Loan Mount Prediction

Geo M Benny

Predicting loan approval is a critical challenge for financial institutions aiming to assess the creditworthiness of applicants efficiently. Accurate prediction models help streamline the approval process, minimise financial risks, and enhance customer satisfaction by providing quicker and more informed decisions. This project focuses on developing a machine learning model to predict whether a loan application will be approved based on various applicant features.

The dataset used in this analysis includes records with key features such as Gender, Married, Loan Amount, Credit History, Education, and Property Area. These features provide insights into the applicant's background and financial status, which are essential for building a robust prediction model. By analysing and preprocessing this data, the goal is to train models that can accurately predict loan approval and thus assist in making more reliable lending decisions.

Data Overview

The dataset used for this analysis contains [number] records with various features relevant to loan approval predictions. Below is a brief description of each variable:

1. Gender: Categorical feature indicating the applicant’s gender (Male or Female). Encoded numerically for model processing.

2. Married: Categorical feature showing marital status (Yes or No). Converted to numeric values to represent the applicant’s marital status.

3. Dependents: Numeric feature representing the number of dependents (0, 1, 2, or 3+). Encoded to reflect family size.

4. Education: Categorical feature indicating education level (Graduate or Not Graduate). Converted to numeric values for model compatibility.

5. Self_Employed: Categorical feature showing employment status (Yes or No). Encoded numerically to represent self-employment status.

6. Property_Area: Categorical feature representing the type of property area (Urban, Semi urban, Rural). Encoded to reflect different property areas.

7. Credit_History: Numeric feature indicating the applicant's credit history, typically a binary variable where 1 indicates a positive credit history and 0 indicates a negative history.

8. Loan_Amount: Numeric feature representing the loan amount requested by the applicant.

9. Loan_Amount_Term: Numeric feature indicating the term of the loan in months.

This dataset provides a comprehensive view of the applicant's background and financial situation, essential for building predictive models for loan approval.

Exploratory Data Analysis (EDA)

1. Distribution of Key Variables:

Loan Amount: The distribution shows a right-skewed pattern, with a concentration of lower loan amounts. Many applicants request smaller loans, with fewer applicants seeking high amounts.

Credit History: Most applicants have a positive credit history (1), indicating a generally good credit background among the dataset's applicants.

Loan Term: The distribution shows that most loans are for terms of 360 months, indicating standard long-term loan requests.

2. Patterns Observed:

Marital Status and Dependents: Married applicants with more dependents tend to have a higher likelihood of loan approval. This pattern suggests that family size might impact financial stability and loan eligibility.

Education and Employment Status: Graduates and employed individuals are more likely to have their loan applications approved, highlighting the role of education and stable employment in loan approval decisions.

3. Relationships and Correlations:

Credit History and Loan Approval: A strong positive correlation exists between a positive credit history and loan approval, indicating that applicants with better credit histories are more likely to be approved.

Loan Amount and Income: There is a moderate correlation between the loan amount requested and the applicant's income. Higher income often corresponds to larger loan requests, reflecting a relationship between financial capacity and loan size.

Data Preprocessing

1. Handling Missing Values:

Missing values in Credit_History, Married, Gender, Dependents, Self_Employed, LoanAmount, and Loan_Amount_Term were filled using appropriate strategies:

Categorical Variables: Missing values in categorical features like Credit_History, Married, and Gender were filled with the most frequent category or imputed using random values (0 or 1) when applicable.

2. Encoding Categorical Variables:

Categorical variables were converted to numerical form using mapping: Gender, Married, Dependents, Education, Self_Employed, and Property_Area were mapped to binary or ordinal numeric values.

For example, Gender was mapped as Male: 1 and Female: 0, while Property_Area was mapped as Urban: 2, Semi urban: 1, and Rural: 0.

3. Feature Scaling:

To ensure uniformity in data ranges, particularly for continuous variables, feature scaling was applied. Loan amounts and applicant incomes were scaled to normalise their distribution and reduce the impact of outliers.

4. Feature Engineering:

Loan Duration: The loan duration was converted from categorical values (e.g., 2 months, 6 months) into numerical values representing the total duration in days or months.

Model Development

In the development of the loan prediction model, two machine learning algorithms were used: **Logistic Regression** and **Support Vector Classifier (SVC)**.

1. **Logistic Regression:**

   - This model was chosen for its simplicity and effectiveness in binary classification problems. It calculates the probability of a loan being approved (loan status Y or N) based on the input features. The model was trained on the processed dataset and showed reasonable performance in predicting loan approval outcomes.

2. **Support Vector Classifier (SVC):**

   - SVC was employed to further enhance the model's prediction ability by finding an optimal hyperplane that maximises the margin between the two classes. This method is particularly effective in cases where the data is not linearly separable.Which in fact I used to create a model.

Model Evaluation

Both models were evaluated using common performance metrics like **accuracy score**, **confusion matrix**, and **classification report**. The models were trained on the training dataset and evaluated on the test dataset to ensure generalisation.

Logistic Regression produced straightforward, interpretable results with good accuracy around 83% **. SVC** provided robust performance, particularly for more complex patterns in the data.

Results and Discussion

The Logistic Regression model achieved an accuracy of 83% in predicting loan approval, indicating a strong performance in classifying applicants as either approved or not approved. This level of accuracy suggests that the model successfully captured key patterns within the dataset, such as credit history, income, and property area, which significantly influence loan decisions.

The model's confusion matrix revealed a good balance between true positives and true negatives, with relatively few misclassifications. However, some false positives and false negatives still occurred, which could be attributed to noise or non-linear relationships in the data.

Overall, the results highlight the importance of applicant credit history, income, and employment status in predicting loan approvals, with potential future enhancements focusing on improving feature engineering and addressing non-linear relationships.

Conclusion

This project aimed to develop a machine learning model to predict loan approvals based on applicant information. Using Logistic Regression and SVC, the model effectively classified loan approval status with an accuracy of 83%, highlighting key factors such as credit history, income, and property area as influential predictors.

For the Streamlit app deployment I used SVC algorithm to create the model for Loan Mount Prediction.

References

- Skill Vertex
- Scikit - Machine Learning Book
- Documentation of Streamlit and Scikit.
- Youtube