Project Report: Loan Default Prediction

Main Objective of the Analysis

The objective of this analysis is to **predict loan default** using historical data on loan applicants. Our goal is to build a classification model that can determine whether an applicant is likely to default, which is essential for financial institutions to mitigate risk and make informed lending decisions. This model will help automate approval pipelines, minimize losses, and improve the bank's credit strategy.

Our focus will be on **predictive performance** while balancing **interpretability**, so stakeholders can trust and act on the predictions.

Description of the Dataset

We used the publicly available **Loan Prediction Dataset** from Analytics Vidhya. It contains records of loan applicants along with attributes such as gender, marital status, education, income, credit history, and whether or not their loan was approved.

Dataset Summary

Attribute	Description
Gender	Applicant's gender (Male/Female)
Married	Marital status
Dependents	Number of dependents
Education	Graduate or not
Self_Employed	Employment type
ApplicantIncome	Applicant's income
CoapplicantIncome	Co-applicant's income
LoanAmount	Requested loan amount
Loan_Amount_Term	Loan term (in months)
Credit_History	Credit score history
Property_Area	Urban/Rural/Semiurban
Loan_Status	Target variable (Loan approved: Y/N)

The dataset consists of around **614 rows and 13 columns**, with a mix of numerical and categorical data.

Data Exploration and Cleaning

We performed initial data exploration and found that the dataset contained several **missing values** in key columns like Gender, Married, LoanAmount, and Credit_History.

We addressed this by:

- Inputting missing values in **categorical columns** with the **mode** (most frequent value)
- Inputting missing values in **numerical columns** with the **median**, as it is robust to outliers

This step ensured our dataset was complete and ready for modeling.

Feature Engineering

To prepare the dataset for machine learning:

Steps Completed:

- **Label Encoding** for binary categorical variables:
  - Gender, Married, Education, Self_Employed, and Loan_Status
- **Re-coded** the Dependents column:
  - Changed '3+' to 3 and converted to numeric
- **One-Hot Encoding** for multi-class categorical variables:
  - Property_Area → created dummy columns for Semiurban, Urban
- **Created clean feature matrix (X)** and **target (y)** ready for model training

Model Building and Evaluation

We trained and evaluated **five classification models** on the dataset using the same training-test split to ensure fairness. These models included:

- **Logistic Regression**
- **K-Nearest Neighbors (KNN)**
- **Decision Tree Classifier**
- **Random Forest Classifier**
- **XGBoost Classifier**

Each model was evaluated on **accuracy, precision, recall,** and **F1 score**. Here is a summary of their performance:

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	**0.86**	0.84	**0.99**	**0.91**
KNN	0.65	0.72	0.81	0.76
Decision Tree	0.73	0.82	0.79	0.80
Random Forest	0.83	**0.85**	0.92	0.88
XGBoost	0.80	0.84	0.87	0.86

Evaluation Metrics Explanation:

- **Accuracy**: How often the model is correct overall
- **Precision**: How many of the "default" predictions were actually correct
- **Recall**: How many of the actual defaulters we caught (VERY important in lending)
- **F1 Score**: A balance between precision and recall

Final Model Recommendation

After comparing the models, **Logistic Regression** was selected as the final model. While more complex models like Random Forest and XGBoost also performed well, Logistic Regression achieved:

- The **highest recall**: 0.99 (capturing nearly all defaults)
- The **highest F1 score**: 0.91
- Excellent interpretability: helps explain the **"why"** behind predictions

This balance between **predictive power** and **explainability** makes it ideal for a business-critical use case like loan default prediction.

Key Findings and Insights

Using Logistic Regression, we observed the following **important predictors** of loan default:

- **Credit History**: The single most important factor — applicants with a poor or no credit history were more likely to default.
- **Loan Amount**: Higher loan amounts slightly increased the chance of default.
- **Self-Employment Status**: Self-employed individuals showed higher risk than salaried applicants.
- **Marital Status and Dependents**: Also showed weak but notable patterns in risk.

These insights can help banks develop more **refined risk profiles** and improve loan approval strategies.

Limitations and Next Steps

Although the model performed well, here are a few limitations and areas for future improvement:

- **Data Imbalance**: The dataset was slightly imbalanced. Techniques like **SMOTE** (Synthetic Minority Over-sampling) could help further improve model performance.
- **Feature Expansion**: We can enhance the model by incorporating additional financial features like savings history, past loan repayment records, or behavioral scores.
- **Hyperparameter Tuning**: Further tuning of model parameters, especially in Random Forest and XGBoost, might yield marginal gains.
- **Explainable AI (XAI)**: Using SHAP or LIME to explain individual predictions would help build more trust in the model.