

Final Project Report

Group 25

Archit Singh

Sancia Saldanha

617-749-8646 (Archit Singh)


617-352-1569 (Sancia Saldanha)

singh.arc@northeastern.edu

saldanha.s@northeastern.edu

Percentage of Effort Contributed by Student1: 50%

Percentage of Effort Contributed by Student2: 50%

Signature of Student 1: 

Signature of Student 2: 

Submission Date: 04-12-2024

I. DATA COLLECTION:

The dataset used for the project ‘Loan Approval Prediction System’, was sourced from [Dataset Link](#). The dataset includes information on loan applicants, their financial records and relevant features for analysis.

II. DATA DESCRIPTION:

The dataset presented contains a comprehensive set of loan application records, with a total of 4269 entries. Each record comprises several financial and personal details of the applicants, which are used to determine their eligibility for loan approval.

Each entry is meticulously detailed, providing not only demographic information such as the number of dependents and education level but also extensive financial data including annual income, loan amount, and asset values across various categories. This rich dataset also includes CIBIL scores, which are critical in assessing the creditworthiness of the applicants.

The data is structured in such a way to give lending institutions a robust framework for decision-making. With fields ranging from basic identification numbers to complex financial metrics, this dataset is a valuable resource for analyzing lending risks and understanding the financial health of loan seekers. The outcome of each application is recorded as either 'Approved' or 'Rejected', offering a clear endpoint to each applicant's narrative within the dataset. One can view all the columns described below:

S.NO	VARIABLE	DATA TYPE	DESCRIPTION
1	loan_id	Numerical	A unique identifier for each loan application
2	no_of_dependents	Numerical	Reflects the number of individuals relying on the applicant's income
3	education	Categorical	Indicates the applicant's level of education, categorized as 'Graduate' or 'Not Graduate'

4	self_employed	Categorical	States whether the applicant is self-employed, with 'Yes' or 'No' as possible values
5	income_annum	Numerical	The yearly income of the applicant, provided in the local currency
6	loan_amount	Numerical	The amount of loan requested by the applicant
7	loan_term	Numerical	Duration of the loan repayment period in years
8	cibil_score	Numerical	The credit score of the applicant which is the creditworthiness of an individual
9	residential_assets_value	Numerical	The monetary value of the residential assets owned by the applicant
10	commercial_assets_value	Numerical	The value of commercial assets owned by the applicant
11	luxury_assets_value	Numerical	The worth of luxury items owned by the applicant
12	bank_asset_value	Numerical	The total value of assets that the applicant has in the bank
13	loan_status	Categorical	The final status of the loan application, either 'Approved' or 'Rejected'

Table 2(a) Variables Description

III. DATA EXPLORATION:

Our exploration started with basic statistics to understand the typical applicant. We understood that most applicants have around 2 to 3 people depending on their income. The income levels and loan amounts requested show a wide range, which means we have a diverse group of applicants. The typical time they'll take to repay the loan usually falls between 2 to 20 years.

	count	mean	std	min	25%	50%	75%	max
no_of_dependents	4269.0	2.498712e+00	1.695910e+00	0.0	1.0	3.0	4.0	5.0
education	4269.0	4.977747e-01	5.000536e-01	0.0	0.0	0.0	1.0	1.0
self_employed	4269.0	5.036308e-01	5.000454e-01	0.0	0.0	1.0	1.0	1.0
income_annum	4269.0	5.059124e+06	2.806840e+06	200000.0	2700000.0	5100000.0	7500000.0	9900000.0
loan_amount	4269.0	1.513345e+07	9.043363e+06	300000.0	7700000.0	14500000.0	21500000.0	39500000.0
loan_term	4269.0	1.090045e+01	5.709187e+00	2.0	6.0	10.0	16.0	20.0
cibil_score	4269.0	5.999361e+02	1.724304e+02	300.0	453.0	600.0	748.0	900.0
loan_status	4269.0	3.778402e-01	4.849042e-01	0.0	0.0	0.0	1.0	1.0
Movable_assets	4269.0	2.010300e+07	1.183658e+07	300000.0	10000000.0	19600000.0	29100000.0	53800000.0
Immovable_assets	4269.0	1.244577e+07	9.232541e+06	-100000.0	4900000.0	10600000.0	18200000.0	46600000.0

Figure 3(a) Mean Attributes

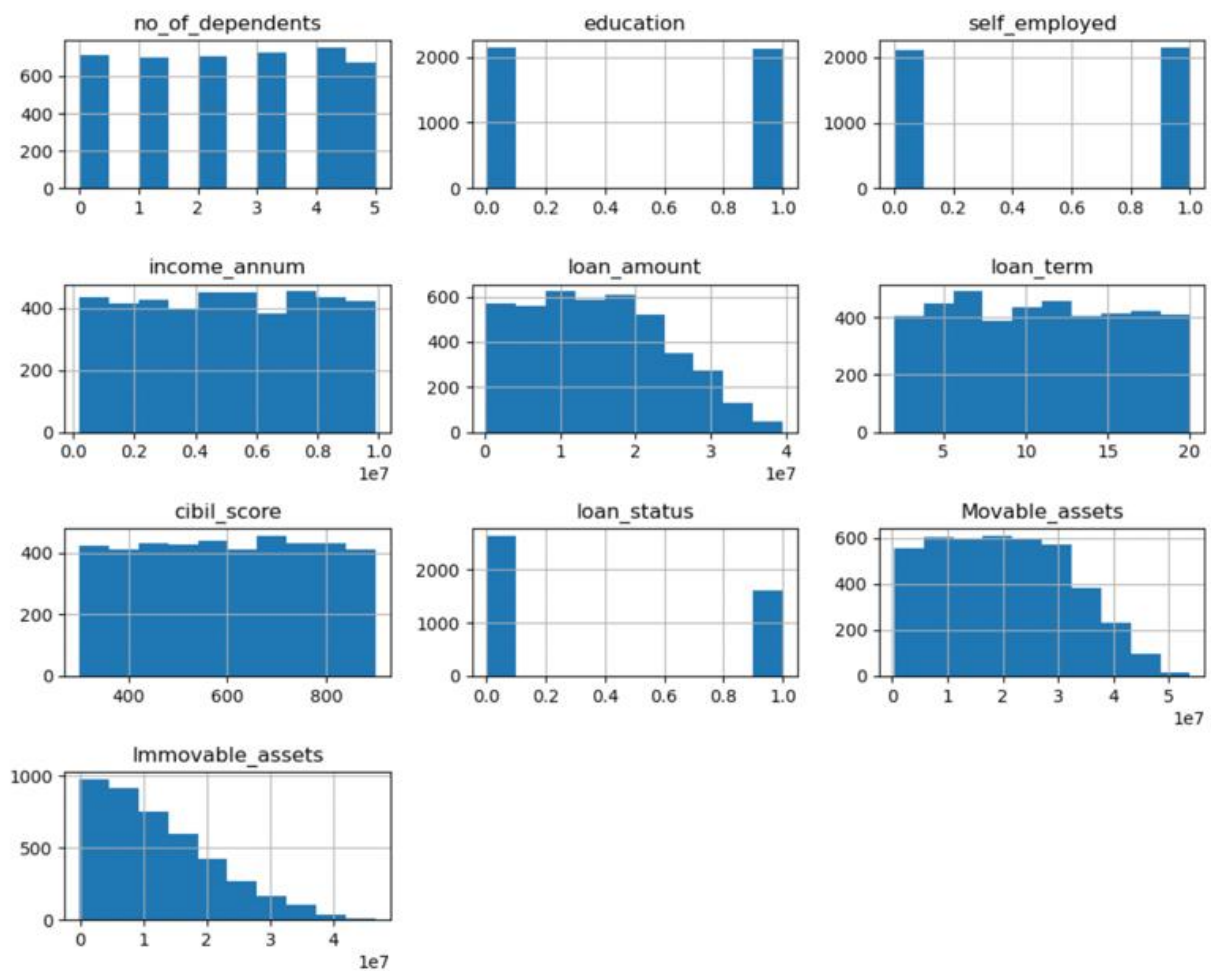


Figure 3(b) Variables Histograms

We also used histograms to visualize how many applicants fell into different categories, such as those with a college education or those who are self-employed. This helps us see the social and economic landscape of our applicants.

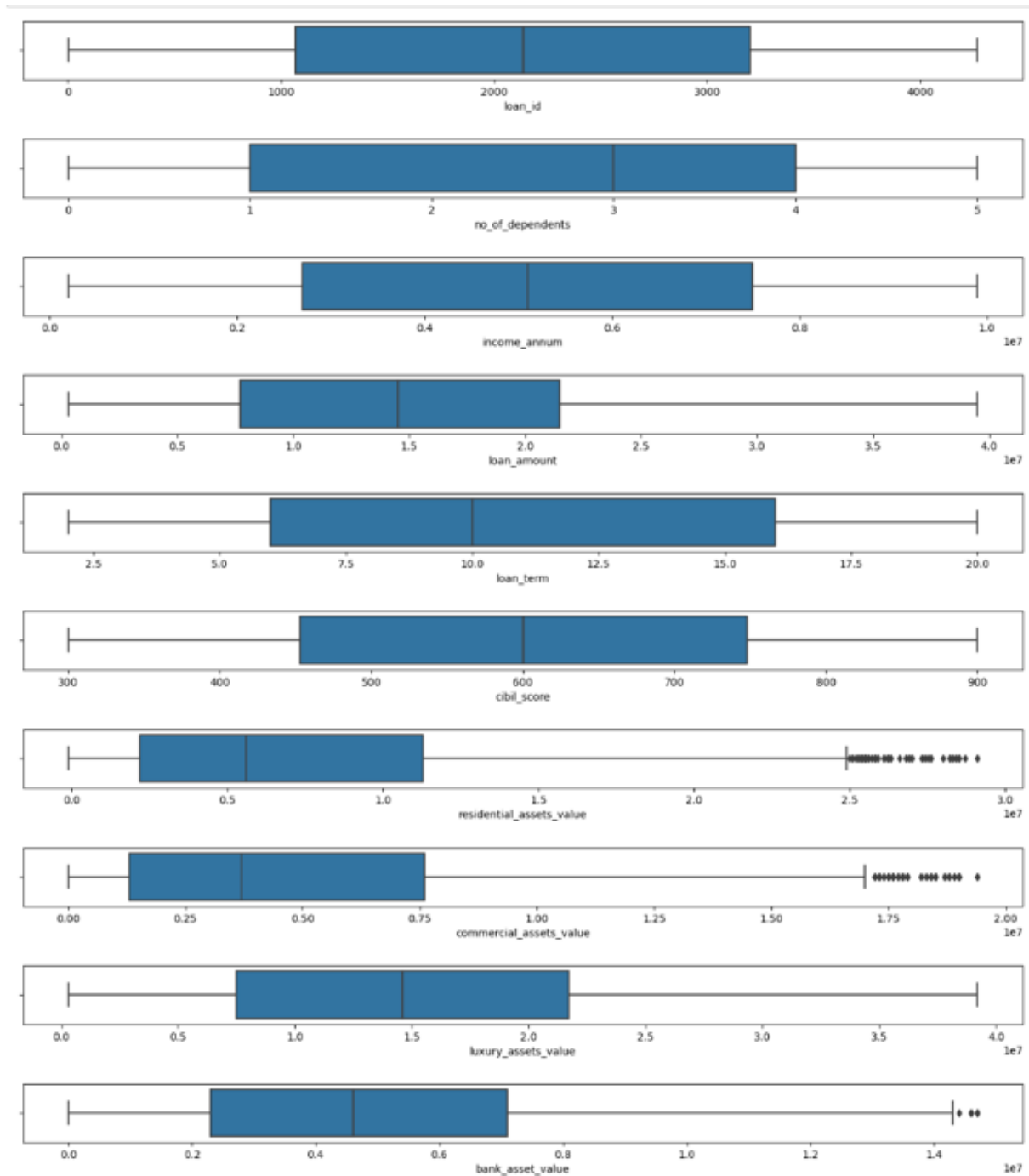


Figure 3(c) Variables Boxplot

The boxplots we created showed us the spread of the values of applicants assets. While we noticed some unusually high or low values, known as outliers, we chose to keep these in our study. These exceptional cases can provide insights into unusual but possible financial situations that we might otherwise overlook.

Furthermore, there seems to be a relatively consistent number of approved and rejected loans across applicants with different numbers of dependents. And the proportion of approved to rejected loans do not appear to vary significantly with the number of dependents, suggesting that the number of dependents might not be a strong standalone factor in the loan approval process.

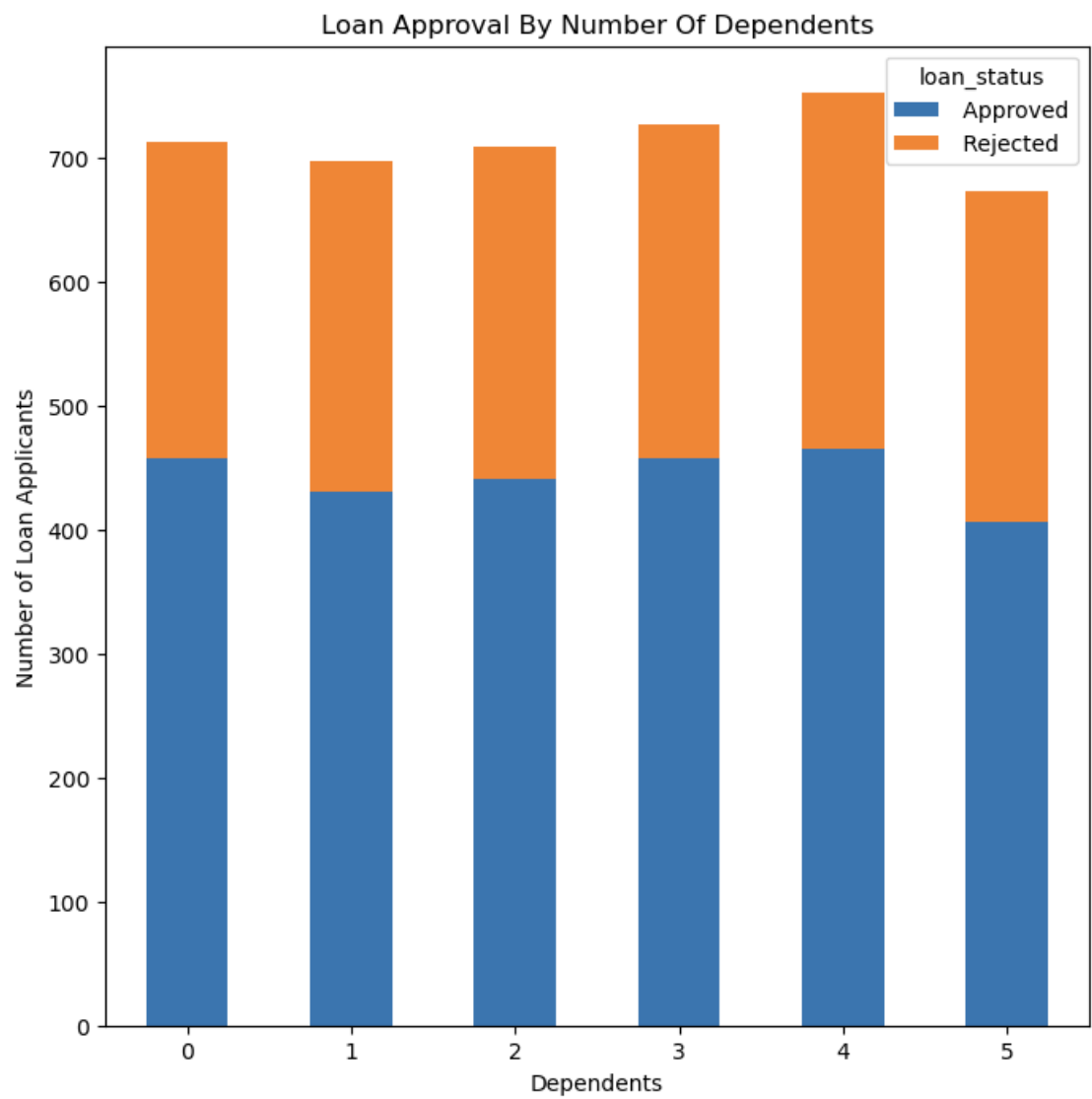


Figure 3(d) Loan Approval by number of dependents

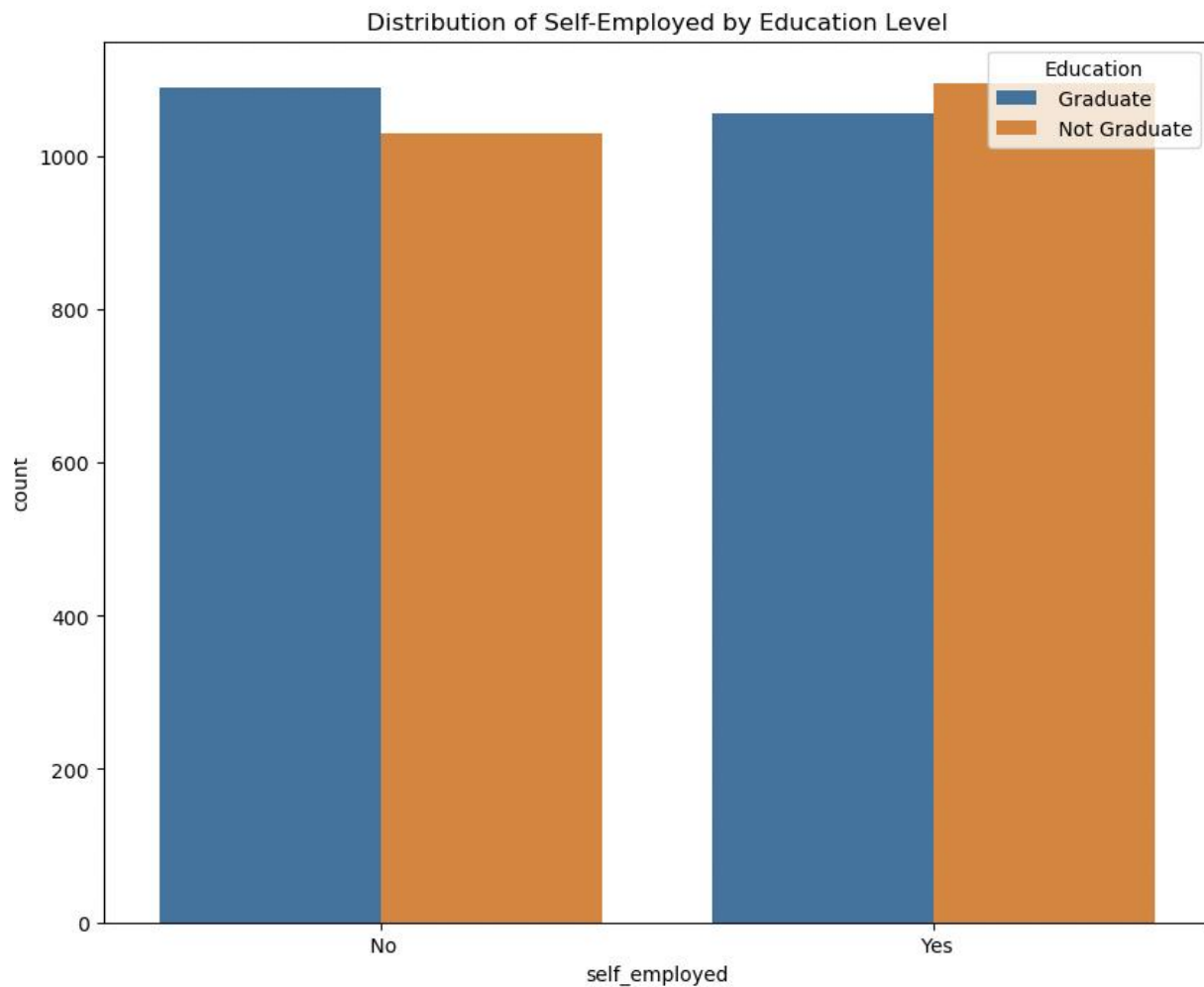


Figure 3(e) Distribution of Self-Employed by Education Level

The majority of applicants are not self-employed regardless of their education level.

There is a relatively similar distribution of self-employed individuals among graduates and non-graduates, indicating that education level might not be a decisive factor in choosing self-employment.

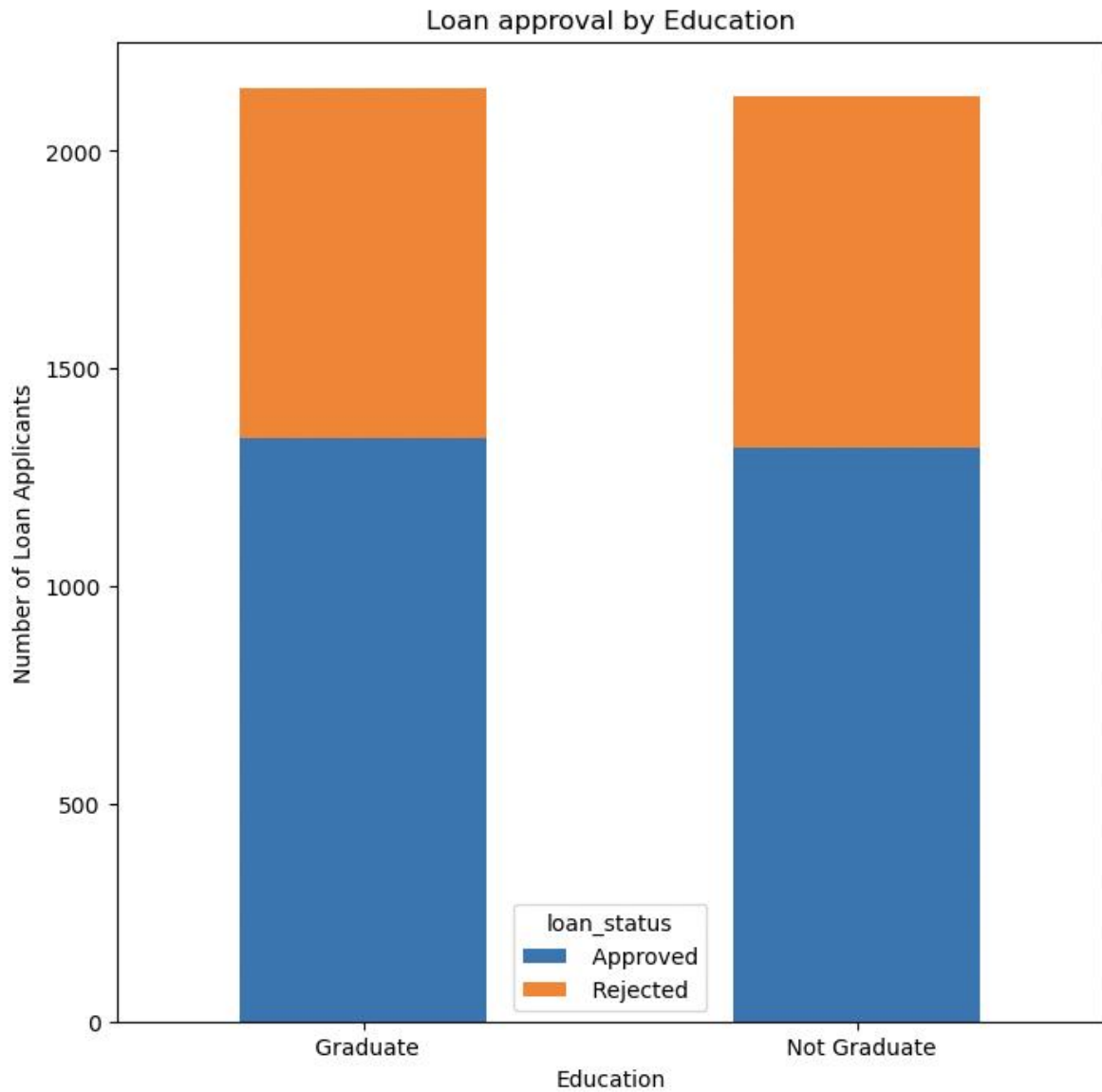


Figure 3(f) Loan approval by education

Both graduates and non-graduates experience loan rejections and approvals, with graduates receiving slightly more approvals than rejections compared to non-graduates.

Education seems to have some influence on loan approval, but there are approvals and rejections within both educational categories

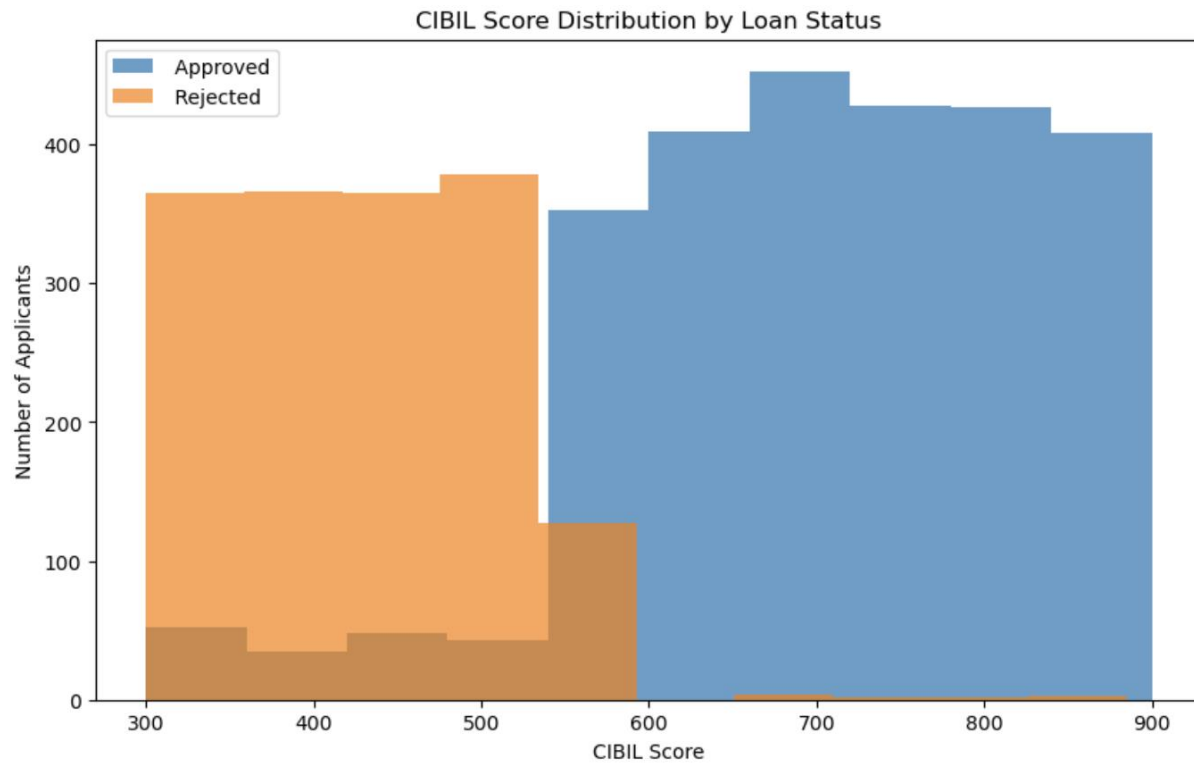


Figure 3(g) CIBIL Score Distribution by loan status

Moreover, we then looked closely at the CIBIL scores, a measure of credit health. Scores in our dataset range from 300, which is considered poor, to 900, which is excellent. The middle score is around 600, suggesting most applicants have a fair credit background.

Lastly, we used a correlation matrix to see how different financial aspects relate to each other. Not surprisingly, we found that people with higher incomes and more valuable assets tend to ask for bigger loans.

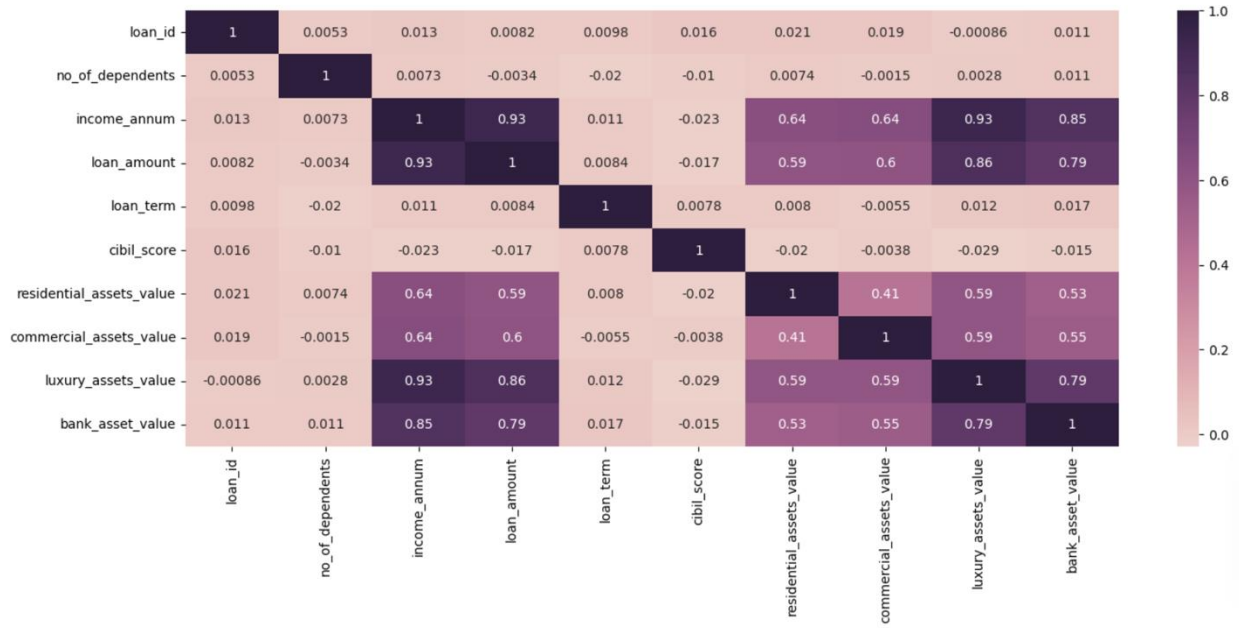
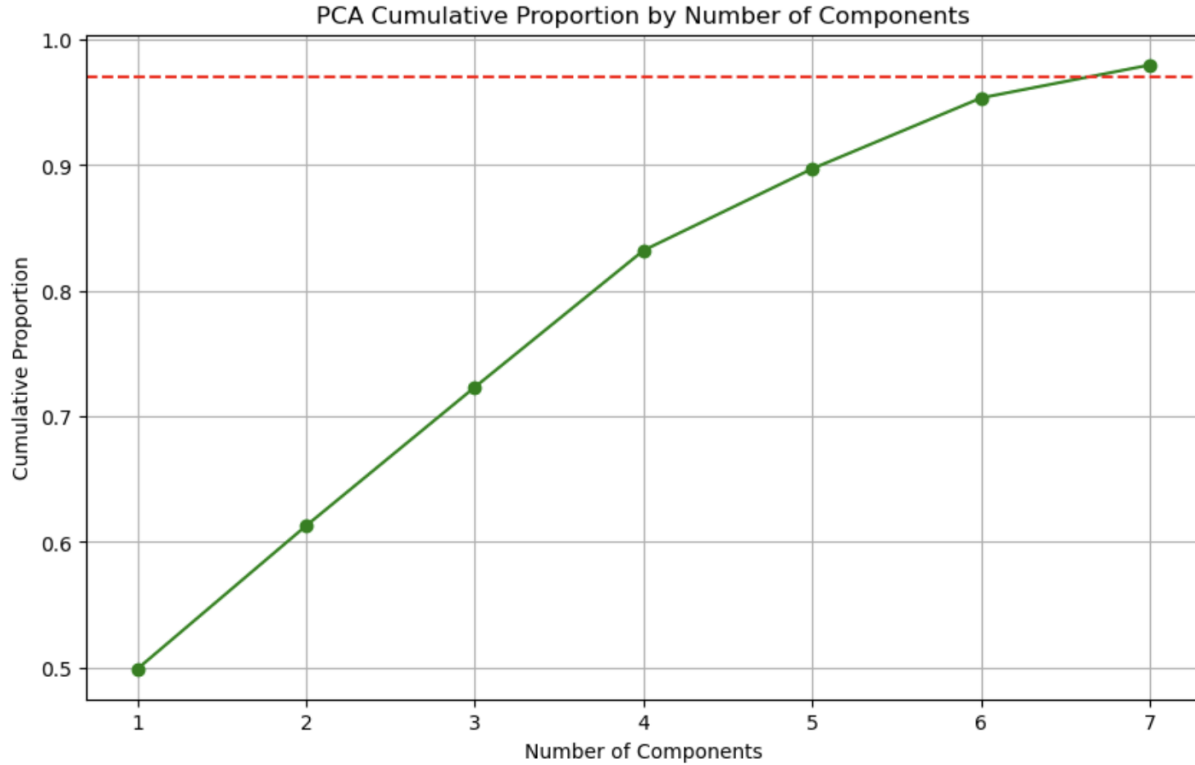


Figure 3(h) Correlation Matrix

IV. DATA PREPROCESSING:

We started by adjusting the scale of the numerical data, such as yearly income and loan amounts. Then, we used a method called Principal Component Analysis to simplify our dataset and to perform dimension reduction.

We understood that seven components capture over 97% of the variance, suggesting that these components can effectively represent the original data with reduced complexity. Post PCA, we mapped PCA components back to the original variables to understand their contribution to each principal component and sorted them in descending order.



Number of components capturing over 97% variance: 7

Figure 4(a) PCA Cumulative Proportion by number of components

```
[('commercial_assets_value', 1.7945595831666978),
 ('residential_assets_value', 1.7661416700989319),
 ('loan_term', 1.7460571236334357),
 ('bank_asset_value', 1.6704041236805671),
 ('no_of_dependents', 1.6364068529044222),
 ('cibil_score', 1.4919605094177963),
 ('luxury_assets_value', 1.0579037626174934),
 ('loan_amount', 1.0430490090380748),
 ('income_annum', 0.8086599832946535)]
```

Figure 4(b) Features sorted in descending order by PCA

But from our understanding of the loan business domain, we know that features such as ‘loan_amount’ and ‘income_annum’ play a significant role in determining whether an applicant's loan application will be accepted or rejected.

Therefore, we decided not to remove them and instead create a copy of the dataframe, namely `df_pca`, in which these columns were dropped to determine if the model performs better without these columns in the future.

Additionally, the column `'loan_id'` was dropped as it is used only as a unique identifier for the applicants.

The categorical variables, such as education, self-employed and loan status, were encoded to transform them into a numerical format. Ordinal encoding was used to convert each unique category into an integer value. This is necessary for most machine learning models, which require numerical input.

Furthermore, we understood that features `'bank_asset_value'` and `'luxury_assets_value'` can be combined to `'Movable_assets'`. And `'residential_assets_value'` and `'commercial_assets_value'` can be combined to `'Immovable_assets'` without losing critical information.

V. DATA MINING MODELS

We have delved into our dataset post-data-preprocessing and cleaning to focus on determining factors that influence loan approval outcomes. This dataset contains detailed attributes for 4,268 loan applicants, including but not limited to the number of dependents, education level, self-employment status, annual income, requested loan amount, loan term duration, CIBIL scores, and the final loan status.

The target variable, `'loan_status'`, originally categorised as Approved or Rejected, has been transformed into a numerical format, with values encoded to 0.0 for Rejected and 1.0 for Approved. This modification aids in applying various machine learning models to predict loan approval outcomes.

We understood that we have to explore the below model to determine the best fit for our dataset:

1. Random Forest
2. Logistic Regression
3. Decision Tree
4. Support Vector Machine

5. K-Nearest Neighbors
6. Neural Networks

- Logistic Regression:

A staple in binary classification tasks, Logistic Regression models the probability of a certain class or event, such as loan approval, based on one or more independent variables. It's well-suited for situations where the outcome is dichotomous. However, it presumes a linear relationship between the independent variables and the logarithm of the odds (log-odds). In our dataset, which exhibits complex, non-linear interactions between factors like income and credit score, this assumption could lead to inadequate modeling of the underlying relationships, thereby skewing the predictions.

- Decision Trees:

This model uses a tree-like graph or model of decisions and their possible consequences. It's particularly intuitive, as it mirrors human decision-making processes. However, while Decision Trees could theoretically capture the nuances in our dataset by segmenting it based on feature thresholds, they are prone to creating overly complex trees that do not generalize well (overfitting), especially with datasets that, like ours, combine diverse types of variables and extensive ranges in values.

- Support Vector Machine (SVM):

SVM is a powerful classifier that works by finding the hyperplane that best divides a dataset into classes. It is particularly adept at handling high-dimensional spaces, which is beneficial for datasets with many attributes. Nevertheless, the application of SVM to our mixed-type dataset would require extensive preprocessing to ensure all data is on a similar scale, and the algorithm's complexity increases significantly with the size of the data, which could make it computationally intensive and less transparent, particularly when explaining loan decisions.

- K-Nearest Neighbors (KNN):

KNN operates on the principle that similar things exist in close proximity. In essence, it looks at the 'k' closest labeled data points and decides the label based on majority vote. While conceptually

simple and effective in classification, KNN's performance heavily depends on the distance metric and the normalization of data scales. Given our dataset's diverse range of financial figures and personal attributes, the model might yield suboptimal predictions without significant preprocessing and could suffer from the curse of dimensionality.

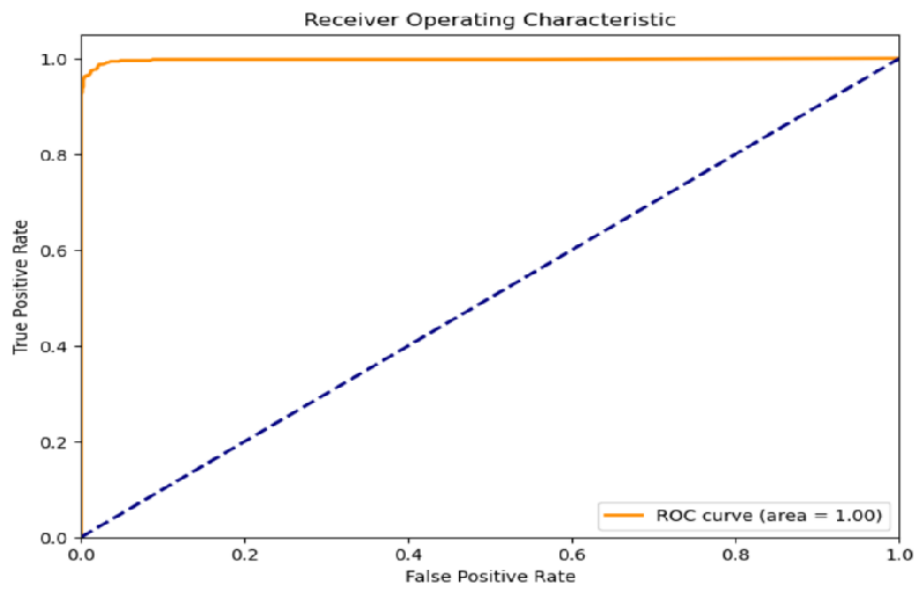
- Neural Networks:

These consist of layers of interconnected nodes or neurons that can learn complex relationships between input variables and outcomes through training. Neural Networks are capable of modeling highly complex, non-linear relationships, making them potentially effective for our multifaceted dataset. However, they require large amounts of data to train effectively, are computationally expensive, and their "black box" nature makes it challenging to interpret which specific features are influencing the loan approval decisions, an essential aspect for transparency in financial services.

After careful consideration, we understood and finalized that the Random Forest approach would be the best possible data mining algorithm. This method constructs multiple Decision Trees during training and outputs the mode of the classes (classification) of the individual trees for a more accurate and robust prediction. It naturally handles mixed data types and is less prone to overfitting compared to individual Decision Trees. Its ensemble nature allows it to capture complex interactions between variables effectively, making it particularly suitable for our dataset, which encompasses a wide range of financial and personal information. Additionally, Random Forest provides insights into feature importance, offering a clearer understanding of what drives loan approval decisions, thus aligning with the need for transparency and interpretability in our analysis.

VI. PERFORMANCE EVALUATION:

1. Random Forest :



Evaluation Metrics:

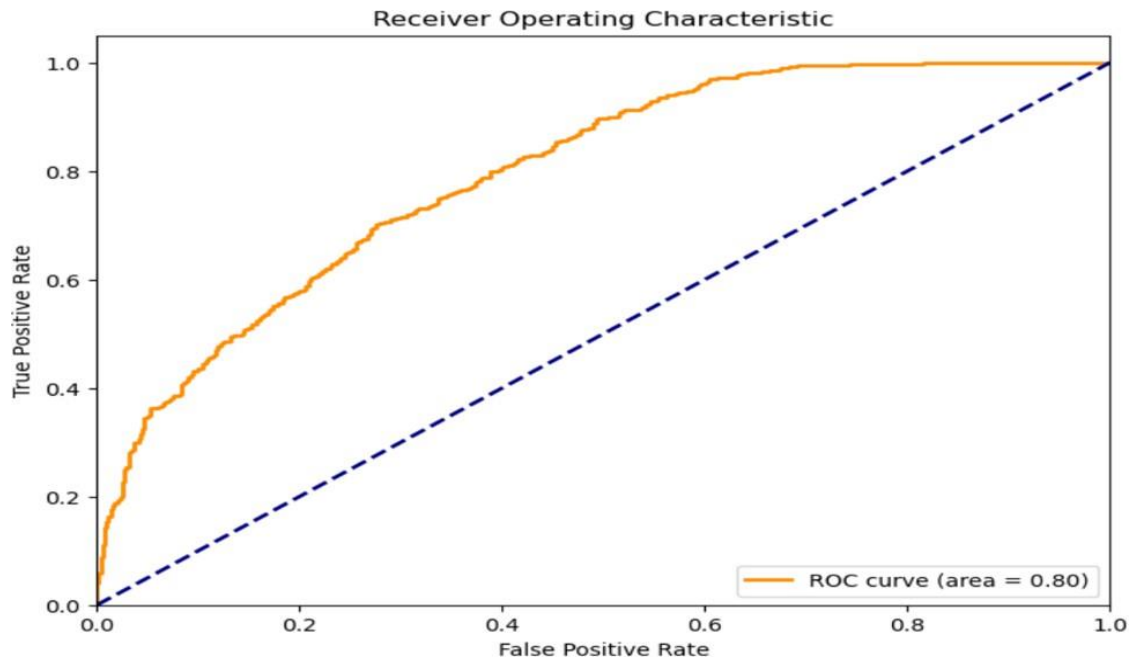
Metric	Value
Training Error	0.0000
Test Error	0.0172
Accuracy	98.28%
Validation Error	0.0172
Sensitivity	96.39%
Specificity	99.38%
F1 Score	97.63%

Confusion Matrix:

	Predicted Positive	Predicted Negative
Actual Positive	805	5
Actual Negative	17	454

Interpretation: The Random Forest classifier demonstrates exceptional performance across all metrics, showing its robustness in accurately predicting both classes with minimal error.

2. Logistic Regression:



Evaluation Metrics:

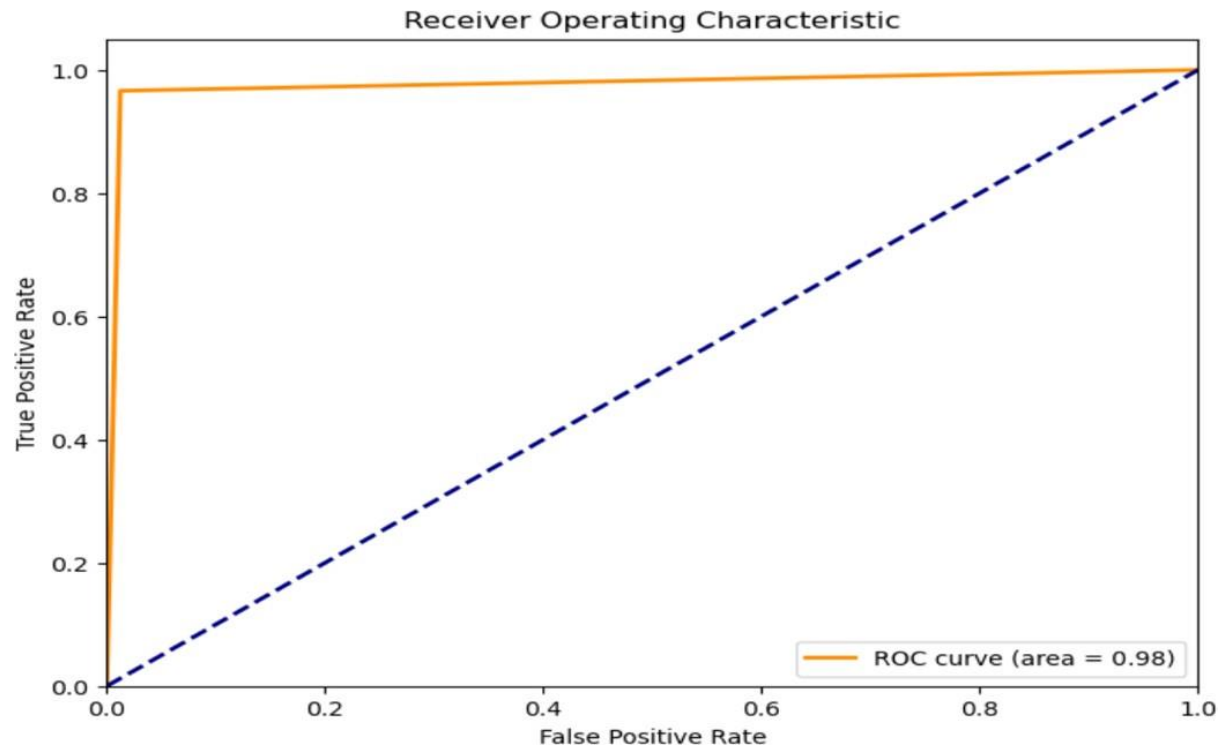
Metric	Value
Training Error	0.2654
Test Error	0.2732
Accuracy	72.68%
Validation Error	0.2732
Sensitivity	37.37%
Specificity	93.21%
F1Score	50.14%

Confusion Matrix:

	Predicted Positive	Predicted Negative
Actual Positive	755	55
Actual Negative	295	176

Interpretation: Logistic Regression shows potential in specificity but needs improvement in sensitivity, highlighting a tendency towards predicting the majority class.

3. Decision Tree:



Evaluation Metrics:

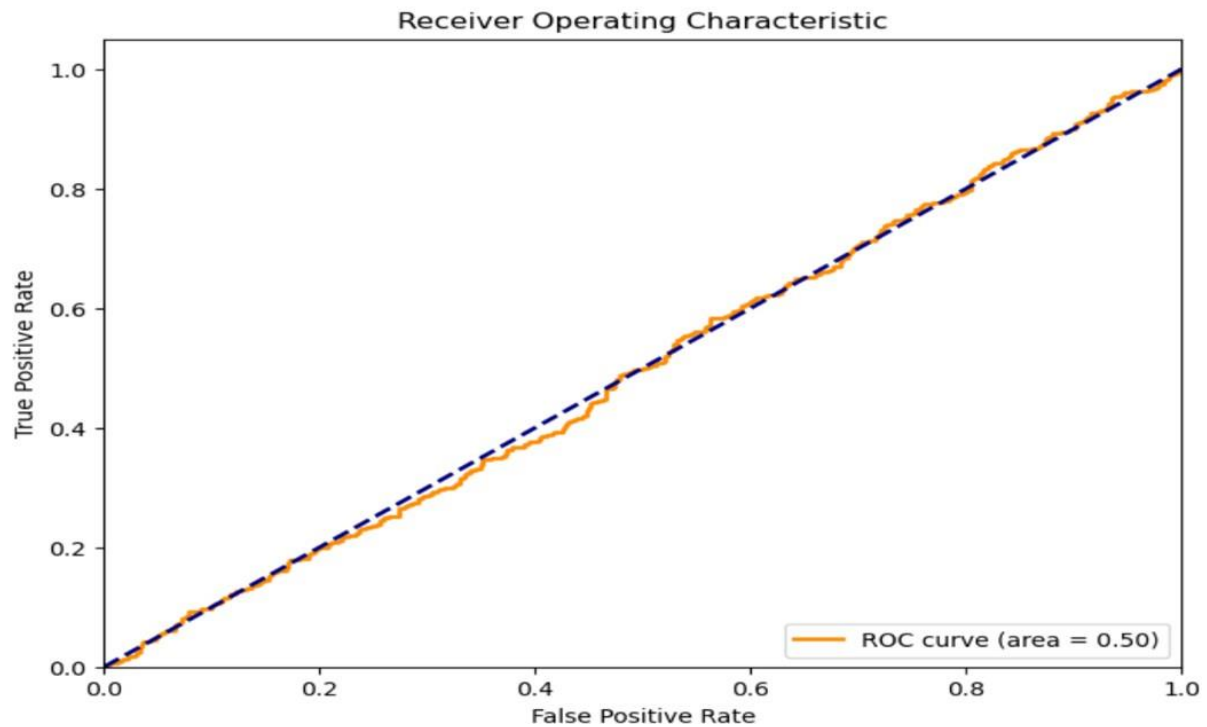
Metric	Value
Training Error	0.0000
Test Error	0.0203
Accuracy	97.97%
Validation Error	0.0203
Sensitivity	96.60%
Specificity	98.77%
F1 Score	97.22%

Confusion Matrix:

	Predicted Positive	Predicted Negative
Actual Positive	800	10
Actual Negative	16	455

Interpretation: The Decision Tree classifier's high performance across all metrics indicates its effective balance in predicting both classes with high accuracy.

4. Support Vector Machine:



Evaluation Metrics:

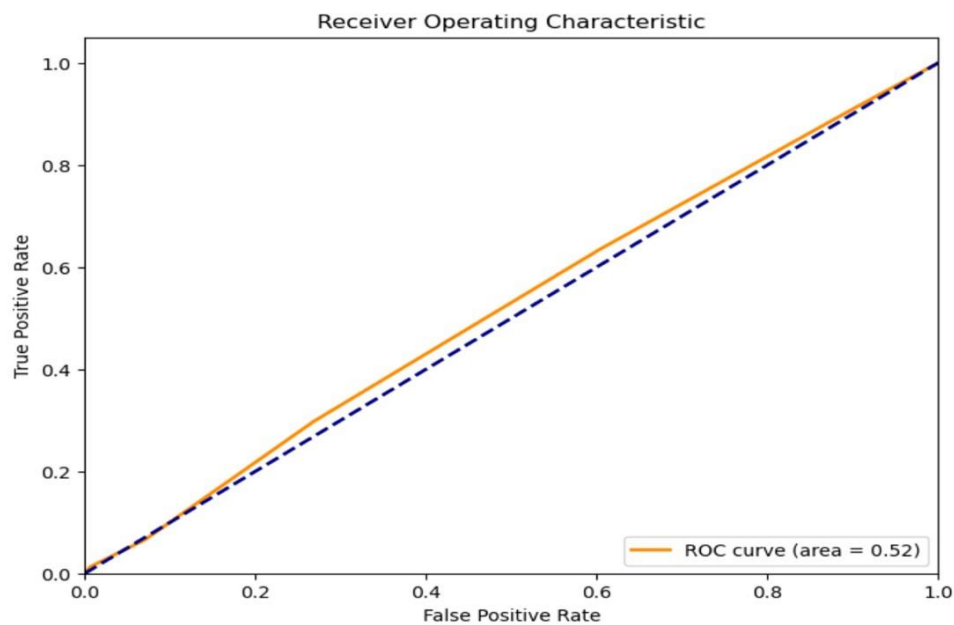
Metric	Value
Training Error	0.3822
Test Error	0.3677
Accuracy	63.23%
Validation Error	0.3677
Sensitivity	0.00%
Specificity	100.00%
F1 Score	0.00%

Confusion Matrix:

	Predicted Positive	Predicted Negative
Actual Positive	810	0
Actual Negative	471	0

Interpretation: SVM's unique outcome of predicting all instances as a single class indicates significant issues, highlighting the need for model re-evaluation or parameter adjustment.

5. K-Nearest neighbors (KNN):



Evaluation Metrics:

Metric	Value
Training Error	0.2855
Test Error	0.4278
Accuracy	57.22%
Validation Error	0.4278
Sensitivity	29.72%
Specificity	73.21%
F1 Score	33.82%

Confusion Matrix:

	Predicted Positive	Predicted Negative
Actual Positive	593	217
Actual Negative	331	140

Interpretation: KNN's lower performance indicates challenges in model accuracy, particularly in correctly identifying positive instances, suggesting a need for model optimization.

VII. PERFORMANCE INTERPRETATION:

1. The Random Forest model has the highest overall accuracy of 98.28% and the lowest validation error of 1.72% in the classification of phishing and legitimate website instances. It also has the highest F1 score of 97.63%. With the highest sensitivity value of 96.39% and specificity of 99.38%, it is evident that the Random Forest Classifier is the most effective model among the ones evaluated for accurately classifying the instances. This model demonstrates a superior balance in recognizing both classes with minimal errors making it the best choice for our application. The best contributors have been evaluated using feature importance scores, as shown in Fig 7 a & b.

2. The Decision Tree classifier follows closely in terms of performance with an accuracy of 97.97%. It exhibits a slightly higher validation error of 2.03% compared to the Random Forest model. The Decision Tree has an F1-score of 97.22%, a sensitivity of 96.60%, and a specificity of 98.77%, showcasing its capability to classify instances with high precision. Its balance in identifying true positives and true negatives nearly matches the Random Forest, positioning it as the second most effective model for this task.

3. Logistic Regression stands out with an accuracy of 72.68%, which is significantly lower compared to the top two models. It has a validation error of 27.32%, a sensitivity of 37.37%, and a specificity of 93.21%. Despite its relatively high specificity, the lower sensitivity and F1-score of 50.14% highlight challenges in correctly classifying positive instances, indicating areas for improvement in model tuning and optimization for better performance.

4. The Support Vector Machine (SVM) model shows a unique outcome with an accuracy of 63.23%, the lowest among the models evaluated. Its validation error stands at 36.77%, and it interestingly reports a sensitivity of 0.00% and a specificity of 100.00%. The model's inability to identify any positive instances while accurately classifying all negative instances suggests significant limitations in its current configuration for this application, requiring substantial adjustments.

5. K-Nearest Neighbors (KNN) has an accuracy of 57.22%, making it the least accurate model in our evaluation. It presents a validation error of 42.78%, a sensitivity of 29.72%, and a specificity of 73.21%. The model's lower F1-score of 33.82% further emphasizes its challenges in effectively classifying instances, particularly in identifying true positives, which marks it as the least suitable model for our specific task based on the current dataset and parameters.

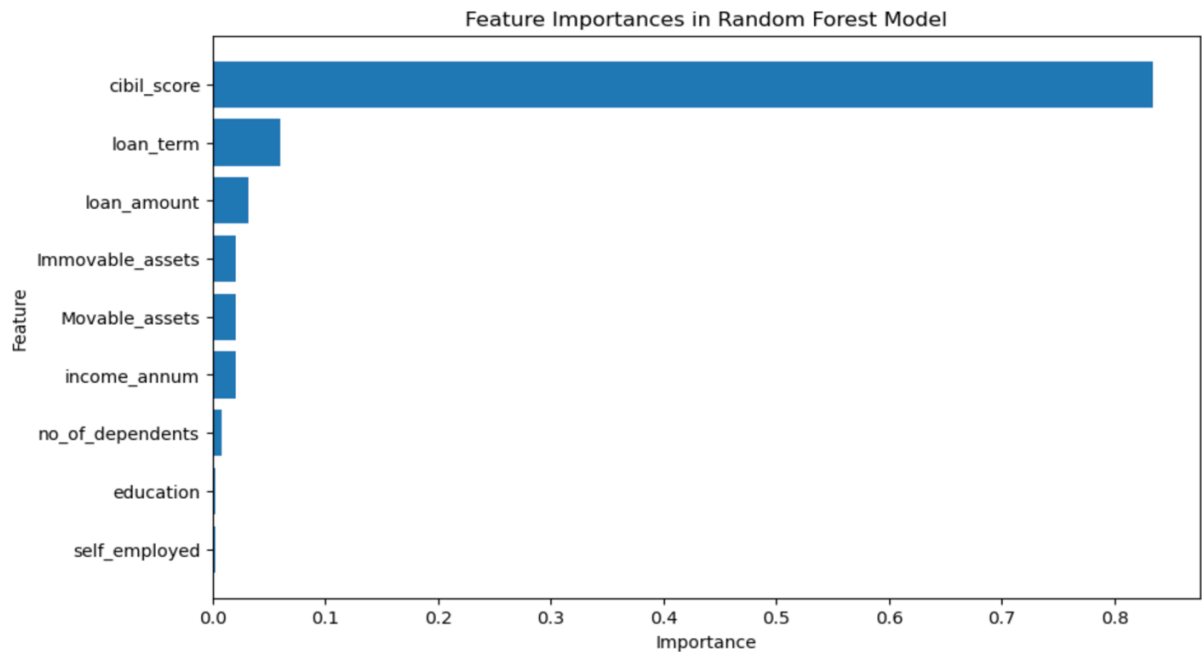


Figure 7 (a) Feature Importance Bar Chart

Feature Importances for Random Forest:		
	Feature	Importance
6	cibil_score	0.834080
5	loan_term	0.060093
4	loan_amount	0.031787
8	Immovable_assets	0.020397
7	Movable_assets	0.020313
3	income_annum	0.020228
0	no_of_dependents	0.008443
1	education	0.002412
2	self_employed	0.002247

Figure 7 (b) Feature Importance for Random Forest

Evaluating Overfitting and Underfitting:

From Fig 7 (c), it is evident that these classification models exhibit low training and low test error. This implies that these models do not underfit or overfit the data. Hence, they can be used for the classification of phishing and legitimate website instances.

	Model Name	Training Error	Test Error
0	Random Forest	0.000000	0.017174
1	Logistic Regression	0.265395	0.273224
2	Decision Tree	0.000000	0.020297
3	Support Vector Machine	0.382195	0.367681
4	K-Nearest Neighbors	0.285475	0.427791

Figure 7 (c), Comparison Chart of the Train and Test Error for all different models

VIII. IMPACT OF THE PROJECT OUTCOMES

The deployment of this predictive lending framework heralds a transformative shift in evaluating loan applications. Anchored by sophisticated analytical models, the framework scrutinizes an applicant's financial footprint, fusing variables like repayment histories, income streams, and requested loan parameters. The strategic incorporation of a multifaceted algorithm, akin to the Random Forest technique, elevates the accuracy of forecasts. This algorithm thrives in complex data terrains, adeptly sidestepping the pitfalls of overfitting while juggling a myriad of input factors. By amalgamating predictions across a spectrum of decision trees, it yields a more nuanced and stable prediction continuum.

The ultimate aim transcends mere risk reduction for lending institutions. It encapsulates a dual-fold purpose: safeguarding the lender's assets while democratizing loan accessibility for bona fide

applicants. This equilibrium not only stabilizes the financial ecosystem but also engenders a paradigm of equitable opportunity in financial undertakings. The sophistication of this system lies not in its ability to predict but in its potential to learn and adapt, ensuring that decisions evolve in tandem with emerging economic patterns and borrower behaviours.