Loan Approval Prediction Using Machine Learning

NAME : VIGNESH MADIVALA

Abstract

This project builds a machine learning model to predict whether a loan application will be approved or rejected based on applicant details. The dataset contains both categorical and numerical features, such as gender, marital status, income, loan amount, and credit history. Various preprocessing techniques were applied to handle missing values and encode categorical variables. Multiple machine learning models, including Logistic Regression, Decision Tree, Random Forest, and XGBoost, were trained and evaluated. The Random Forest Classifier achieved the highest accuracy of approximately 81%, making it the best-performing model.

Introduction

Financial institutions require fast and accurate tools to evaluate loan applications. Traditional manual methods are prone to delays and inconsistencies. Machine learning provides an automated, data-driven approach to predict loan approval outcomes. This project aims to:

1. Preprocess applicant data for training ML models.
2. Train multiple classifiers to predict loan approval status.
3. Evaluate and compare model performance.

Dataset Description

- Source: Loan dataset (CSV file).
- Features:
    - Categorical: Gender, Married, Education, Self_Employed, Property_Area.
    - Numerical: ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term, Credit_History.
- Target: Loan_Status (Approved = 1, Not Approved = 0).

Preprocessing Steps

1. Handled missing values using imputation.
2. Encoded categorical variables using LabelEncoder and pd.get_dummies.
3. Split dataset into training (80%) and testing (20%) sets.
4. Normalization/scaling was not applied (since tree-based models are scale-independent).

Models Implemented

- Logistic Regression → Baseline linear model.
- Decision Tree Classifier → Rule-based tree splits.
- Random Forest Classifier → Ensemble of decision trees (best performance).
- XGBoost Classifier (optional) → Gradient boosting method.

Evaluation Metrics

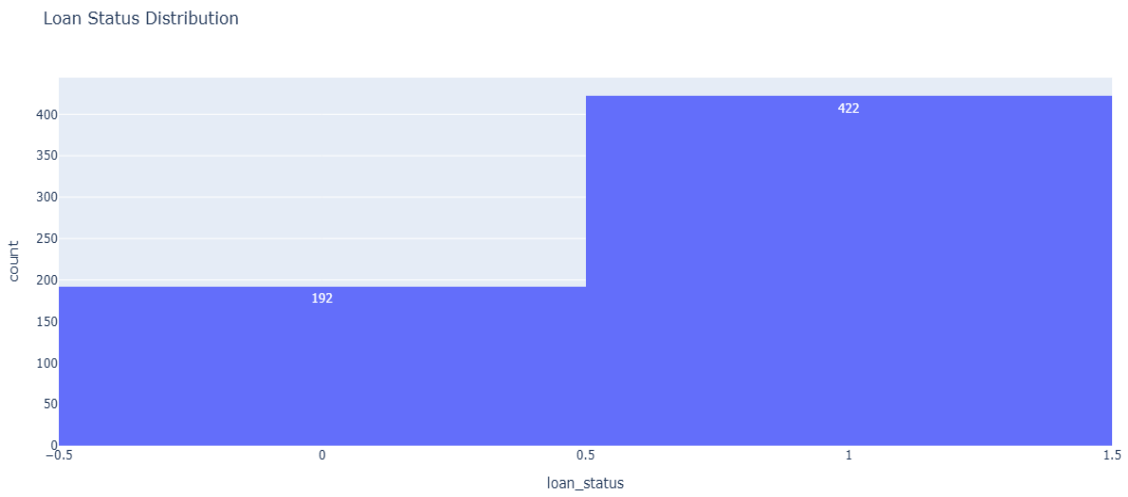
- Accuracy Score
- Confusion Matrix
- Classification Report (Precision, Recall, F1-score)

Results

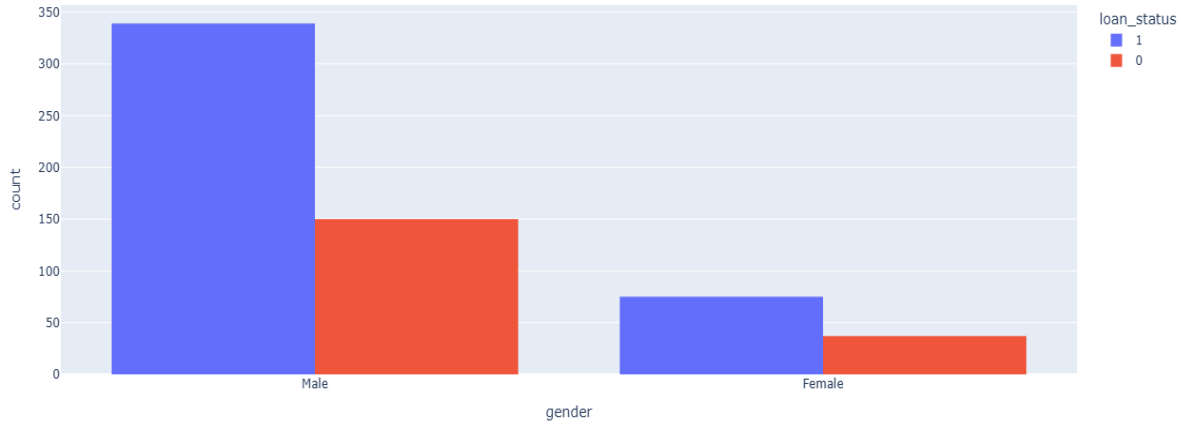
Model	Accuracy	Notes
Logistic Regression	~77%	Baseline performance
Decision Tree	~79%	Slightly better than Logistic
Random Forest	~81%	Best-performing model
XGBoost	~80%	Close to Random Forest

Visualizations

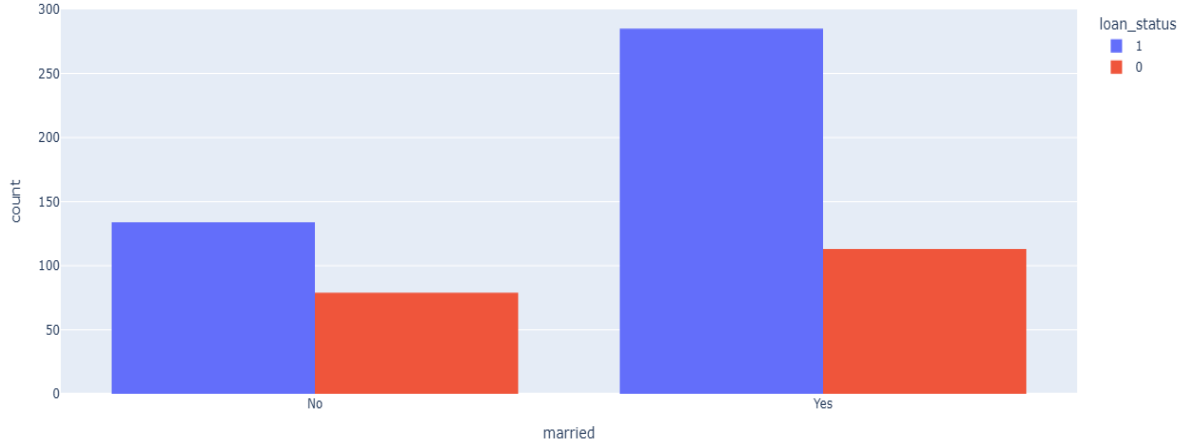
1. EDA Graphs: Countplots for categorical variables



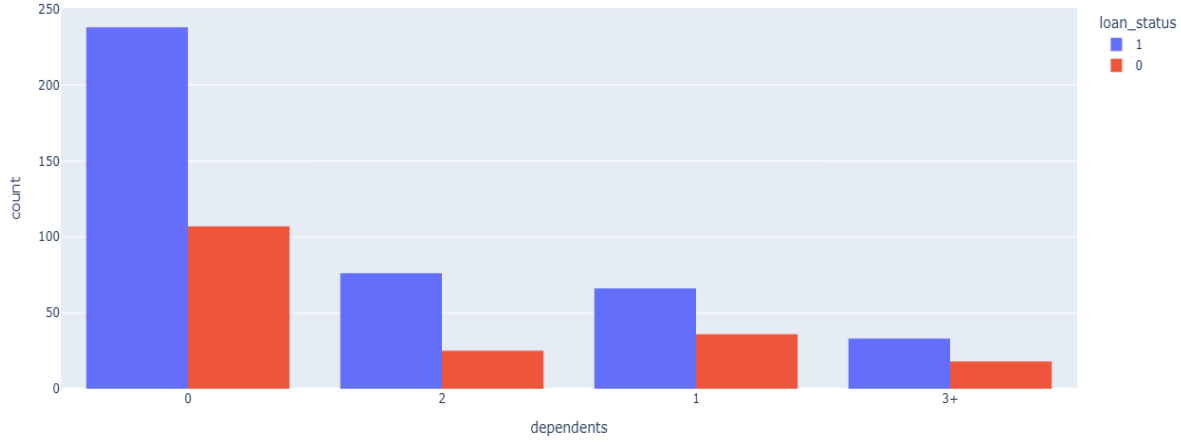
gender vs Loan Status



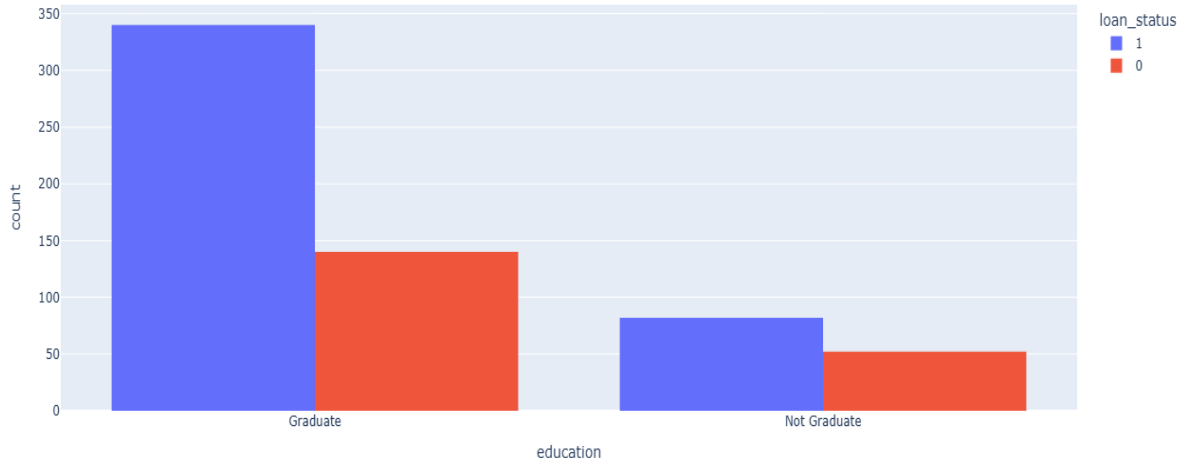
married vs Loan Status



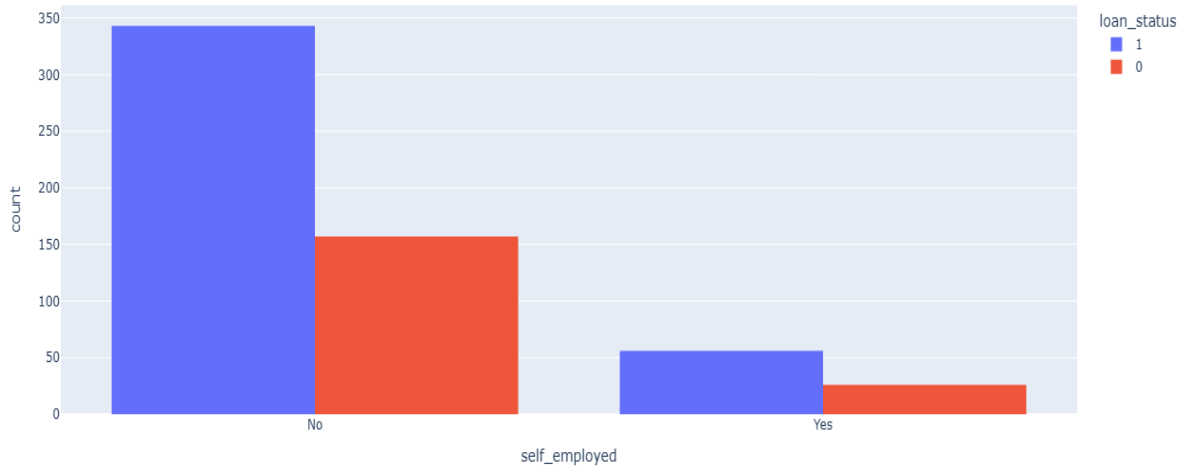
dependents vs Loan Status



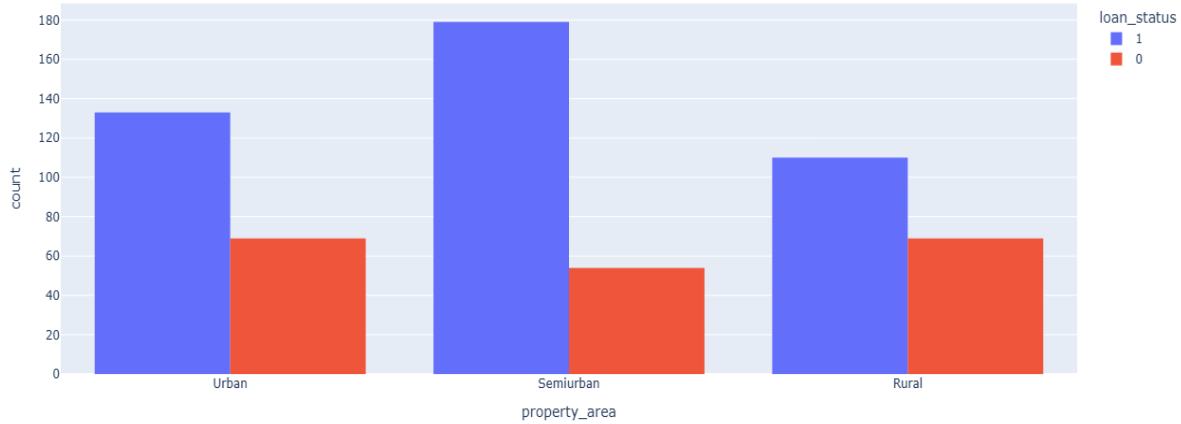
education vs Loan Status



self_employed vs Loan Status

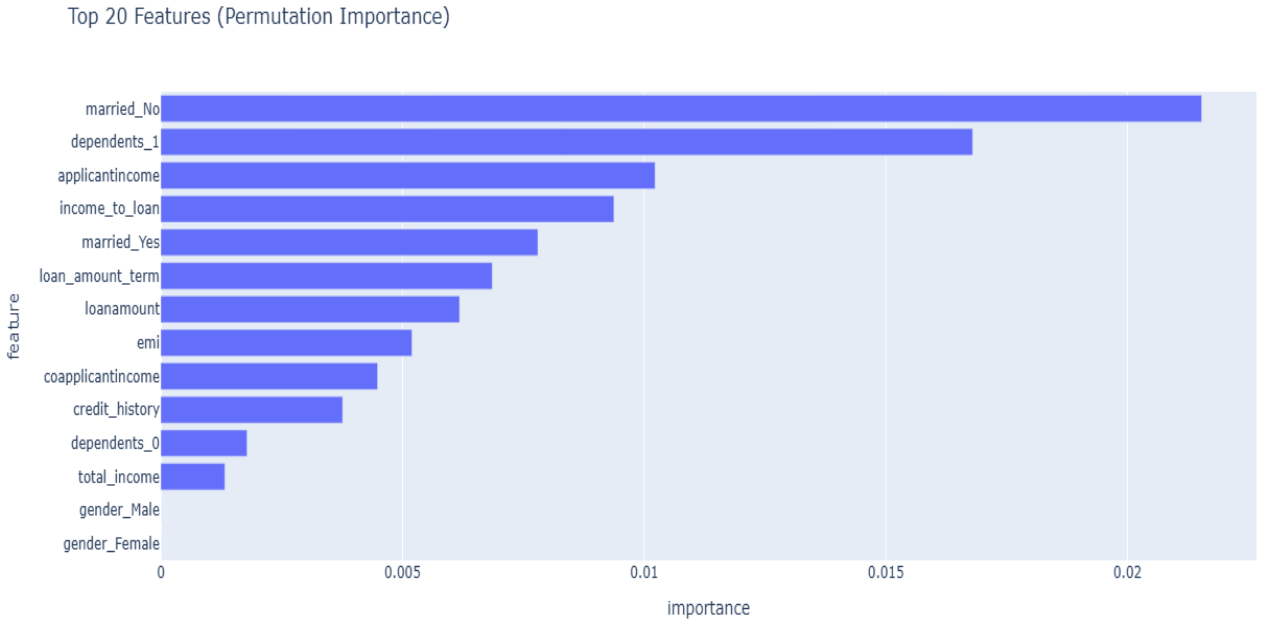


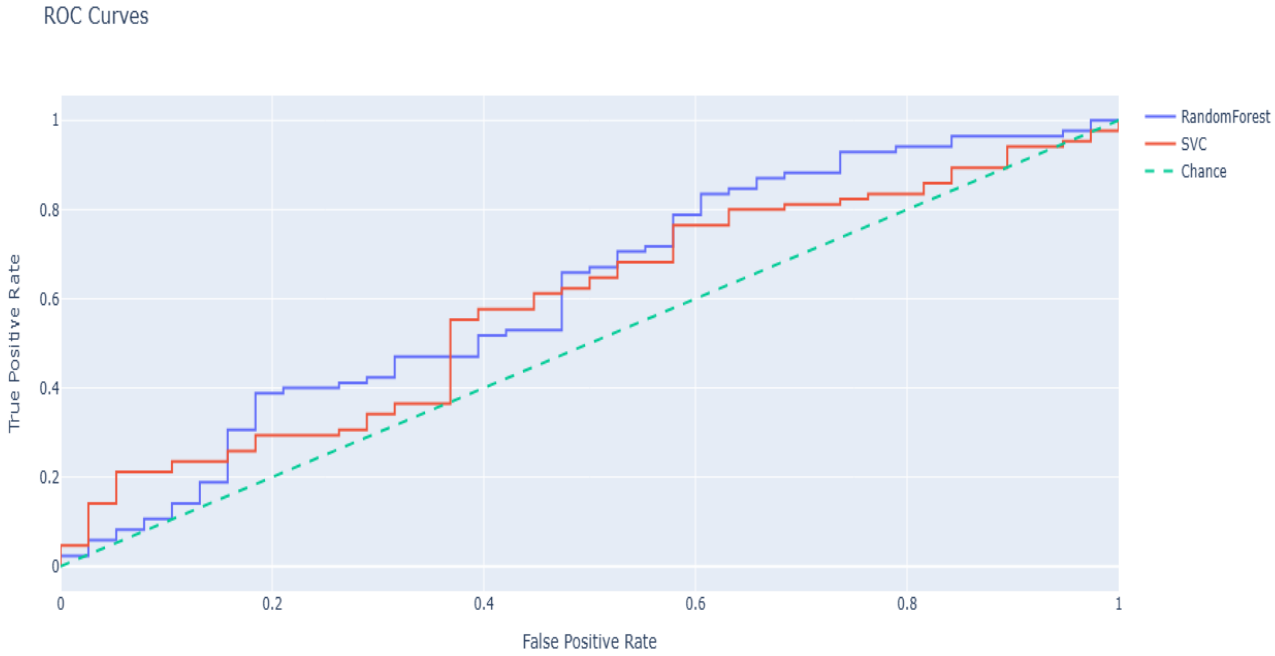
property_area vs Loan Status



2. Heatmap: Correlation between numerical features.

3. Optional: Add ROC curve and Feature Importance plot for Random Forest.





Conclusion

The project demonstrates the effectiveness of machine learning in predicting loan approvals. Random Forest provided the highest accuracy (81%), outperforming Logistic Regression and Decision Tree models. The model can serve as a decision-support tool for banks, though improvements can be made with cross-validation, hyperparameter tuning, and feature engineering.