Loan Approval Prediction using Machine Learning

Sayan Rana

Postgraduate Department of Data Science
St. Xavier's College (Autonomous) - Kolkata
Supervisor's Name:- Dr. Ayan Chandra

Table of Content

Introduction

- In today's fast-paced world, loan approval is a crucial service provided by banks and financial institutions to ensure timely and accurate decisions.
- Traditional manual methods relied heavily on credit scores, income, and employment details. These approach were quite subjective, time-consuming, and inefficient for large volumes of applications.
- The advent of Machine Learning (ML) revolutionized the loan approval process by enabling automation and uncovering hidden patterns in historical data to predict loan repayment likelihood.
- ML models improve the speed, accuracy, and objectivity of loan decisions by analyzing complex interactions between borrower characteristics.

Data Description

The dataset has been downloaded from Kaggle which comprises 598 observations & 13 variables. The target variable is Loan_Status which helps to determine whether an applicant's loan can be approved or denied based on the features (predictors). The variables include:-

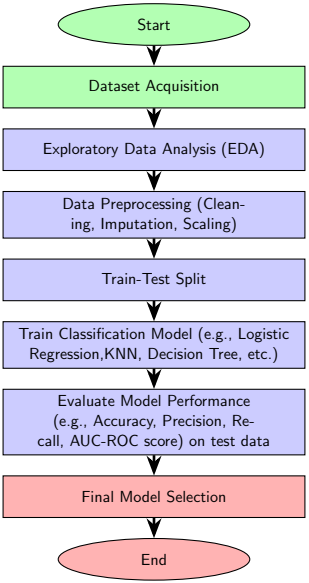
Applicant Information

1. Gender
2. Married
3. Dependents
4. Education
5. Property_Area
6. Applicant_Income

Loan Information

1. Loan_ID
2. LoanAmount
3. Loan_Amount_Term
4. CoApplicantIncome
5. Self_Employed
6. Credit_History
7. Loan_Status

Pipeline for ML algorithm



Exploratory Data Analysis (EDA)

Exploratory Data Analysis is being performed to visualize key patterns in the data.

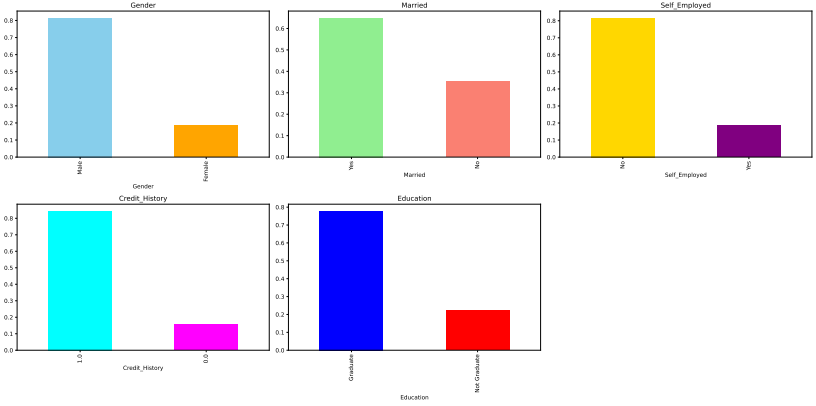


Figure: Visualization of Categorical Features

Exploratory Data Analysis (EDA) contd.

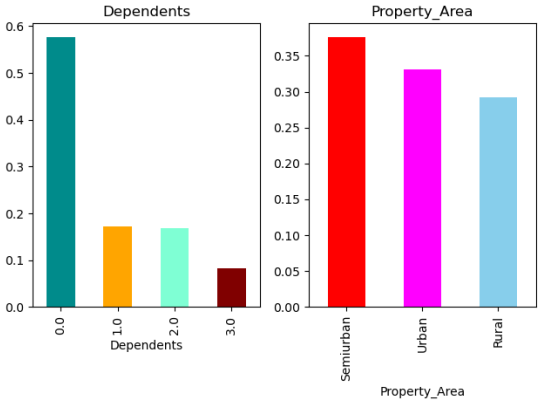
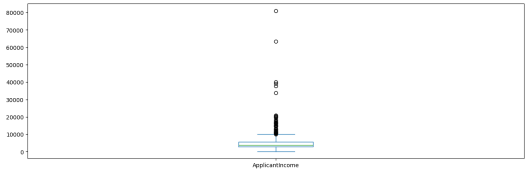


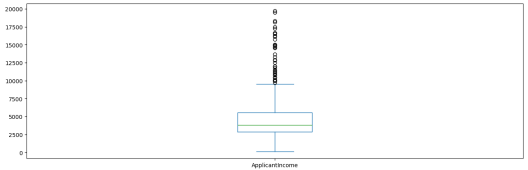
Figure: Visualization of Categorical Features contd.

Data Preprocessing

The vital step before performing any Machine Learning (ML) algorithm is data pre-processing. It involves handling of missing data, checking for outliers, and encoding categorical variables into a numeric format, which enhances less memory utilization.



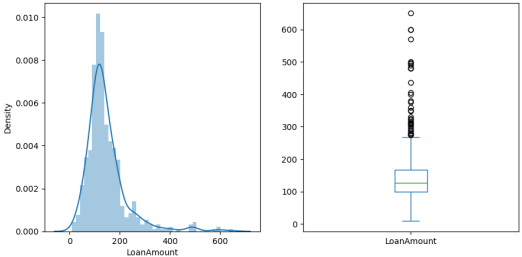
(a) Before pre-processing ApplicantIncome



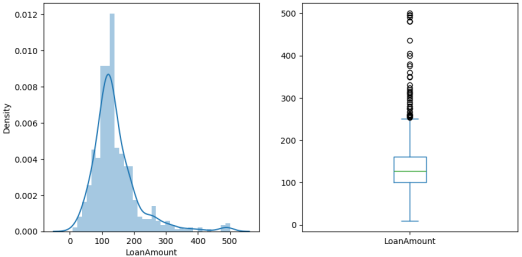
(b) After pre-processing ApplicantIncome

Figure: Comparison of ApplicantIncome before and after pre-processing

Data Preprocessing



(a) Before pre-processing Loan Amount



(b) After pre-processing Loan Amount

Figure: Comparison of Loan Amount before and after pre-processing

Data Preprocessing

Checking for Multicollinearity:-

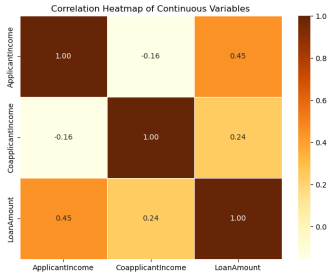


Figure: Heatmap for continuous variables

Remark: The heatmap indicates low multicollinearity among the variables, as none of the pairwise correlations are close to 1 or -1. This suggests that the variables can be used together in a classification model without serious multicollinearity issues.

Data Preprocessing

Encoding Categorical Variables to Numeric Format:-

Column_Name	Original	Encoded
Gender	Male	1
	Female	0
Married	No	0
	Yes	1
Education	Graduate	0
	Not Graduate	1
Self_Employed	No	0
	Yes	1
Property_Area	Urban	2
	Rural	0
	Semiurban	1
Loan_Status	Y	1
	N	0

Model Training

The dataset is divided into two parts: Training (70%) & Test (30%) sets respectively. Various Machine Learning algorithms have been trained on the training data independently. These include:

1. **Logistic Regression**
2. **K-Nearest Neighbor Classifier (KNN)**
3. **Decision Tree**
4. **Random Forest**
5. **Naïve Bayes Classifier**
6. **Support Vector Classifier (SVC)**

Model Evaluation

Since it is a binary classification problem, **Accuracy** & **Area Under the Receiver Operating Characteristic Curve (AUC-ROC)** score are used as metrics to evaluate which model performs best. The AUC-ROC curve for various ML models is given below:-

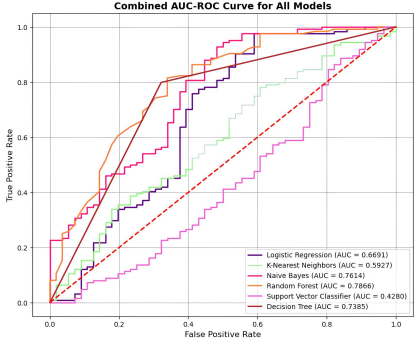


Figure: AUC-ROC Curve for different models

Model Performance Comparison

Method	Accuracy Score (%)	AUC-ROC Score
Logistic Regression	79.44	0.6691
K-Nearest Neighbor Classifier	69.44	0.5927
Decision Tree	76.11	0.7385
Random Forest Classifier	76.67	0.7866
Support Vector Classifier	68.89	0.4280
Naïve Bayes Classifier	81.11	0.7614

Table: Performance of Different Classification Models

Conclusion

This dissertation concludes that **machine learning** techniques can effectively automate and optimize the loan approval process. A detailed data preparation strategy, including handling missing values, detecting *outliers*, and checking for **multicollinearity**, ensured robust modeling. Multiple classification algorithms were evaluated using **accuracy** and **AUC-ROC** metrics. The **Naïve Bayes Classifier** emerged as the best performer with 81% *accuracy* and a strong *AUC-ROC* score. The results show that machine learning improves efficiency, accuracy, and reduces human bias in loan approval. Implementing such models can greatly improve *institutional risk assessment*, *decision-making*, and *overall resource management* in the financial services sector.

Future Scope

Naïve Bayes Classifier provides the best accuracy with an *accuracy* score of 81% for the test data. In order to obtain better results, other ensemble techniques like **Bagging** and **Boosting** can also be implemented.

References

- Breiman, L. (2001). *Random Forests*. *Machine Learning*, 45, 5–32. https://doi.org/10.1023/A:1010933404324
- Adankon, M., Cheriet, M. (2009). *Support Vector Machine*. In: Li, S.Z., Jain, A. (eds) *Encyclopedia of Biometrics*. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-73003-5_299
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.).
- GeeksForGeeks. https://www.geeksforgeeks.org/
- Datacamp. https://www.datacamp.com/
- Kaggle (for datasets). https://www.kaggle.com/
- Medium. https://medium.com

Thank You