### SOHOM MUKHERJEE

### <https://www.sohommukherjee.wordpress.com>

### Loan Default Prediction System – Project Summary

Project goal:  
The primary objective of this project is to build a machine learning–based system to predict whether a loan applicant is likely to repay the loan or default. Using historical applicant data, the model helps financial institutions make data-driven lending decisions, reducing risk and improving approval efficiency.

Dataset:  
We used the loan\_data.csv dataset containing applicant demographic details, financial history, and loan attributes. The dataset has categorical and numerical variables, including Gender, Education, Employment status, ApplicantIncome, CoapplicantIncome, LoanAmount, Loan Term, Credit History, and Property Area, with the target variable Loan\_Status (Y/N).

Approach:

1. Data preprocessing: Removed irrelevant fields (Loan\_ID), handled missing values with imputation, applied label/one-hot encoding for categorical variables, and scaled numerical features.
2. Exploratory data analysis (EDA): Visualizations revealed key trends—applicants with a positive credit history and higher incomes had higher approval chances. Correlation analysis confirmed Credit\_History as the most influential feature.
3. Model building: Implemented and compared three models—Logistic Regression, Decision Tree, and Random Forest—using an 80-20 train-test split. Metrics used: accuracy, precision, recall, and F1-score.
4. Model selection: The Random Forest model achieved the highest F1-score (~0.82), balancing precision and recall, making it the best choice for production.
5. Tuning and deployment: Hyperparameter tuning with GridSearchCV improved the model’s performance. The final pipeline was saved as loan\_default\_final\_model.joblib for deployment.

Key insights:

* Credit history strongly influences loan repayment predictions.
* High-income applicants with smaller loan amounts show lower default probability.
* Categorical factors like property area and education have secondary importance compared to credit data.

Conclusion:  
The Random Forest–based Loan Default Prediction System provides reliable predictions, helping lenders minimize risk and optimize loan approval processes. Future work can include class imbalance handling (SMOTE), XGBoost models, feature engineering (loan-to-income ratio), and deployment via a web API.