.

#### 3.3. ****Model Selection and Training:****

* **Model Selection:** Several machine learning models were selected for this project, including:
  + **Logistic Regression:** A simple and interpretable model often used for binary classification tasks.
  + **Support Vector Machines (SVM):** Known for its effectiveness in high-dimensional spaces.
  + **Decision Trees:** A model that splits the data based on feature values to make predictions.
  + **Random Forest:** An ensemble method that builds multiple decision trees and merges them to get a more accurate and stable prediction.
  + **Gradient Boosting:** Another ensemble method that builds models sequentially, correcting the errors of the previous models.

**3.4. Hyperparameter Tuning:**

Hyperparameter tuning is the process of optimizing the settings (hyperparameters) that govern the behaviour of a machine learning model, such as learning rate, regularization strength, or maximum tree depth. Proper tuning helps balance bias and variance, improving a model’s generalization and performance on unseen data.

Hyperparameter tuning was performed to optimize the performance of various models: Decision Tree, Random Forest, Logistic Regression, and SVM. `GridSearchCV` was used to systematically search for the best combination of hyperparameters by evaluating the model’s performance on each parameter set. For each model, a parameter grid was specified, and the models were tuned based on specific evaluation metrics, such as recall or accuracy.

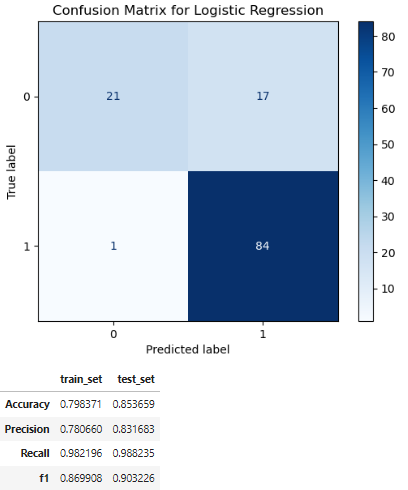
An example showcasing hyperparameter tuning: 

Fig 3.1: Confusion Matrix for Logistic Regression before hyperparametric tuning

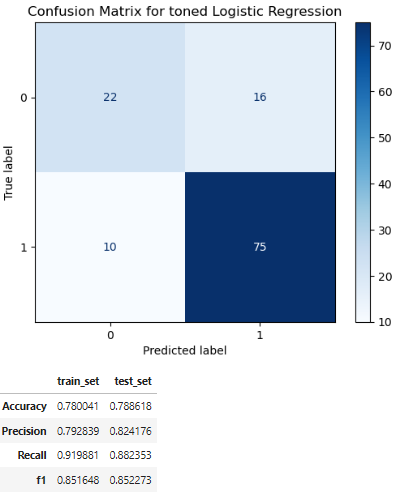


Fig 3.2: Confusion matrix for logistic regression after hyperparametric tuing

3.5. **Model Evaluation:**

* **Evaluation Metrics:** The models were evaluated using the following metrics:
  + **F1 Score:** A measure of a model's accuracy that considers both precision and recall.
  + **Recall:** The ability of a model to identify all relevant cases (i.e., all actual positive instances).
  + **Accuracy:** The overall correctness of the model in predicting loan approvals.
  + **ROC AUC:** The Area Under the Receiver Operating Characteristic curve, which measures the model's ability to distinguish between classes.

**4. Result and Analysis**

**4.1. Model performance:**

The performance metrics for each model—Logistic Regression, Decision Tree, Random Forest and Support Vector Machine (SVM) are summarized in the table below:

Table 4.1: Model Performance Comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | F1 | Recall | Accuracy | ROC AUC |
| Logistic Regression | 85.2% | 91.8% | 78.9% | 73.1% |
| Decision Tree | 86.7% | 92% | 80.5% | 73.5% |
| Random Forest | 90.3% | 98.8% | 85.4% | 77.0% |
| Support Vector Machine | 90.3% | 98.8% | 85.4% | 77.0% |

#### 4.2. ****Model Deployment:****

* **Saving the Model:** The final Random Forest model was serialized using the pickle library, making it ready for deployment. This model can be integrated into a real-world system to automate the loan approval process.

#### 4.3 ****Conclusion:****

* The methodology applied in this project enabled the creation of a robust loan prediction model. By systematically approaching data preprocessing, model selection, and evaluation, the project successfully identified a highly effective model that can be used to support loan approval decisions.

### ****Conclusion:****

This project successfully demonstrated the application of machine learning techniques to predict loan approvals, providing a data-driven approach to a traditionally manual and subjective process. By leveraging historical loan application data, several machine learning models were developed and evaluated, with the Random Forest classifier emerging as the most effective model. This model exhibited strong performance across various metrics, including accuracy, F1 score, recall, and ROC AUC, making it a reliable tool for predicting loan approval outcomes.

The project began with thorough data preprocessing and exploratory analysis, which were crucial in understanding the data and preparing it for model training. Feature engineering, scaling, and encoding were performed to ensure that the models could learn from the data effectively. The use of GridSearchCV for hyperparameter tuning further optimized the models, leading to better performance.

The final model was not only accurate but also robust, with the ability to generalize well to new, unseen data. This highlights the potential of machine learning to improve the efficiency and accuracy of loan approval processes in financial institutions. The model was saved and can be deployed to automate the decision-making process, reducing the risk of human error and bias.

In conclusion, this project showcases the power of machine learning in enhancing financial decision-making. It provides a strong foundation for further refinement and implementation in real-world scenarios, where the integration of such models could lead to significant improvements in the speed and reliability of loan processing. Future work could involve expanding the dataset, incorporating additional features, or exploring more advanced machine learning techniques to further enhance model performance