**Loan Approval Prediction**

Project Report

**Team Members**

Avanthika Kanna (11594068)

Sumith Dugyala (11697983)

Kalyan Krishna Karumuri (11605269)

Naresh Chigurla (11608346)

**Abstract**

Using machine learning algorithms, this project presents a way to assist financial institutions in automating their lending processes and optimizing them. The main aim is to produce predictive models that would be able to decide, with great accuracy, whether a loan application should be accepted or rejected based on a number of applicant attributes. Multiple classification models are applied, and results are evaluated on standard evaluation metrics that include Decision Trees, Random Forests, Support Vector Machines (SVM), and Naive Bayes. Results showed that SVM- model to be the best among tested algorithms, achieving the highest accuracy and precision scores, trailed very closely by Random Forest, whereas Decision Tree and model, followed with much poorer performance. From the exploration of feature importance through their significance values, it has been demonstrated that the credit history, total income, debt to income ratio, and loan amount are still important features for predicting the loan approval outcome. The project hence explores how effective machine learning techniques are at loan approval prediction and comes up with insights that the model may inform upon the decision-making processes in financial institutions.

**Introduction**

Approval of loan is an indispensable process in today's financial landscape for individuals and institutions requiring to mitigate the risk involved. On the other hand, digitalization has brought on board an enormous amount of application for a loan, such that the traditional way of approval is not effective; a lot of errors have been observed in decision-making. The project would thus address this pressing problem through the use of machine learning techniques in predicting loan approval to up the efficiency of the process and, hence, decision accuracy.

The aim of the project is to create a predictive model capable of predicting whether a loan will approve for a loan applicant based on several attributes from his demographics and financial background. The goal here is to categorize the applicants into approved and non-approved by training the machine learning model from historical data, including applicant profiles with corresponding loan outcomes. This augurs well for quicker loan approval, but it also reduces the chance of human errors and biases forming part of a manual assessment by its inherent nature. This project tries to close the traditional gap between the methods of loan approval and the modern, data-driven approaches which obviously offer a long-standing problem in the financial industry a much-needed, scalable, and efficient solution. It is an aim of ours to participate actively in the further development of lending practice, which should be ever more effective, just, and sustainable through interdisciplinary competences and team working.

**Background**

Loan issuance is a very critical domain bringing together finance, data science, and technology in a very critical view of individual and institutional interests in loan approvals. However, from a traditional point of view, it is quite cumbersome to determine the eligible or rather creditworthy client that can be able to access the loan through manual methods. Machine learning (ML) in today's digitalized world is a game-changer in this process, using abundance in historical data on loans for decision automation. On the other hand, ML brings speed, reduction in cost, effective risk management, and at the same time, overcomes biases and subjectivity to bring about fairer lending. However, ML systems pose serious ethical, regulatory, and interpretability issues; hence, they need to be transparent and require oversight. The goal of the present work was to continue researching the application of ML in loan approval so as to further responsible AI deployment in finance via developing interpretable, fair, and effective models.

**Literature Review**

The use of machine learning to predict loan approval has gained much traction in the recent past, bearing in mind that it may improve efficiency in decision-making systems and accuracy developed within financial institutions. A great deal of work has been done in this area to make this system more efficient and exact.

Diwate et al. (2021) in conducted a study on the basis of loan approval prediction using machine learning techniques in the International Research Journal of Engineering and Technology. They have also empirically validated the algorithm performance and finally emphasize that there is room for feature selection and model evaluation to further increase predictive accuracy.

In a doctoral dissertation, Cherif and Berkane (2023) show that their use is very promising in the field. He made a review that consisted of building predictive models in relation to particular needs of the financial institutions, in turn, the credit history and levels of income and loan terms.

Comparative study of machine learning algorithms in the prediction of customer loan approval was introduced by Tumuluru et al. (2022) against the background of the second ICAIS. In the study, therefore, it informed the effectiveness of different algorithms and feature engineering techniques on the prediction of the loan approval outcome.

In another research work, Kadam et al., 2021, have made predictions for loan approval using different machine learning algorithms. They have pointed out the issue of rigorous data preprocessing and model selection that has to be done to get a strong model for making predictions about loan approval.

**Experiment Methodology**

The following are the major steps to this project: data collection, preprocessing of the collected data, feature engineering, model selection, model training, and model evaluation. Each of them is so designed that, from the design to the details, the experimental result can be robust, reproducible, and valid. The experimental design is detailed, as shown below:

1. Data Collection:

This project was done using data sourced from Kaggle on past loan application records with a varied demographic and financial record. After obtaining the dataset, it had 381 rows, and 13 columns of features included ApplicantIncome, LoanAmount, Credit\_History among others.

1. Data Preprocessing:

Handles the null values in the dataset by removing rows that have missing values. All the categorical variables are encoded using one-hot encoding for the conversion of the categorical data into a numerical format in preparation for model training. Scaling of numerical features was done using min-max scaling in order to make the numerical features be within a uniform range.

1. Feature Engineering:

More derived features such as TotalIncome and Debt\_Income\_Ratio are prepared from existing features to prepare the input data for the models. After finalization of the dataset, unrequired columns like Loan\_ID, ApplicantIncome, and CoapplicantIncome are dropped.

1. Model Selection:

Decision Trees, Random Forest, Support Vector Machines (SVM), and Naive Bayes are the four classification algorithms selected to train and test the model. Each approach of these has some particular strengths of its own and is appropriate in dealing with our problem of predicting loan approval.

1. Model Training and Evaluation:

In the ratio of 80:20, the dataset was divided into training and testing sets for developing the model and performance estimation, respectively. While building the model on this dataset, the performance evaluation of the model was computed with the evaluation parameters as accuracy, precision, recall, F1-score, and ROC-AUC. The performance scores are checked for each model by incorporating the training and testing data, ensuring that robustness is kept and overfitting is not done.

1. Experiment Setup and Parameters:

Default parameters are used to train every model at the first go, just to get an idea of the base performance of the model. Some of the hyperparameter tuning can be done using grid search or random search. Cross-validation methods will further help in the validation of the generalization ability and hence help reduce variance of the models, such as k-fold cross-validation.

1. Experiment Iterations:

Based on these initial results, further hyperparameter tuning, exploration of different feature sets, or alternative algorithm assessment, which would improve model performance, should be considered in further experiments. Sensitivity analyses can be pursued to understand the different impacts of the change in parameters and features from the outcomes of the model.

1. Results Analysis:

Results of each iteration of the experiment are analyzed in great detail to identify trends, patterns, and areas of improvement. Such insights direct the next iterations, guiding decisions about model selection, feature engineering, and hyperparameter tuning.

Comprehensive exploratory data analysis was done on the project. First, univariate analysis was done to visualize numerical variables:

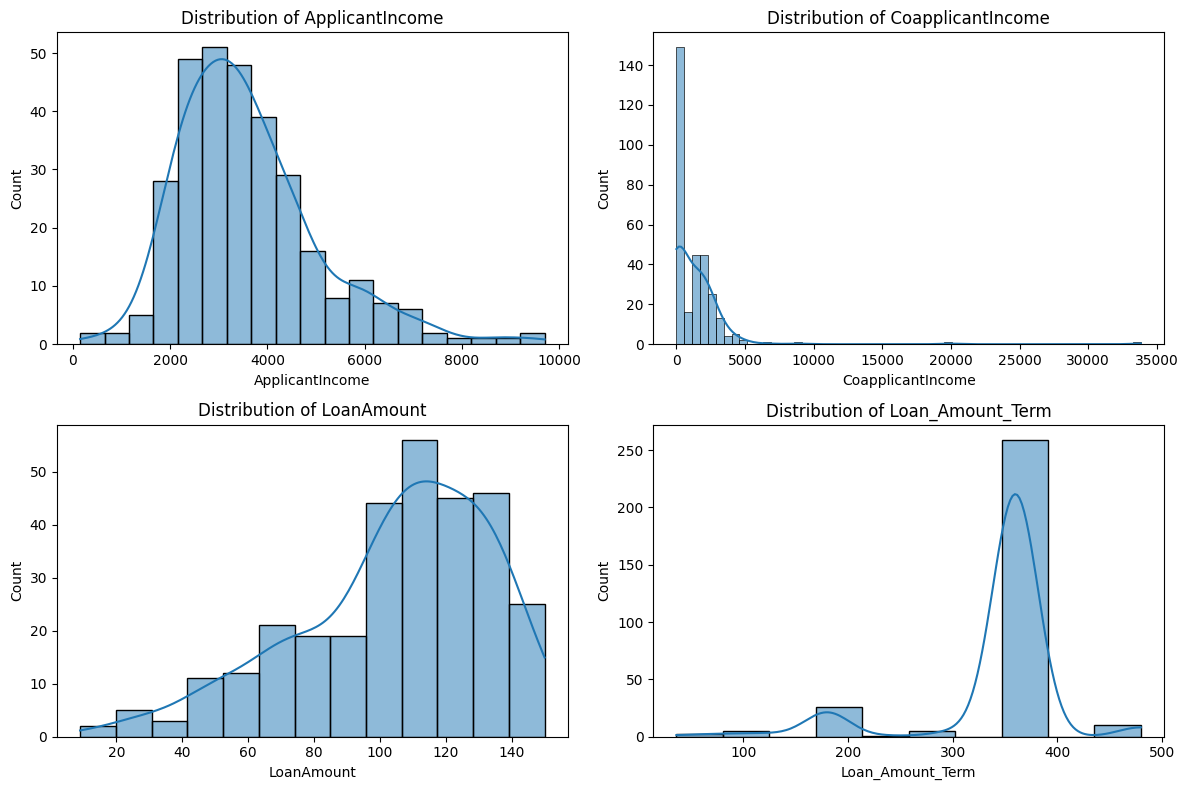


Figure 1 - Numerical Variables

Next, categorical variables were also visualized and it was observed that the dataset was biased towards the male gender as there were significantly more males than females in the dataset as can be shown in the following plot:

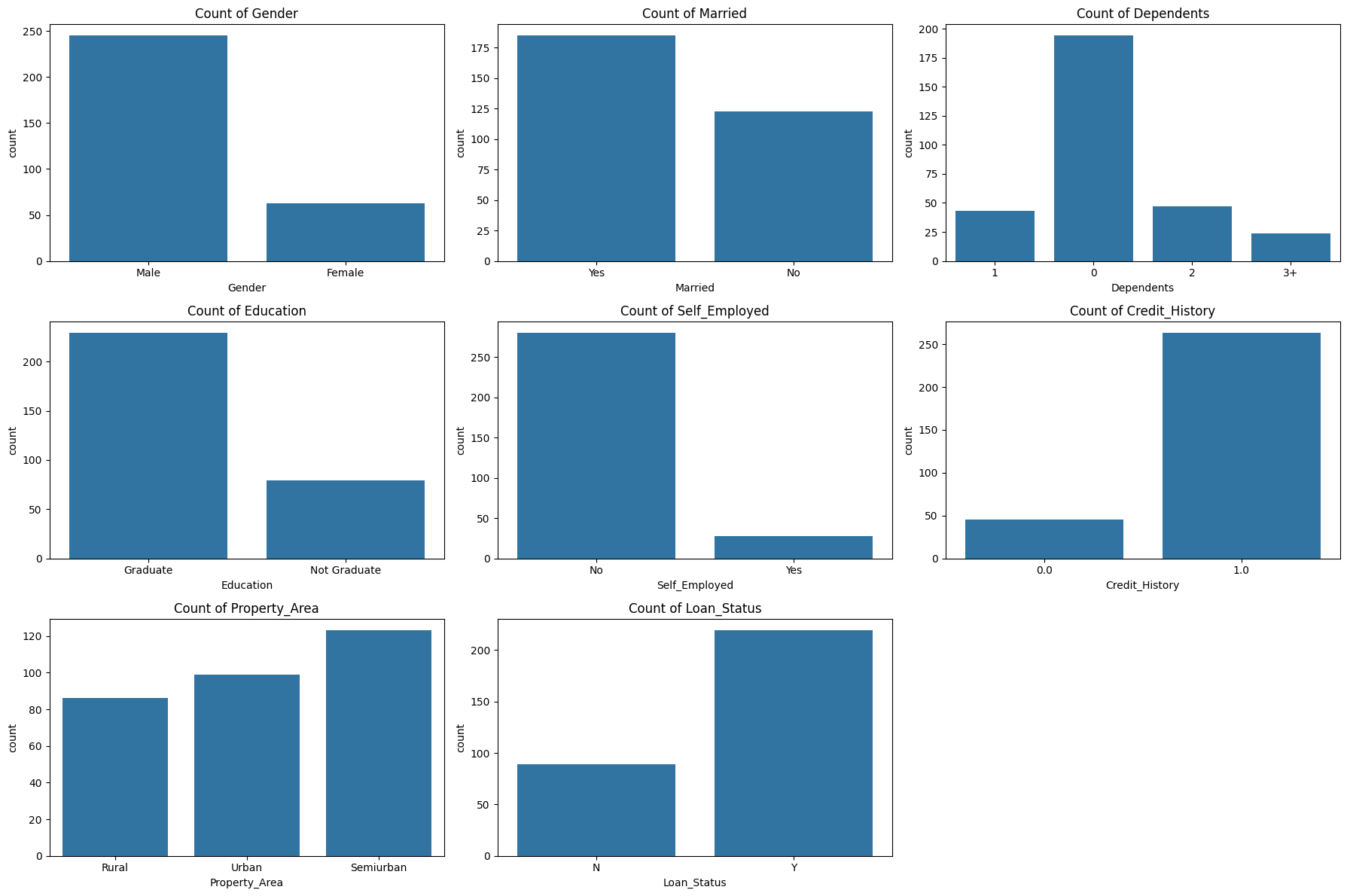


Figure 2 - Categorical Variables

Afterwards, the relationship between numerical variables and the loan status was deemed necessary and thus visualized during the bivariate analysis as shown in the following visualization:

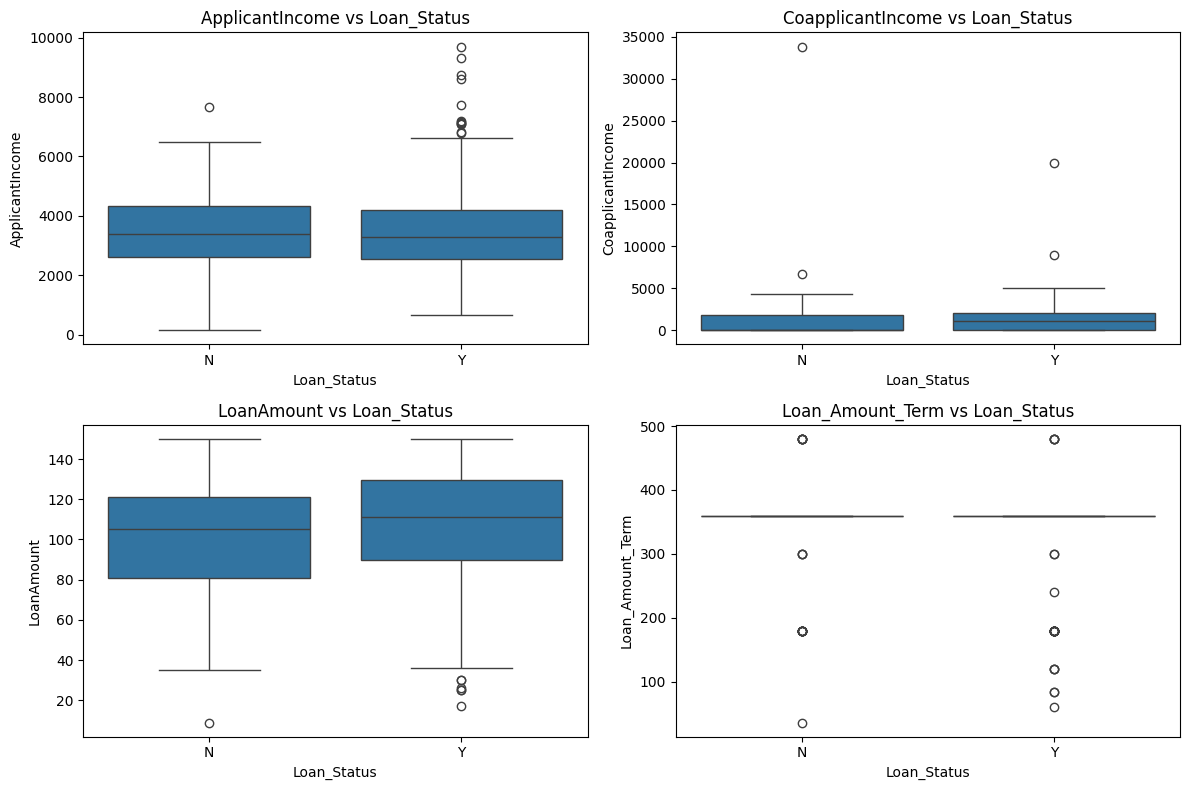


Figure 3 - Relationship between numerical variables and Loan\_Status

The same was done for the categorical variables as shown below as well:

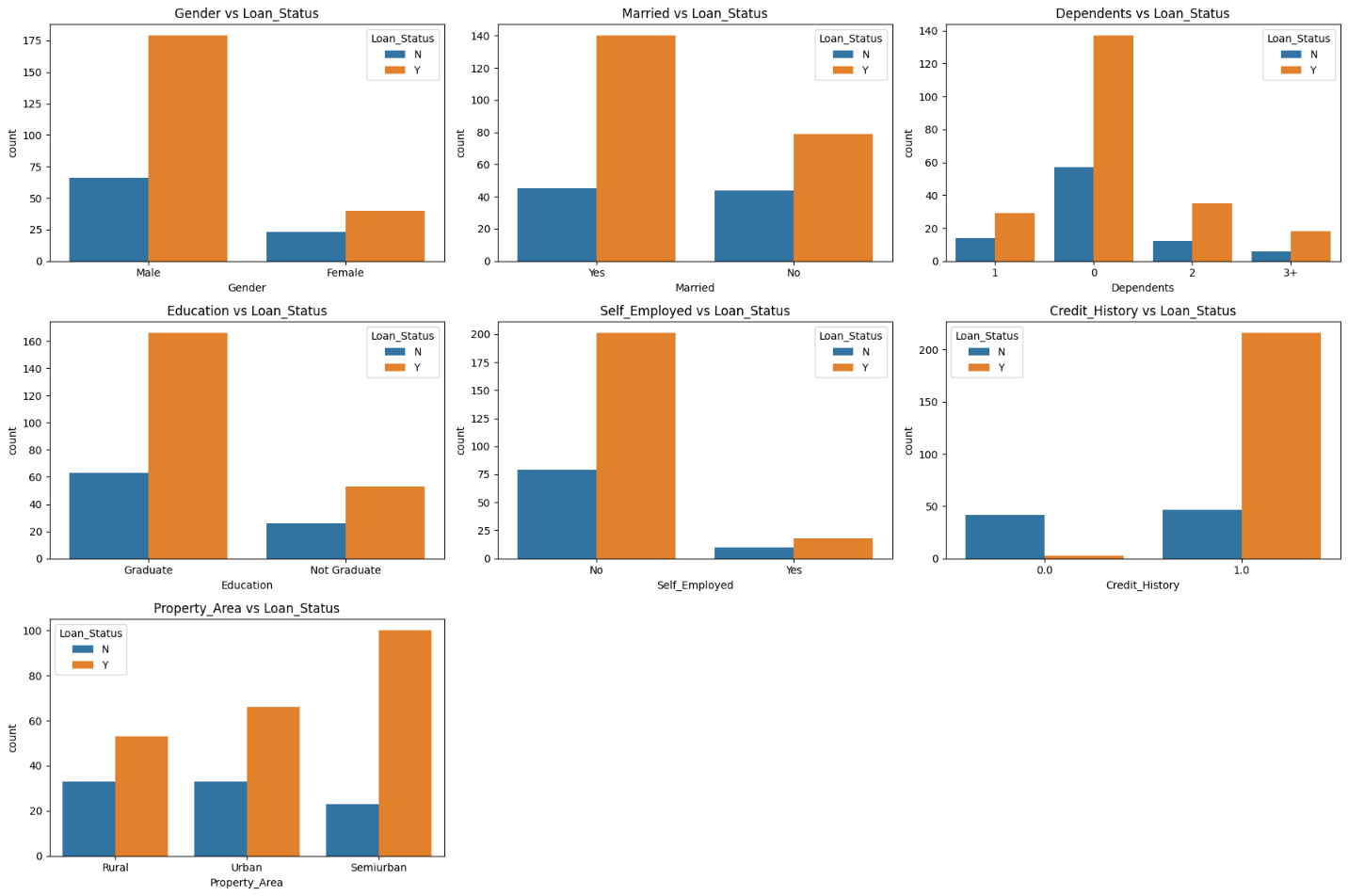


Figure 4 - Relationship between categorical variables and Loan\_Status

Multivariate analysis was then conducted by generating a pairplot for numerical variables as can be seen in the following visualization:

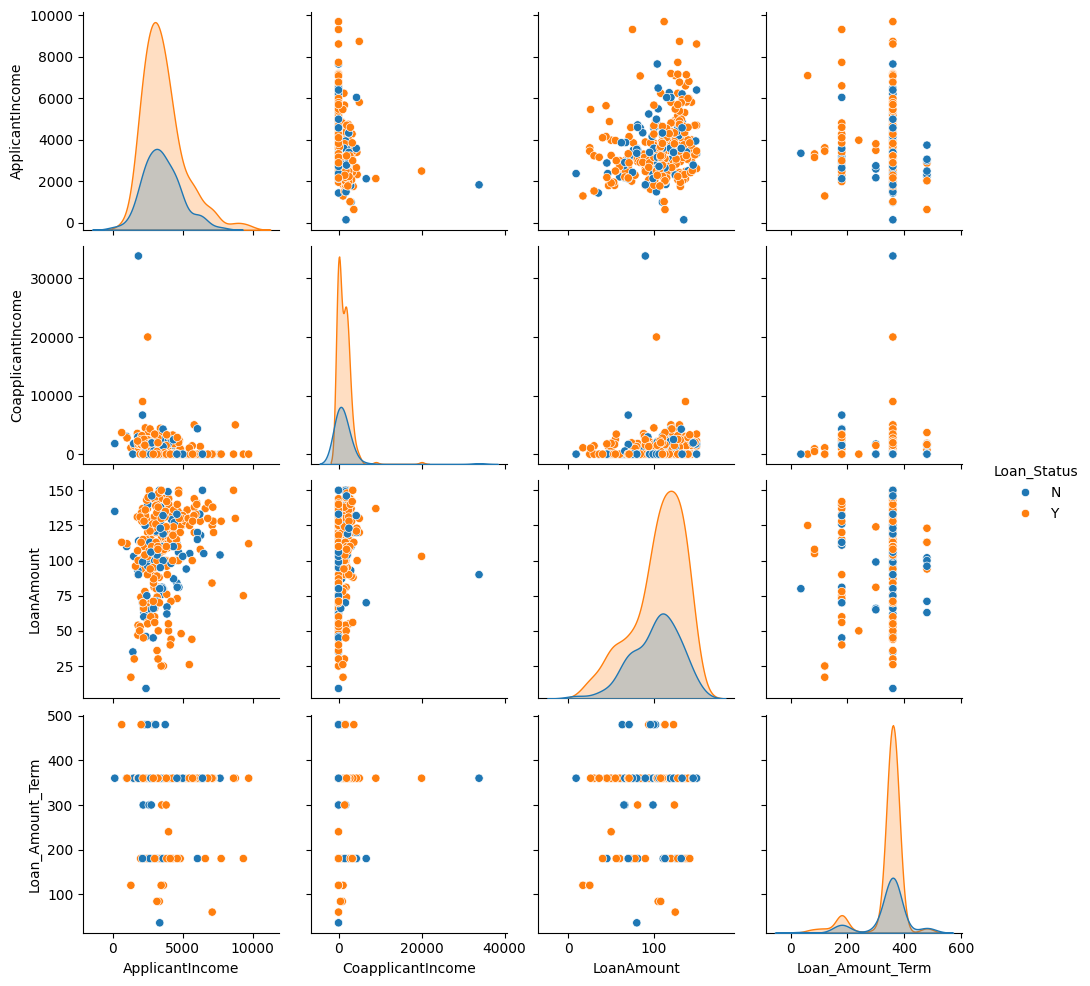


Figure 5 - Pairplot for numerical variables

Finally, a correlation matrix was visualized to observe the various relationships among the engineered features for the analysis:

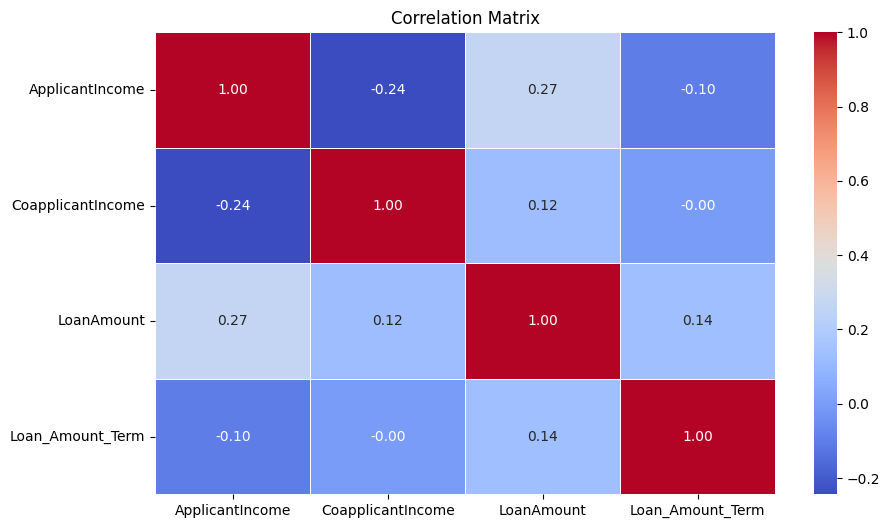


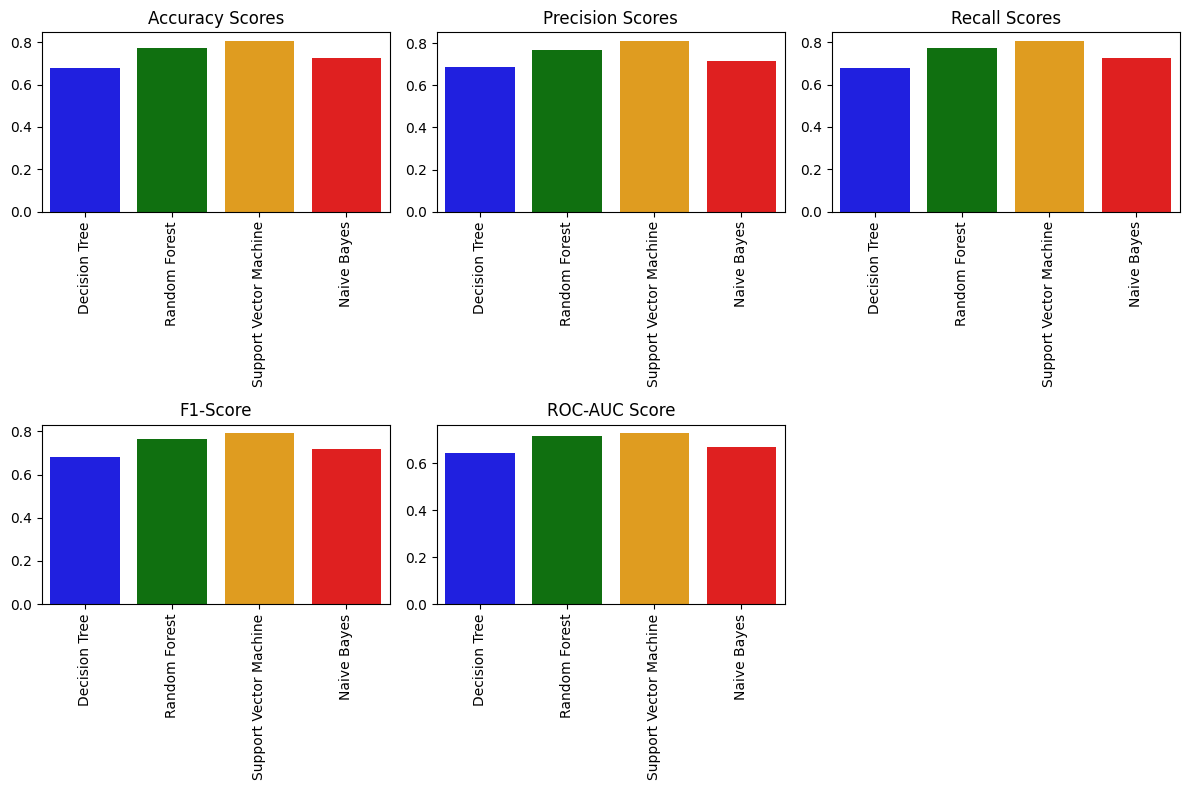
Figure 6 - Correlation Matrix

By adopting such a rigorous experiment methodology, we have developed very robust and reliable predictive models for the prediction of loan approval ultimately, and hence contribute to more efficient and more equitable lending practices in the financial industry.

**Results**

The experiment results provide insights into the performance of the four classification models in predicting loan approval. Additionally, feature importances from Decision Tree and Random Forest models offer valuable information about the significance of different features in the prediction process.

**Overall Models Evaluations:**



The table below summarizes the evaluation metrics for each model:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **ROC-AUC** |
| Decision Tree | 0.68 | 0.69 | 0.68 | 0.68 | 0.64 |
| Random Forest | 0.77 | 0.77 | 0.77 | 0.77 | 0.72 |
| Support Vector Machine | 0.81 | 0.81 | 0.81 | 0.79 | 0.73 |
| Naive Bayes | 0.73 | 0.72 | 0.73 | 0.72 | 0.67 |

**1. Decision Tree Classifier Feature Importances:**

|  |  |
| --- | --- |
| **Feature** | **Importance** |
| Credit\_History | 0.3647 |
| TotalIncome | 0.1959 |
| Debt\_Income\_Ratio | 0.1260 |
| LoanAmount | 0.0970 |
| Loan\_Amount\_Term | 0.0326 |
| Property\_Area\_Rural | 0.0312 |
| Gender\_Female | 0.0299 |
| Married\_Yes | 0.0282 |
| Dependents\_2 | 0.0268 |
| Education\_Graduate | 0.0215 |

**2. Random Forest Sorted Feature Importances:**

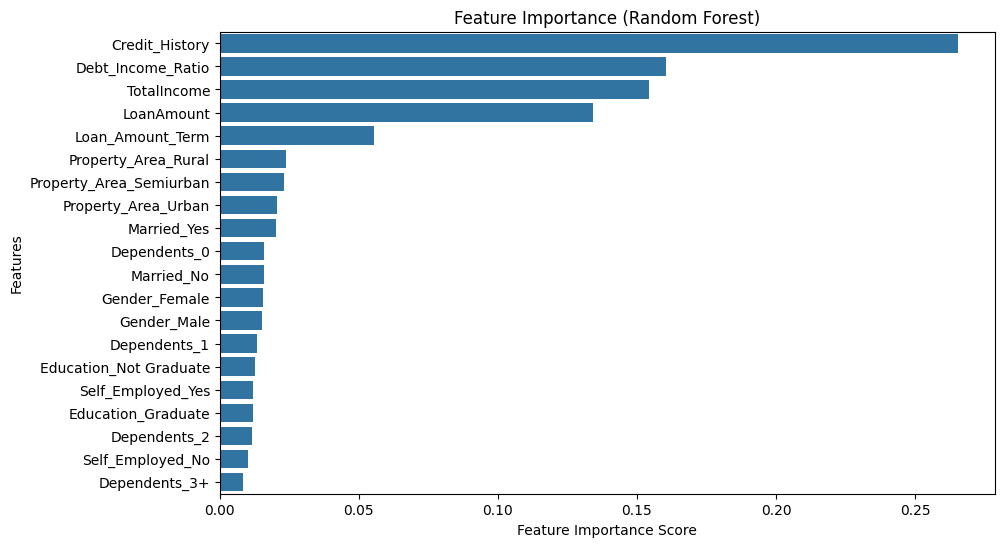


Figure 7 - Random Forest feature importance

The sorted feature importances from the Random Forest model are presented below in the above visualization.

**Discussion and Analysis**

* **Model Performance:** It is of high accuracy, precision, and recall among other models applied, which clearly denotes that the SVM model brings effectiveness in forecasting loan approval. The random forest follows in terms of giving a reasonably good score from all metrics, and decision tree and naive Bayes models have shown relatively poor performance compared to those above.
* **Feature Importance:** Credit\_History turned out to be the most dominating feature in both the Decision Tree and Random Forest models, hence bringing out its importance in the prediction of loan approval. Other important features are TotalIncome, Debt\_Income\_Ratio, LoanAmount, which show importance towards financial and demographic factors in the decision process.
* **Comparison between Models:** SVM shows the best accuracy and most of the evaluation metrics over other models, hence pointing out its suitability for the loan approval prediction task. The next-best performer, Random Forest, exhibits strong performance relative to SVM against the widest set of metrics. It only showed moderate performance; probably the model should be further optimized or feature engineered to improve prediction from Decision Tree and Naive Bayes.

Finally, the project allowed implementation of real-world prediction of approval and assessment of each of the model’s evaluation by prompting the user to enter various inputs related to the prediction of the loan and then predicting whether the user was approved for a loan or not based on the supplied features. The following screenshot showcases this in action:

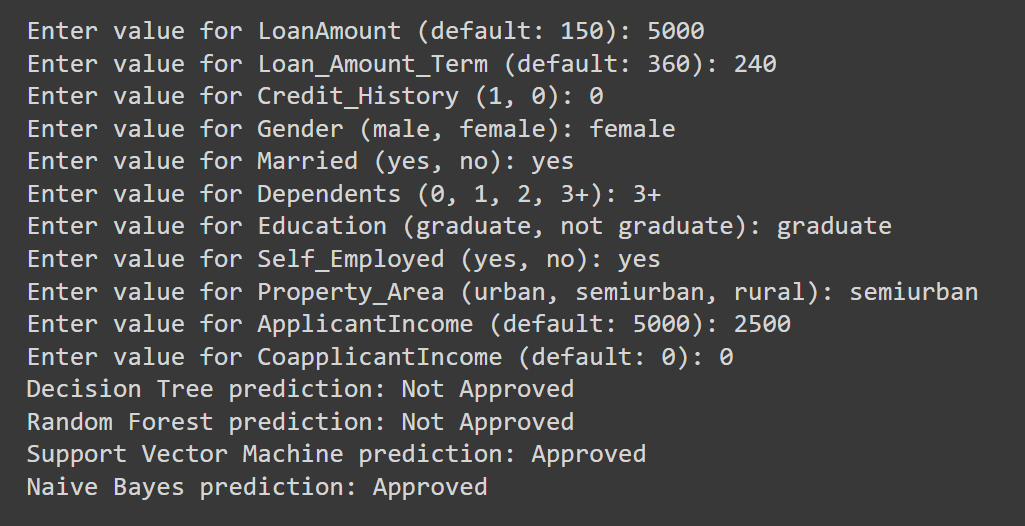


Figure 8 - Sample prediction based on user input

In summary, the experiment results highlight the effectiveness of Support Vector Machine and Random Forest models in predicting loan approval, underscoring the importance of feature engineering and model selection in building accurate and interpretable predictive models for financial decision-making.

**Conclusion**

It can be concluded that the SVM model outperforms other algorithms and is characterized by the highest number of accuracy and precision. On its part, Random Forest has shown competitive performance by several metrics. The Decision Tree and Naptive Bayes models also presented slightly inferior performance compared with the SVM and Random Forest. Feature importance analysis identifies important attributes being credit history, total income, debt to income ratio, and loan amount. These can be helpful for the financial institution to enable them to make wise decisions about the approval of loans.

However, this project is not without certain limitations and with some promising results. These can make additional data sources allow better model performance, while the entire set of the used dataset may not cover all the relevant factors affecting loan approval decisions. Besides, the performance of predictive models may not represent an appropriate assessment, and hence external validation over diversified datasets is required. The above probably can be taken up with other alternate sources of data, maybe from social media activity and transaction history, so as to further enhance the model strength in predicting than what it stands with this data set alone. New machine learning algorithms, that is ensemble methods and deep learning models, can be researched in the future to see where they make improvements in the model. In overall, this project provides valuable insight into the application of machine learning for prediction of loan approval and sets a stage of applicability of future research in the domain.Top of Form

**References:**

1. Diwate, Y., Rana, P., & Chavan, P. (2021). "Loan Approval Prediction Using Machine Learning." International Research Journal of Engineering and Technology (IRJET), 8(05), 1741-1745.
2. Cherif, A., & Berkane, N. (2023). "Loan Approval Prediction Using Machine Learning" (Doctoral dissertation).
3. Tumuluru, P., Burra, L. R., Loukya, M., Bhavana, S., CSaiBaba, H. M. H., & Sunanda, N. (2022, February). "Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms." In 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS) (pp. 349-353). IEEE.
4. Kadam, A. S., Nikam, S. R., Aher, A. A., Shelke, G. V., & Chandgude, A. S. (2021). "Prediction for loan approval using machine learning algorithm." International Research Journal of Engineering and Technology (IRJET), 8(04).