Credit Scoring and Loan Default Prediction

PRESENTED BY : SHRUTI BADAGANDI

Project Goal

Project Goal The objective of this project is to build a predictive model that can help financial institutions assess the risk of loan defaults. Loan defaults are a significant financial risk, and predicting them using historical data can greatly improve decision-making and risk management. **This project utilizes machine learning techniques along with interactive data visualization tools Like – PYTHON For Data cleaning , visualization. LIBRARIES Are Pandas, Matplotlib. Power BI For Building Dynamic Dashboard . Platform is VS Code.** It also helps stakeholders understand patterns through interactive dashboards.

Dataset Overview

The dataset used for this project is named **CreditScoring-LoanDefault_Dataset.csv.**

It contains information about various applicants including demographic **details (like age and gender), financial metrics (such as income and loan amount), and credit history.** The primary target variable is Loan Default, which indicates whether the loan was defaulted (1) or not (0). This structured data allows us to train models that learn patterns associated with defaults.

Data Cleaning & Preprocessing

Data preprocessing is a critical step to ensure the quality and consistency of input data. I handled missing values by replacing them with the median (for numeric columns) or a placeholder such as 'N/A' (for categorical columns).

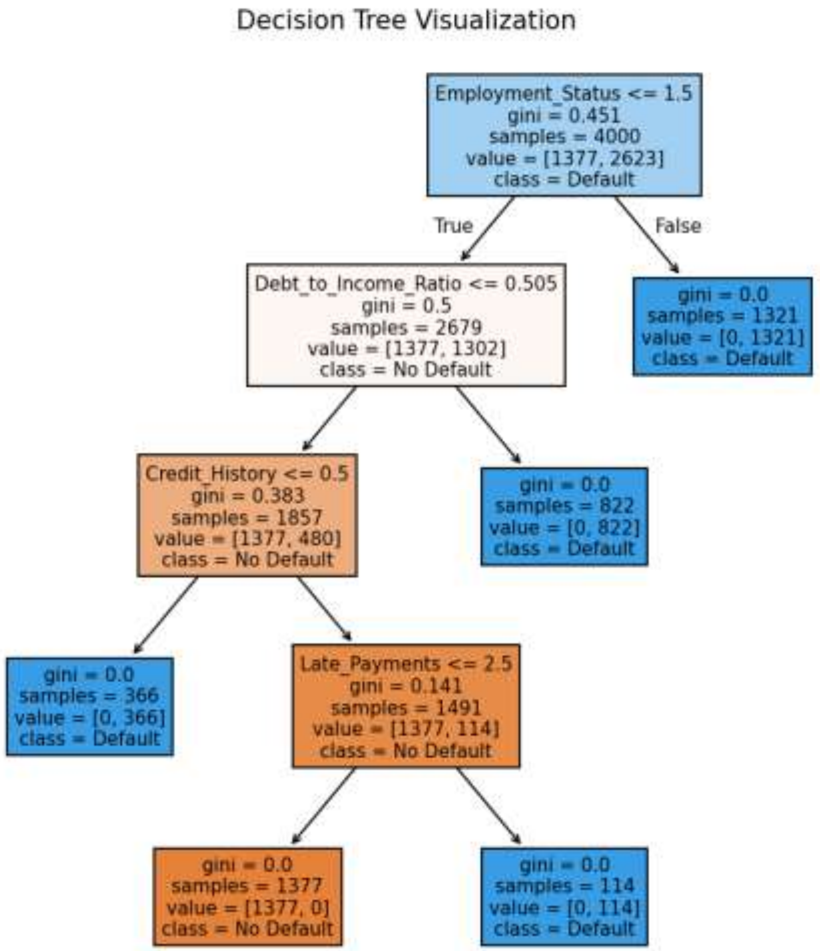
Categorical features were encoded numerically using Label Encoding. After cleaning, the data was split into training and testing sets in an 80:20 ratio to validate model performance effectively. Preprocessing ensures the model receives structured, meaningful input for learning.

Model Training

I Chose the Decision Tree Classifier for its simplicity, interpretability, and ability to handle mixed data types. It builds decision rules based on the most significant features, **helping us visualize and understand the decision-making process.** I defined the features (independent variables) and set Loan Default as the target. The model was trained using the training data and then tested on unseen data to evaluate accuracy and performance.

Decision Tree Visualization

The trained Decision Tree model was visualized using matplotlib, to understand the logic and rules it used to make predictions. Nodes in the tree split based on feature thresholds, such as income level or credit history. For example, a decision path may show that if the applicant has no credit history and a loan amount above a certain threshold, the risk of default is high. This transparency allows stakeholders to see not only what the model predicts, but why it makes those predictions.



Streamlit Web Application

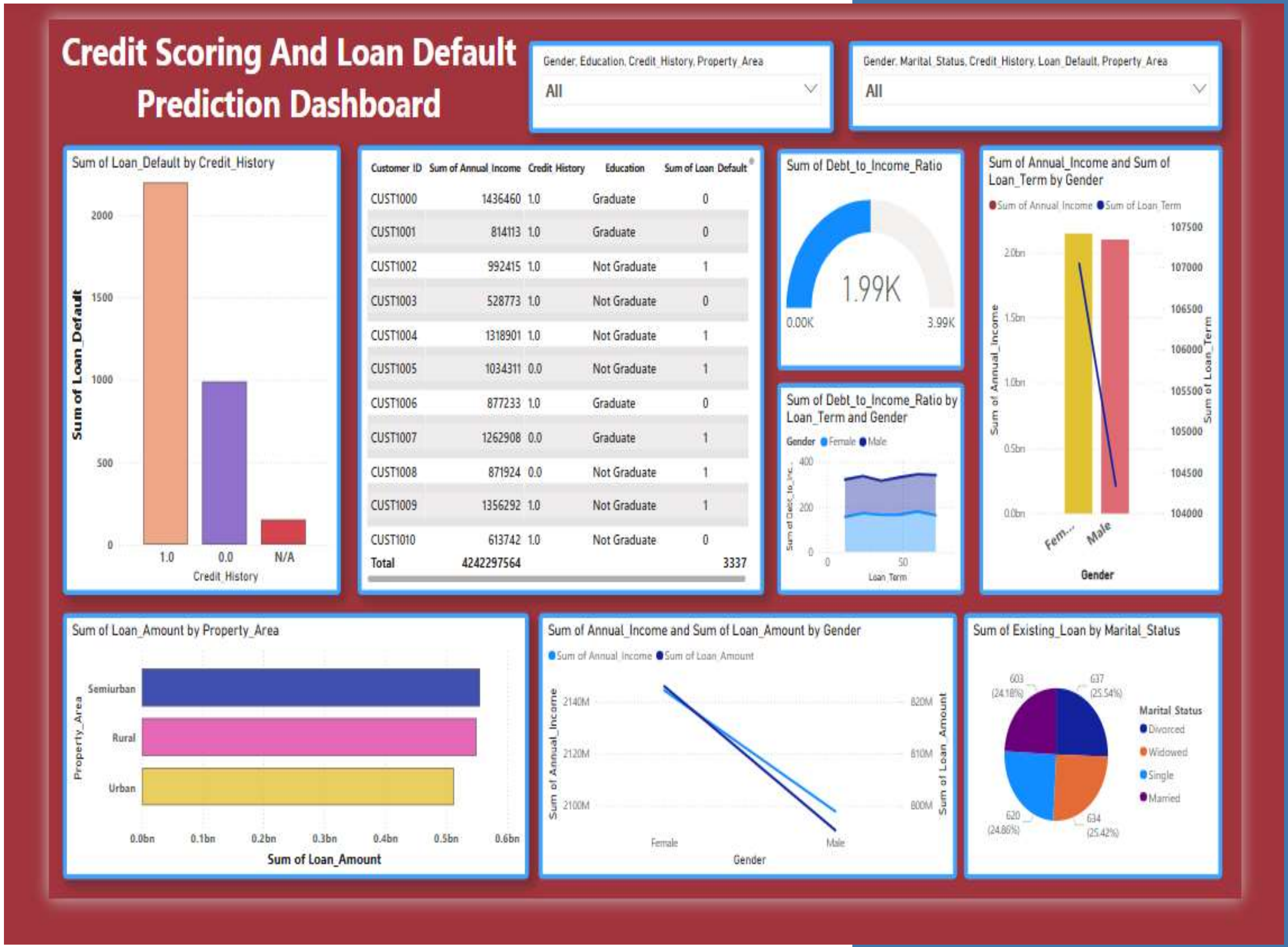
To make our model interactive, I developed a web application using Streamlit. This app allows users—like loan officers or analysts—to input applicant data and get real-time predictions on loan default. The app calculates both the binary result (default/no default) and the probability of default. This tool bridges the gap between technical models and practical usage, offering quick and user-friendly insights for risk-based decision-making.

Power BI Dashboard

The Power BI dashboard provides a visual interface to explore loan data and model outcomes. It includes:

- Pie charts showing overall default distribution
- Bar graphs comparing default rates across genders, employment types, income groups, and regions
- Line charts showing trends over time
- Slicers/filters for drill-down by age, marital status, credit history, and loan amount

Business users can quickly identify high-risk applicant profiles or observe how default trends change over time. For instance, filtering by credit history reveals default rates spike when credit history is poor or missing. The dashboard is not just a visualization tool—it acts as a live reporting platform for stakeholders to make data-informed policy and credit decisions.



Model Evaluation

The model was evaluated using metrics like Accuracy, Confusion Matrix, and ROC AUC Score

The Confusion Matrix showed a good balance of true positives (correctly identified defaults) and true negatives (correctly identified non-defaults).

The ROC AUC score of ~0.86 signifies strong discriminatory ability of the model, meaning it reliably distinguishes between good and bad applicants.

These evaluations confirm that the model is reliable for real-world deployment.

Conclusion

This project successfully applied a full data science pipeline—from data preprocessing and model building to application deployment and dashboard visualization.

I developed an interpretable machine learning model, integrated it into a user-friendly app, and visualized it using a dynamic dashboard. The combination of predictive analytics and business intelligence enables better decision-making and operational efficiency in the credit industry.

THANK YOU