Report on Loan Approval Prediction Model

1. Problem Statement

Can we assess whether a loan will be approved or declined based on the data input by an applicant during the online application, including both the applicant's and loan information?

The objective of this project is to assess the likelihood of loan approval based on applicant and loan information, and this can potentially help in automating the decision-making process in loan approvals, reducing bias and increasing efficiency.

2. Data Sources

Source of the data is here. The data for this project was downloaded from Analytics Vidhya's DataHack platform. Analytics Vidhya is a well-known community in the data science field, offering a platform for learning and practicing data science and analytics. The features of this dataset includes: [Loan_ID, Gender, Married, Dependents, Education, Self_Employed, ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term, Credit_History, Property_Area, Loan_Status]. Within the list of features, ['ApplicantIncome', 'CoapplicantIncome', 'LoanAmount'] are numerical variables while others are categorical variables.

3. Feature Engineering and Preprocessing

a. Handling Missing Values

In the preprocessing stage, special attention was given to handling missing values in the dataset, which is crucial for maintaining the integrity and reliability of the machine learning models. The approach varied based on the nature of the variables:

- Categorical Variables: For missing values in categorical variables, the mode was used for imputation. This method is effective as it preserves the common trends in the data without introducing significant bias.
- Numerical Variables: The median was chosen for imputing missing values in numerical variables, and this is because of the presence of outliers. The median is less affected by outliers compared to the mean.

b. Categorical Data Encoding

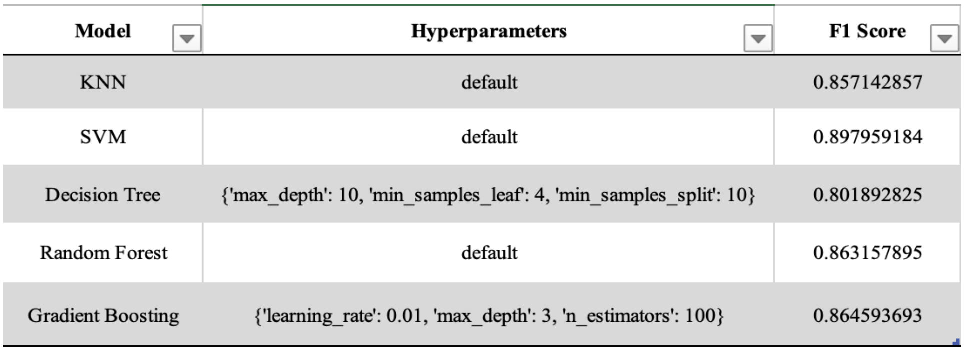
Label Encoding was used to convert categorical variables into a numerical format. This technique assigns a unique integer label to each category within a categorical variable, simplifying data for machine learning models.

c. **Feature Scaling**

Numerical variables were standardized using the standardization technique. This process ensures that variables with different units and ranges do not unduly influence machine learning algorithms, leading to more robust and accurate model predictions.

4. Model Results

In the context of assessing loan approvals, I have chosen to focus on Precision and Recall, and their balance as represented by the F1 Score, rather than solely on Accuracy. Accuracy, while providing an overall success rate, can be misleading in cases of imbalanced datasets. Precision is crucial in this scenario as it reflects the ability of the model to correctly identify true loan approvals, minimizing the risk of approving unqualified applicants. Recall, on the other hand, ensures that genuinely eligible applicants are not denied, maintaining the integrity and inclusiveness of the loan process. The F1 Score harmonizes these two metrics, offering a single measure that balances the need to minimize false positives (approving bad loans) and false negatives (rejecting good loans). This balanced approach is essential in the financial domain, where both types of errors carry significant consequences.



Model	Hyperparameters	F1 Score
KNN	default	0.857142857
SVM	default	0.897959184
Decision Tree	{'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 10}	0.801892825
Random Forest	default	0.863157895
Gradient Boosting	{'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 100}	0.864593693

5. Hyperparameter Search Results

Grid search was employed to fine-tune the hyperparameters for both the Decision Tree and Gradient Boosting models. The best hyperparameters and their corresponding F1 scores are as follows:

- Decision Tree:

  Parameters: {'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 10}

  F1 Score: 0.802
- Gradient Boosting:

  Parameters: {'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 100}

  F1 Score: 0.865

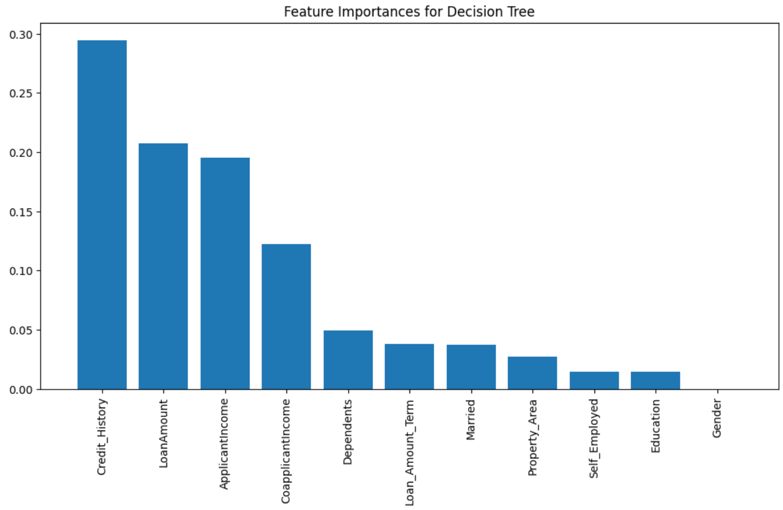
These results demonstrate the effectiveness of grid search in optimizing the models for the given problem, resulting in improved predictive performance.

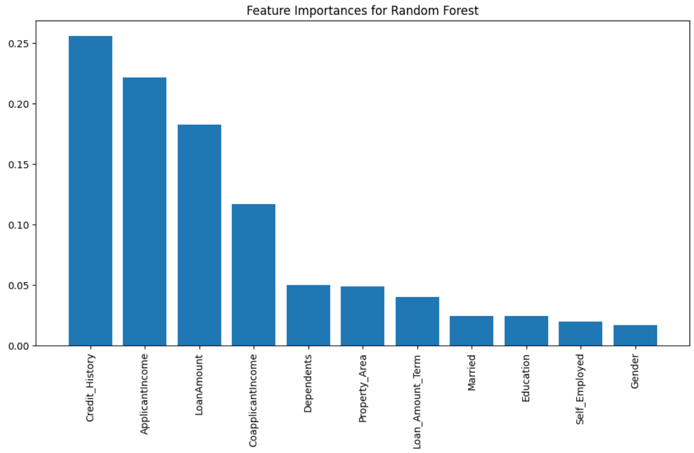
6. Conclusions

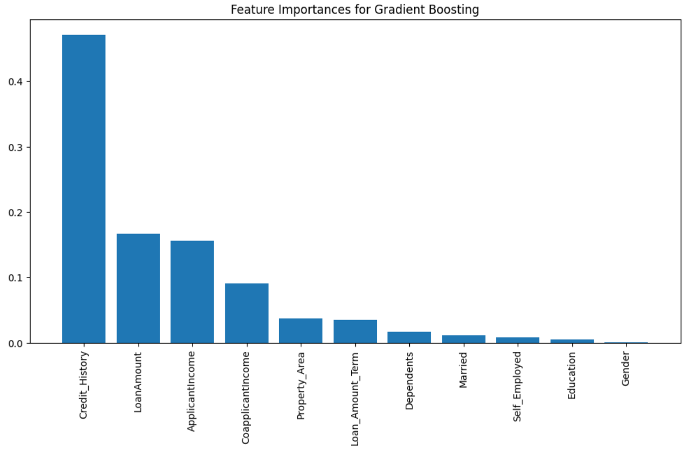
a. Model Comparison

Among the models evaluated in this project, the Support Vector Machine (SVM) emerged as the top-performing model with the highest F1 score (0.897959184). SVM's robustness in handling complex decision boundaries and its ability to capture subtle relationships in the data contributed to its superior performance.

b. Feature Importance







From the feature importance analysis, it is evident that several key factors significantly influence the loan approval decision. Credit history, loan amount, applicant income, and co applicant income were identified as the most influential features in the classification process. These features carry the most weight in determining whether a loan application will be approved or declined, providing valuable insights into the decision-making process.

c. Applicability

These models find real-world use in financial institutions and online loan applications, automating assessments for streamlined, unbiased, and efficient approvals. Yet, they have limitations, including data quality, bias in historical data, and external economic factors. Ongoing monitoring and refinement ensure their relevance in dynamic lending environments.

7. GitHub Repository Link

Link: https://github.com/Allison67/data-classification-loan-prediction