DATATHON 2024

TEAM ARAVATA

VISHWAKRMA INSTITUTE OF INFORMATION TECHNOLOGY

Contents

Objective:

- Develop a robust machine learning-based system to **predict loan eligibility** accurately.
- Explore various machine learning models and **techniques** to get best result possible.
- Utilize available data to build predictive models that assist in **decision-making processes** for loan approval.
- Enable financial institutions to make informed lending decisions based on **data-driven** insights.

Observations:

- The **Applicant's Income** and **Co-applicant's Income** columns display wide income variations among applicants. The average applicant income is **$5403.46**, with some individuals earning up to **$81000**.

- The **Loan Amount** varies from **9** to **700**, with an average of about **146.41**, indicating **diverse loan requirements** among applicants.

- The **Loan Amount Term** mostly centers around **360** (which could be in months, implying 30 years if so), but it ranges from **12 to 480**, showing different loan repayment terms.

- The **Credit History** has a mean of **0.84**, suggesting that a majority of applicants have a **good credit history**.

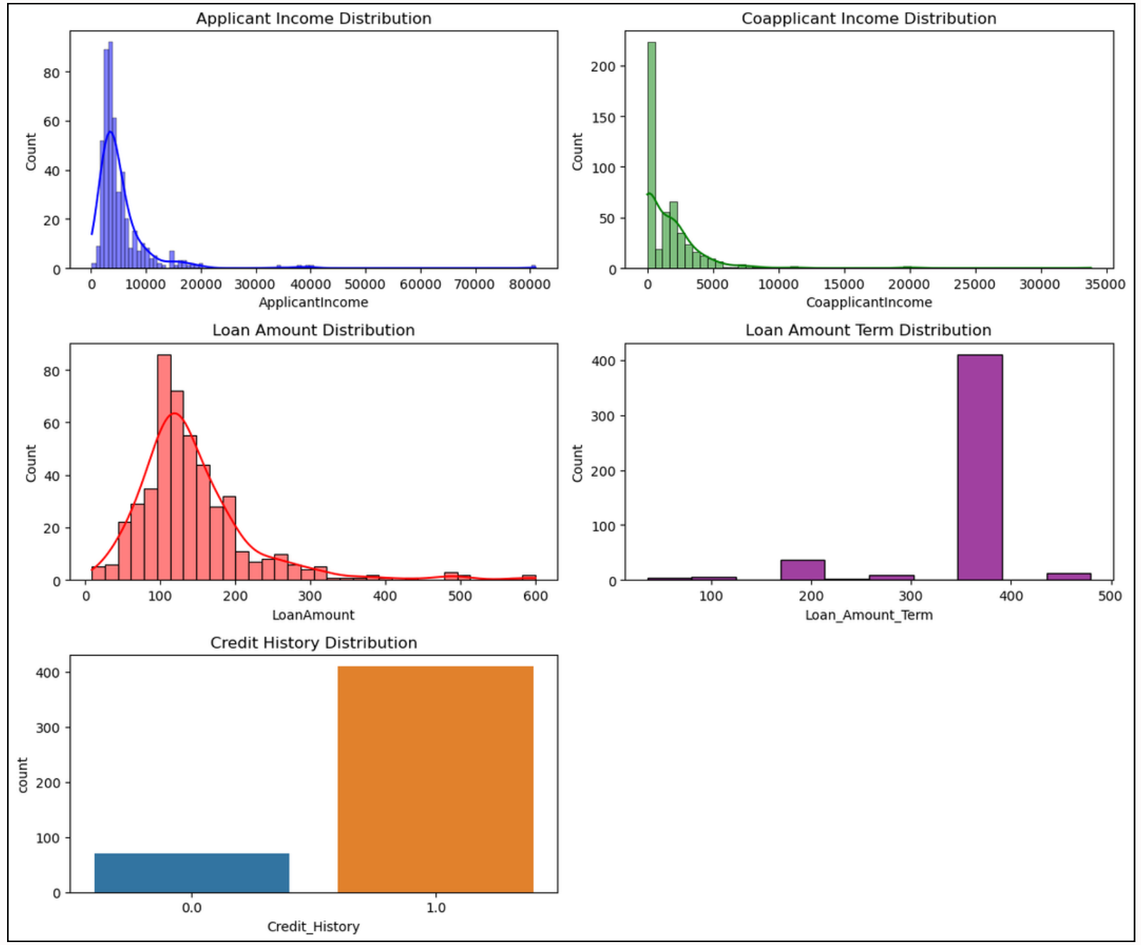

Fig. 1 Exploratory Data Analysis

Understanding the correlations:

- **Credit History:** There's a strong positive correlation between *'Credit_History'* and '**Loan_Status**'. This suggests that applicants with a good credit history are more likely to have their loans approved.

- Marital Status: '**Married**' shows a positive correlation with '**Loan_Status**', indicating that married individuals might have a slightly higher chance of loan approval compared to unmarried individuals..

- Education: There's a positive correlation between being a **graduate** and '**Loan_Status**'. Graduates are more likely to get their loans approved than non-graduates.

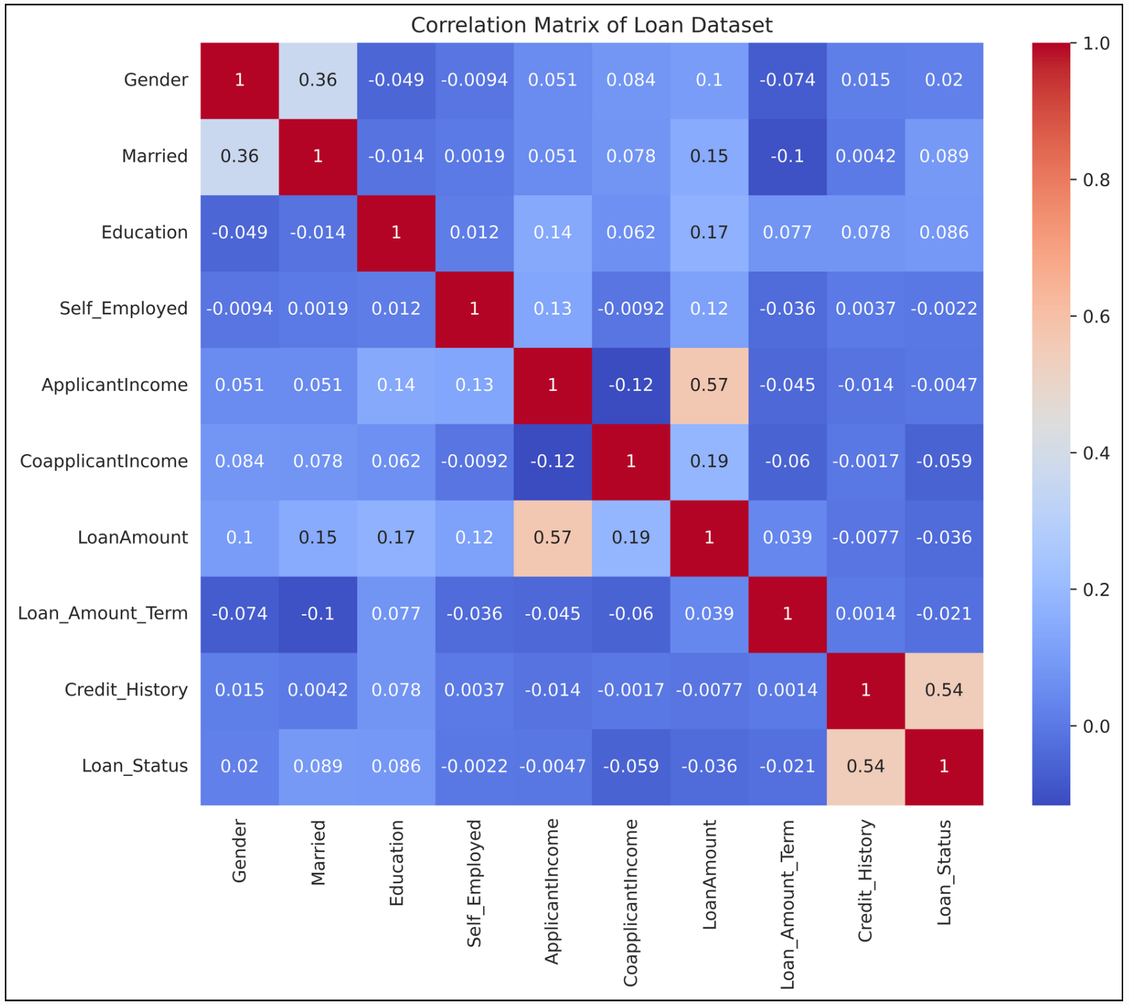

Fig. 2 Correlation Matrix of loan dataset

Understanding the correlations:

- Applicant Income: As expected, there's a positive correlation between '**ApplicantIncome**' and '**LoanAmount**', indicating that higher income applicants tend to apply for and receive larger loan amounts.

- Loan Amount and Co-applicant Income: There's also a positive correlation between '**LoanAmount**' and '**CoapplicantIncome**', suggesting that loans with higher amounts are associated with higher coapplicant incomes, possibly to ensure repayment capability.

Proposed Approach:

Method 1: (Neural Networks)

- **Neural Network Architecture Design:**
  - Utilizes a feedforward neural network with multiple dense layers.
  - Adjusts the number of nodes in hidden layers and applies **ReLU** activation functions.
  - Implements dropout layers to mitigate **overfitting.**

- **Hyperparameter Optimization:**
  - Engages in grid search to optimize hyperparameters such as **node count, dropout probability**, **learning rate**, and **batch size**.
  - Evaluates model performance based on validation loss.
  - Identifies the best model configuration with the **lowest validation loss.**

- **Result:**
  - **Accuracy: 75%**
  - **For "Yes":** Precision- 0.56, Recall- 0.68, f1-score - 0.61
  - **For "No":** Precision- 0.85, Recall- 0.78, f1-score - 0.82

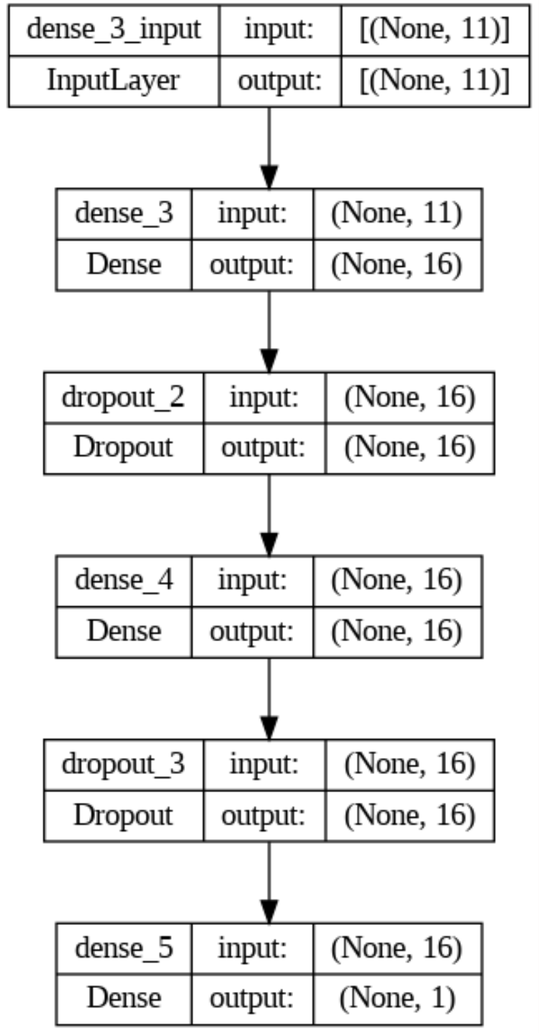

Fig. 3 Neural Network Architecture Design

Method 2: (Random Forest Classifier)

With 200 decision trees and a maximum of 10 leaf nodes per tree, the model was trained on the training data and evaluated on the test set. The achieved accuracy and precision were reported as performance metrics.

- **Result:**
  - **Accuracy: %**
  - **For “Yes”:** Precision- 0.56, Recall- 0.68, f1-score - 0.61
  - **For “No”:** Precision- 0.85, Recall- 0.78, f1-score - 0.82

Method 3: (Gradient Boosting)

It is an ensemble technique that builds a strong predictive model by combining multiple weak learners, typically decision trees, in a sequential manner

Result:

- Accuracy:83%
- Precision:86.7%
- Recall:90.9%
- F1:88.2%

Method 4-(KNN)

1. Instance-Based Learning:
   - KNN belongs to the family of instance-based learning algorithms, also known as lazy learning algorithms.
   - Unlike model-based algorithms that learn explicit patterns from the training data to make predictions, KNN relies on the entire training dataset during prediction time.
2. Distance-Based Classification:
   - In KNN classification, the algorithm predicts the class of a new data point by identifying the majority class among its K nearest neighbors in the feature space.

RESULT:

- Accuracy:67.7%
- Precision:80.4%
- Recall:77.2%
- F1-score:90.1%

Method 5- Logistic Regression

Logistic Regression statistical method for binary classification tasks in machine learning. Hence it is relevant for our model. Despite its name containing "regression," it is primarily used for classification rather than regression tasks.

Result:

- Accuracy: 80.21%
- Precision: 90.90%
- Recall: 85.71%
- F1-Score: 100%

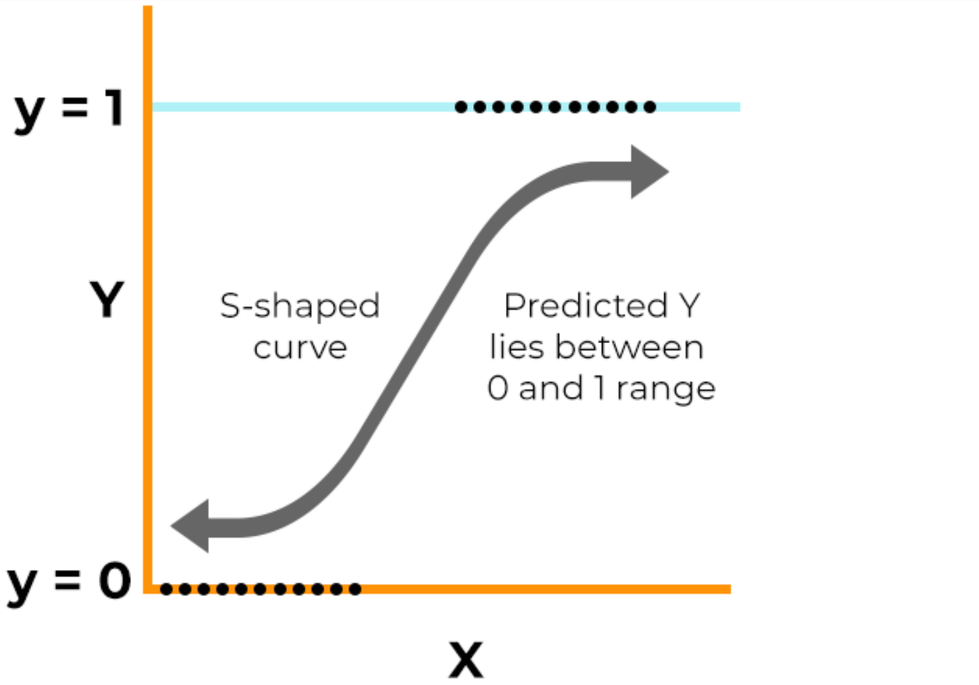

Fig. 4 Logistic Regression

Solution:

- We employed a diverse ensemble of machine learning models, including **RandomForest**, **GradientBoosting**, **AdaBoost**, and **LogisticRegression.**

- Leveraged the **VotingClassifier** ensemble technique to combine predictions from individual models, enhancing overall prediction accuracy.

- Trained individual classifiers on the training data to capture diverse patterns in the dataset.

- Demonstrated the effectiveness of the approach by evaluating and comparing the accuracy of each model separately.

- Achieved superior prediction performance with the **ensemble method**, as evidenced by the higher accuracy of the **VotingClassifier**.

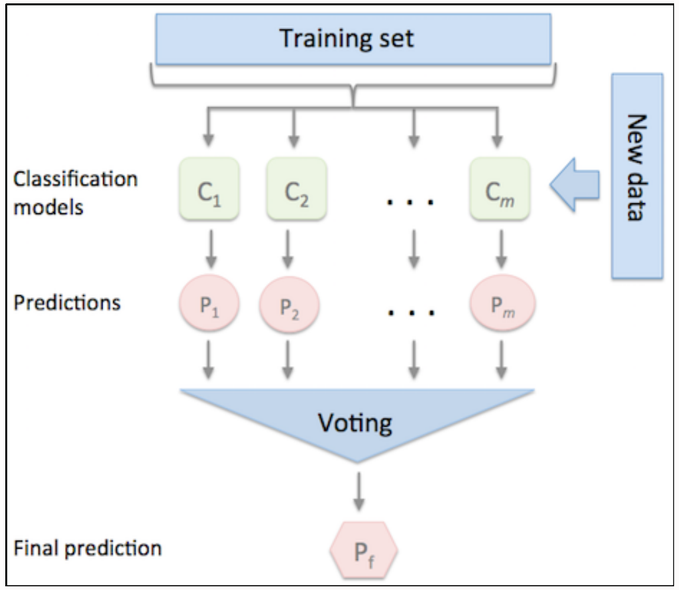

Fig. 4 Working of Voting Classifier

Image source: Mlxtend

Tech Stack:

Language:

Frameworks:

Tools:

Future Scope:

- **Model Ensemble Refinement:** Experiment with different ensemble techniques can be done such as stacking or blending to further improve predictive performance.

- **Feature Engineering:** Additional features or transformations can be explored to enhance the predictive power of the models.

- **Advanced Hyperparameter Tuning:** Utilize more sophisticated hyperparameter optimization techniques like Bayesian optimization or genetic algorithms to fine-tune model parameters effectively.

- **Deployment of app-based Interface**: Create a user-friendly mobile application where users can input their information and receive instant loan eligibility predictions.

Thank You