International Journal of Banking, Risk and Insurance
12 (2) 2024, 92-101
http://publishingindia.com/ijbri/
doi: https://doi.org/10.21863/ijbri/2024.12.2.008

Submitted: 10 September, 2023
Accepted: 18 December, 2023

Unlocking Loan Approval: Data-Driven Insights and Machine Learning Precision

Sameer Jain*

Abstract

This research study explores the realm of loan approval prediction, leveraging a comprehensive dataset encompassing a wide array of applicant-related variables, including loan parameters, credit scores, education, and asset holdings. The primary objective is to construct an effective predictive model to aid lending institutions in making informed decisions regarding loan applications. Furthermore, this study identifies key factors that significantly influence loan approval determinations, offering insights into prioritising services for applicants with higher approval probabilities. The dataset employed in this study comprises critical financial information typically used to assess loan eligibility, such as CIBIL scores, income, employment status, loan terms, loan amounts, asset values, and loan status. Employing advanced machine learning and data analysis techniques, this research develops predictive models capable of estimating the likelihood of loan approval based on these features. Key findings reveal that a higher CIBIL score positively correlates with increased chances of loan approval, while a larger number of dependents diminishes approval probabilities. Additionally, applicants with more substantial assets, encompassing both movable and immovable holdings, are more likely to secure loan approval. Furthermore, individuals requesting higher loan amounts with shorter tenures exhibit greater odds of approval. The research employs machine learning models, including the Decision Tree Classifier and Random Forest Classifier, to forecast loan approval outcomes. These models yield impressive accuracies of 92.98% and 90.98%, respectively, with the Decision Tree Classifier outperforming the Random Forest Classifier within this context.

Keywords: Loan Approval Prediction, Credit Risk Assessment, Machine Learning Models, Financial Decision-Making, Lending Institutions

Introduction

In the contemporary financial landscape, the efficient and accurate evaluation of loan applications is of paramount importance to lending institutions. Several factors connected to the applicant, including credit scores, income, loan terms, and asset holdings, interact in a complicated way while deciding whether to accept or reject a loan application. Accurate forecasting of loan approval results not only reduces credit risk but also ensures that financial resources are utilised in the best possible way. Therefore, the current study aims to address this significant problem by utilising cutting-edge data analysis methods and machine learning models to provide a predictive framework for loan acceptance.

The primary objective of this work is to develop predictive models that assist lending institutions assess the probability of loan acceptance using a large dataset that includes a variety of applicant-related information. This research intends to enhance the decision-making processes of banks and financial institutions in order to enable them to simplify operations, decrease human labour, and give priority services to applicants with a higher likelihood of loan acceptance. In order to achieve this goal, the study also aims to pinpoint important elements that have a substantial impact on loan approval choices, illuminating the crucial elements that lenders should consider when assessing loan applications. Researchers and financial professionals have given the creation of predictive models for loan acceptance a lot of attention. Based on application characteristics, these algorithms calculate the likelihood of loan acceptance using past data. In order to develop precise prediction models, a number of machine learning approaches, such as Decision Trees, Random Forests, Logistic Regression, and Neural Networks, have been used (Jiang & Wei,

* Faculty, NICMAR Business School, National Institute of Construction Management and Research (NICMAR) University, Balewadi, Pune, Maharashtra, India. Email: sameerjain@nicmar.ac.in

2018; Thomas & Indirubin, 2019). These models have been shown to have the ability to considerably increase the effectiveness of the loan approval process and lower the risk of defaults. A well-established factor in loan acceptance is creditworthiness as determined by credit scores (Agarwal & Tufano, 2018). Higher credit scores are often perceived viewed as less hazardous and increase the likelihood that loan applications will be accepted. Also crucial in loan acceptance choices are income and work status (Mester & Nakamura, 2018). Applicants with steady job and sufficient money to cover their repayment commitments often receive preference from lenders. The value of applicants' assets has been acknowledged as a significant determinant in loan acceptance decisions in addition to conventional criteria (Barr, 2020). Applicants with significant assets, such as real estate holdings and liquid assets, are frequently viewed as being in a stronger solid financial position and may be granted loans with more advantageous conditions. The accuracy of loan approval projections has shown incredible potential when using machine learning approaches. Widely used in this setting are decision tree models, which are renowned for being easy to understand and utilise (Mishra et al., 2017). Due to its capacity to manage complicated data patterns, Random Forests, an ensemble learning approach, have also grown in favuor (Huang et al., 2018). These machine learning models offer an efficient way to combine multiple factors in order to make complex and informed loan approval choices.

A machine learning technique called a decision tree classifier is utilised for both classification and regression tasks. It creates a model in the shape of a tree-like structure, where a leaf node represents the result (a class label in classification or a numeric value in regression), an edge indicates a decision rule, and an inside node represents a feature or attribute. A decision tree classifier, when used for classification tasks, recursively divides the dataset into subsets based on the values of input features until it reaches leaf nodes, where the final class label is assigned based on the majority class of the instances within that leaf (Breiman et al., 1984).

Proposed Methodology

- *Data Collection:* The thorough data gathering is the initial phase in this study process. A dataset will be created that contains numerous applicant-related factors, such as loan terms, credit ratings, educational background, and asset holdings. This dataset will be derived from historical records kept by lending institutions, guaranteeing that it includes a representative sample of loan applications and results.
- *Data Pre-Processing:* Pre-processing will be conducted to tidy up and get the acquired data ready for analysis. This process comprises dealing with missing values, outliers, and, if necessary, data encoding. Techniques such as standardization and normalization will be employed to ensure the dataset's consistency.
- *Exploratory Data Analysis (EDA):* To learn more about the dataset, exploratory data analysis will be done. To comprehend the distribution of the variables, spot patterns, and evaluate the links between the variables, descriptive statistics, visualizations, and correlation analysis will be employed. EDA will be extremely important in locating possible influencing elements.
- *Feature Selection:* In order to pinpoint the most pertinent elements that substantially affect loan approval decisions, feature selection approaches will be used. To choose a subset of features for model development, methods such feature importance from machine learning models or statistical tests will be applied.
- *Model Development:* Predictive models will be created using cutting-edge machine learning methods, notably the Decision Tree Classifier and Random Forest Classifier. The pre-processed dataset will be used to train these models, and hyperparameter tuning will be done to enhance their performance.
- *Model Evaluation:* The accuracy, precision, recall, F1-score, and support of the assessment metrics used to evaluate the performance of the prediction models will be used. To make certain that the models generalize effectively to new data, cross-validation techniques will be used.
- *Interpretation of Findings:* Key findings, includings the impact of variables on loan approval decisions, will be interpreted. Insights gained from the models will be used to understand which factors influence

approval probabilities and their interactions.

- *Ethical Considerations:* Ethical considerations will be addressed, including data privacy, fairness, and bias. Steps will be taken to ensure that the research and its applications do not perpetuate discrimination or unfair practices in lending.
- *Practical Application:* The research findings and predictive models will be applied practically to lending institutions. Guidelines for using the models to assess loan approval probabilities will be developed, and recommendations for prioritising services to applicants with higher approval probabilities will be provided.

Step 1: Obtain the dataset and load the data using 'Python' tool.

Step 2: Explore, clean and pre-process the data.

Attributes Name and their Description:

- loan_id: A unique identifier for each loan application.
- no_of_dependents: The number of dependents of the applicant.
- education: The education level of the applicant.
- self_employed: A binary indicator of whether the applicant is self-employed (yes/no).
- income_annum: The annual income of the applicant.
- loan_amount: The loan amount requested by the applicant.
- loan_tenure: The tenure of the loan requested by the applicant, measured in years.
- cibil_score: The CIBIL score of the applicant, a credit score used for assessing creditworthiness.
- residential_asset_value: The value of the residential asset owned by the applicant.
- commercial_asset_value: The value of the commercial asset owned by the applicant.
- luxury_asset_value: The value of the luxury asset owned by the applicant.
- bank_assets_value: The value of the bank assets owned by the applicant.
- loan_status: The status of the loan application, which can be either "Approved" or "Rejected."

Step 3: Data Pre-Processing

```
In [3]: # Checking the shape of the dataset
        df.shape
Out[3]: (4269, 13)
```

The dataset consists of 4,269 rows and 13 columns, with a total of 55,497 records. To streamline the data for analysis, the unnecessary "loan_id" column, which serves as an identifier, will be removed.

```
df.drop(columns='loan_id', inplace=True)

# Checking for null/missing values
df.isnull().sum()

no_of_dependents            0
education                   0
self_employed               0
income_annum                0
loan_amount                 0
loan_term                   0
cibil_score                 0
residential_assets_value    0
commercial_assets_value     0
luxury_assets_value         0
bank_asset_value            0
loan_status                 0
dtype: int64
```

Checking the data types of the columns

```
# Checking the data types of the columns
df.dtypes

no_of_dependents             int64
education                   object
self_employed               object
income_annum                 int64
loan_amount                  int64
loan_term                    int64
cibil_score                  int64
residential_assets_value     int64
commercial_assets_value      int64
luxury_assets_value          int64
bank_asset_value             int64
loan_status                 object
dtype: object
```

The dataset includes four types of assets: Residential, Commercial, Luxury, and Bank. These assets are being categorized into two groups: Movable assets and Immovable assets. Specifically, Residential and Commercial assets are classified as Immovable assets, while Luxury and Bank assets are categorized as Movable assets.

Step 4: Descriptive Statistics provide a summary of key characteristics for each variable in a dataset. These descriptive statistics offer a snapshot of the central tendencies and variability of the specified variables in the dataset. They provide valuable insights into the characteristics and distributions of the data (Fig. 1-3).

	no_of_dependents	income_annum	loan_amount	loan_term	cibil_score	Movable_assets	Immovable_assets
count	4269.000000	4.269000e+03	4.269000e+03	4269.000000	4269.000000	4.269000e+03	4.269000e+03
mean	2.498712	5.059124e+06	1.513345e+07	10.900445	599.936051	2.010300e+07	1.244577e+07
std	1.695910	2.806840e+06	9.043363e+06	5.709187	172.430401	1.183658e+07	9.232541e+06
min	0.000000	2.000000e+05	3.000000e+05	2.000000	300.000000	3.000000e+05	-1.000000e+05
25%	1.000000	2.700000e+06	7.700000e+06	6.000000	453.000000	1.000000e+07	4.900000e+06
50%	3.000000	5.100000e+06	1.450000e+07	10.000000	600.000000	1.960000e+07	1.060000e+07
75%	4.000000	7.500000e+06	2.150000e+07	16.000000	748.000000	2.910000e+07	1.820000e+07
max	5.000000	9.900000e+06	3.950000e+07	20.000000	900.000000	5.380000e+07	4.660000e+07

Fig. 1: Descriptive Statistics

Step 5: Exploratory Data Analysis: During the exploratory data analysis phase, we'll concentrate on several crucial elements. In order to understand how the data is distributed and its central trends, we will first look at how the data is distributed across the variables. The correlations between the independent variables and the target variable will next be examined in order to better understand how the predictors affect the desired outcome. Additionally, We'll examine the correlations between the variables to look for any patterns or dependencies. I will depict the dataset visually using data visualization tools so that I may look for potential trends and patterns in the data. These visualizations serve as effective tools for uncovering obscure information and developing a better comprehension of the traits and behaviour of the dataset.

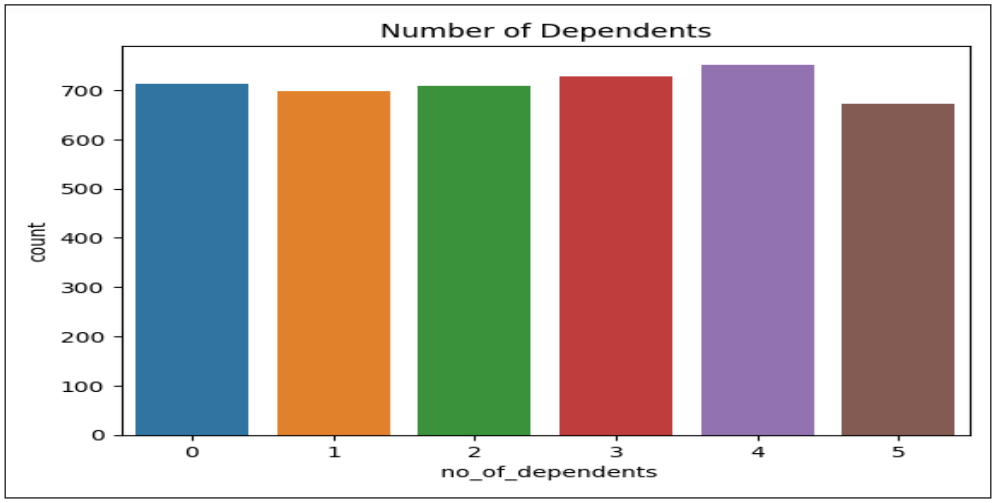


Fig. 2: Number of Dependents

The distribution of dependents connected to loan applications is shown in the graph. A noteworthy finding from the data is that dependents are distributed in a manner that is generally consistent, with applicants who have 4 and 3 dependents having a little larger concentration than those in other categories. It is crucial to keep in mind that the applicant's discretionary income tends to decline as the number of dependents grows. Consequently, it is reasonable to hypothesize that applicants with 0 or 1 dependent may have a higher likelihood of loan approval, given their potentially higher disposable income.

Box-Plot and Violin Plot between Education and Income:

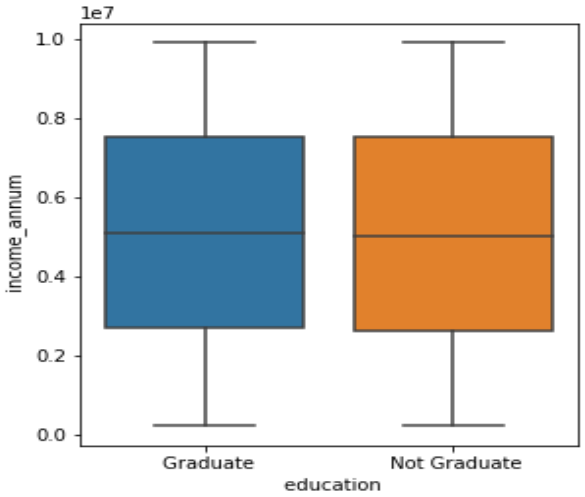


Fig. 3a: Box-Plot

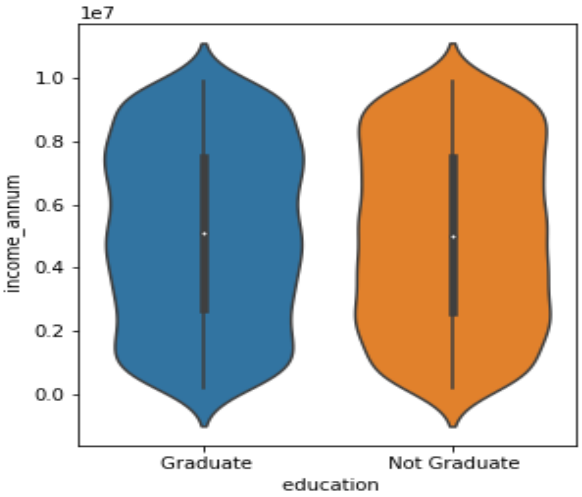


Fig. 3b: Violin-Plot

The pair of visualizations, comprising both a boxplot and a violin plot, offers an insightful portrayal of the relationship between the education levels of loan applicants and their annual income. Upon examining the boxplot, a notable observation emerges: both graduates and non-graduates exhibit nearly identical median incomes, with only a marginal increase in income among the graduates. Furthermore, the violin plot illustrates the income distribution among these two groups. It becomes apparent that non-graduate applicants display a relatively even distribution of income, spanning the range from 2,000,000 to 8,000,000. In contrast, among the graduate applicants, the distribution is more skewed, with a higher concentration falling within the income range of 6,000,000 to 8,000,000. Given the limited disparity in annual income between graduates and non-graduates, it is reasonable to infer that education level may not exert a significant influence on loan approval decisions.

Line Graph for Loan Amount and Tenure: The presented line plot effectively illustrates the relationship between loan amount and loan tenure. Notably, it reveals that for loan tenures ranging from 2.5 to 7.5 years, the loan amount tends to fall within the range of 1,400,000 to 15,500,000. However, a notable departure from this trend is observed for loan tenures of 10 years, where the loan amount experiences a significant increase, indicating a distinct lending pattern for this specific tenure duration (Fig. 4-6).

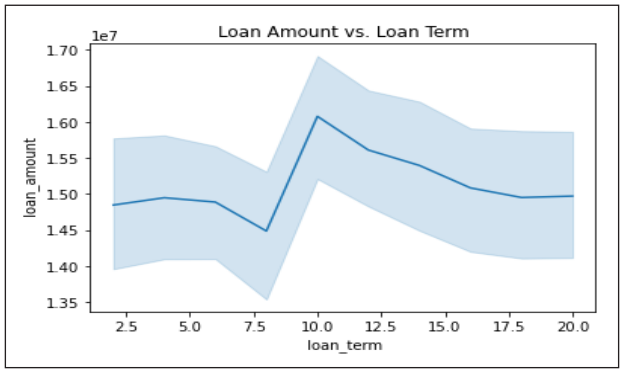


Fig. 4: Line Graph between Loan Amount and Tenure

CIBIL Score Distribution

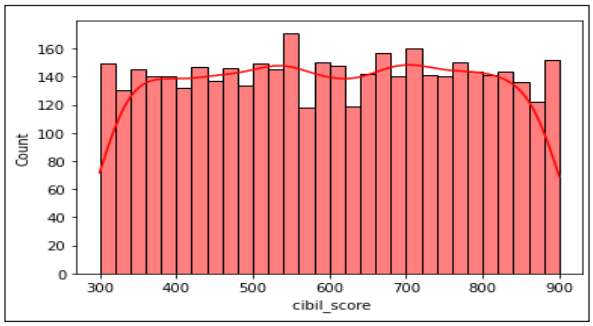


Fig. 5: CIBIL Score Distribution

Before delving into an analysis of individual CIBIL scores, it is essential to understand the CIBIL score ranges

and their corresponding meanings. The CIBIL score, a numerical indicator, is categorized as follows:

- 300-549: This range is associated with a "Poor" credit score, indicating a higher risk profile for the individual.
- 550-649: Falling within this range suggests a "Fair" credit score, reflecting a somewhat improved but still relatively risky credit history.
- 650-749: A credit score within this range is considered "Good," indicating a more stable and reliable credit profile.
- 750-799: CIBIL scores in this range are labelled as "Very Good," signifying a strong credit history and lower credit risk.
- 800-900: The highest range, from 800 to 900, is categorized as "Excellent," indicating an exceptional credit history and the lowest credit risk.

Source: https://www.godigit.com/finance/credit-score/ranges-of-credit-score

These score ranges provide a standardized framework for evaluating an individual's creditworthiness, aiding lenders in making informed decisions regarding loan approvals and interest rates. Referring to the table outlining CIBIL score categories, it becomes evident that a significant portion of customers possesses CIBIL scores below 649, a factor that can potentially hinder the approval of their loan applications. Conversely, there is a notable presence of applicants with CIBIL scores exceeding 649, a positive indicator for the bank. This specific segment of customers presents an opportunity for the bank to target and offer priority services. Additionally, the bank can consider extending special offers and discounts to incentivize these individuals to choose the bank for their loan needs. Based on this observation, a hypothesis can be formulated: Customers with CIBIL scores above 649 are more likely to have their loan applications approved. This hypothesis reflects the potential correlation between a higher credit score and a greater likelihood of loan approval, serving as a valuable insight for the bank's lending strategies (Fig. 7-9).

Line Graph between Loan Amount & Loan Term Vs Loan Status

The depicted graph effectively illustrates the interplay between loan amount, loan tenure, and loan status. A notable trend emerges as approved loans typically exhibit higher loan amounts paired with shorter repayment tenures. Conversely, rejected loans often feature lower loan amounts coupled with longer repayment tenures. This pattern could potentially stem from the bank's lending policies, which might incline towards rejecting loans with extended repayment durations. Additionally, loans with lower amounts may be subject to rejection, possibly because they may not align with the bank's profitability criteria.

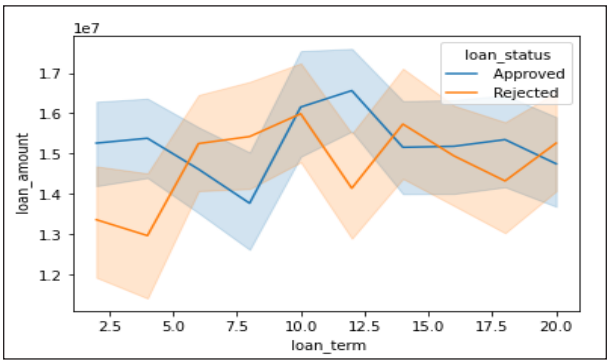


Fig. 6: Line Graph between Loan Amount & Loan Term Vs Loan Status

Correlations Matrix Heatmap

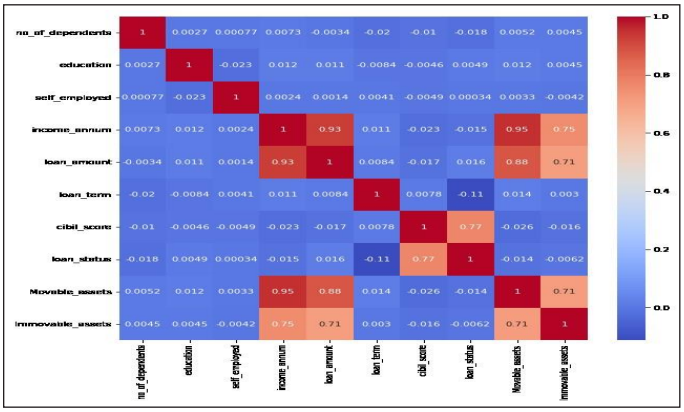


Fig. 7: Correlations Matrix Heatmap

The correlation matrix heatmap reveals several notable strong correlations within the dataset:

- *Movable Assets and Immovable Assets:* These two variables exhibit a strong positive correlation, which is logical since both belong to the category of assets. It's expected that individuals with higher movable assets may also possess higher immovable assets and vice versa.

- *Income and Movable Assets:* There is a strong positive correlation between income and movable assets, which is reasonable because individuals with greater income tend to accumulate more assets.
- *Income and Immovable Assets:* Similarly, income and immovable assets show a strong positive correlation, indicating that higher income is associated with higher immovable assets.
- *Movable Assets and Loan Amount:* Movable assets and loan amount are positively correlated, suggesting that applicants with more movable assets may request higher loan amounts.
- *Immovable Assets and Loan Amount:* Similarly, immovable assets and loan amount also exhibit a positive correlation, indicating that individuals with substantial immovable assets may seek larger loans.
- *Loan Status and CIBIL Score:* The correlation between loan status and CIBIL score highlights that a higher CIBIL score is positively associated with loan approval, which aligns with expectations.
- *Loan Amount and Income:* There is a correlation between loan amount and income, suggesting that applicants with higher incomes may request larger loan amounts.

The strong correlation between assets and income is logical, as individuals with higher incomes often accumulate more assets. Further exploration will focus on the relationships between assets and loan amounts, as well as income and loan amounts. The analysis of the correlation between loan status and CIBIL score has already been addressed in the previous section.

Step 6 and 7: Feature Selection and Model Building

For model building, I will employ two machine learning models to predict loan approval status:

- Decision Tree Classifier
- Random Forest Classifier

These models will be instrumental in evaluating and forecasting the outcomes of loan approval based on the dataset's features and patterns.

To facilitate model training and evaluation, we performed a train-test split using the train_test_split function. The dataset was divided into two subsets:

- X_train: The training set, containing the predictor variables (excluding the 'loan_status' column).
- X_test: The testing set, also comprising predictor variables.

Additionally, two corresponding sets were created to store the target variable, 'loan_status':

- y_train: The target values for the training set.
- y_test: The target values for the testing set.

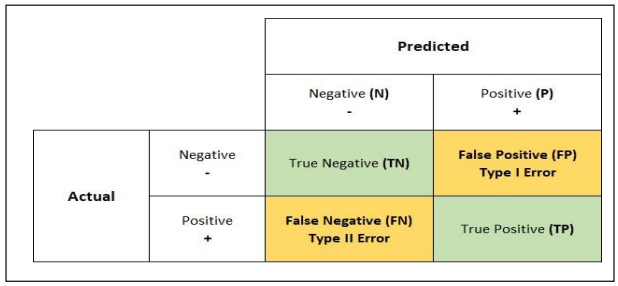
The split allocated 20% of the data to the testing set, ensuring that model evaluation can be conducted independently on this portion. A random seed value of 42 was used for reproducibility.

Step 8: Model Evaluation

Confusion Matrix: The confusion matrix is a table is commonly used to evaluate the performance of a classification model. It provides a breakdown of the predicted and actual class labels for a classification problem. In a binary classification scenario (two classes: positive and negative), the confusion matrix typically consists of four values:

- *True Positive (TP):* This refers to the count of instances that the model correctly predicted as positive. In other words, it represents the number of cases where the model made a positive prediction, and the actual outcome was indeed positive.
- *True Negative (TN):* This signifies the count of instances that the model correctly predicted as negative. It corresponds to the number of cases where the model accurately predicted a negative outcome, and the actual result was indeed negative.
- *False Positive (FP):* This represents the number of instances that the model incorrectly predicted as positive when, in reality, they were negative. It's also commonly referred to as a Type I error, indicating the model's tendency to generate false positive predictions.
- *False Negative (FN):* This denotes the count of instances that the model inaccurately predicted as negative when, in fact, they were positive. It's often referred to as a Type II error, indicating the model's failure to identify positive cases correctly.

Mathematically, the confusion matrix can be represented as follows:



Source: https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5

Fig. 8: Confusion Matrix

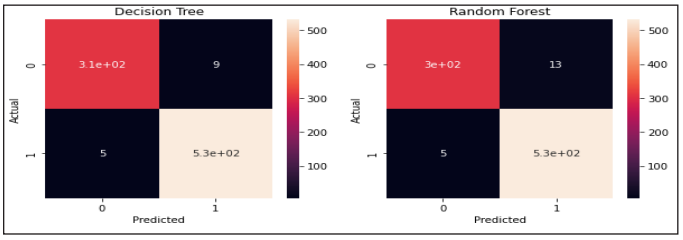


Fig. 9: Confusion Matrix for Decision Tree and Random Forest Classifier

The presented confusion matrix heatmap provides a visual representation of the true positive (TP) and true negative (TN) value counts for both machine learning models. In the case of the Decision Tree Classifier, there are only 14 instances of false positives and false negatives combined. Conversely, the Random Forest Classifier exhibits 18 false positive and false negative values. This observation indicates that the Decision Tree Classifier demonstrates a superior accuracy rate compared to the Random Forest Classifier, as it has a lower combined count of misclassifications.

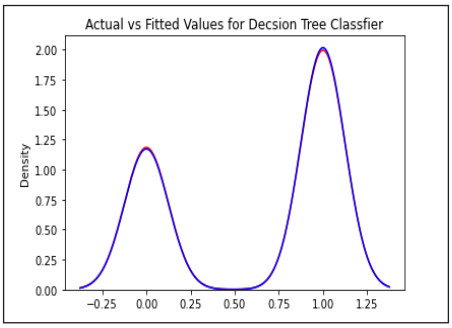


Fig. 10a: Actual vs Fitted Values for Decision Tree Classifier

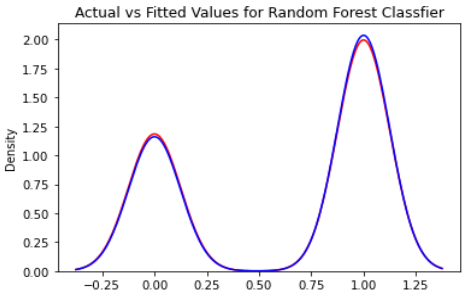


Fig. 10b: Actual vs Fitted Values for Random Forest Classifier

Distribution Plot

The distribution plots of both machine learning models exhibit a striking similarity, with nearly identical distribution densities of predicted and actual values in the Random Forest Classifier. However, a distinctive observation emerges in the case of the Decision Tree Classifier, where the distribution density of predicted values visibly overlaps with the distribution of actual values (Fig. 10). Based on this analysis, it can be reasonably concluded that the Decision Tree Classifier outperforms the Random Forest Classifier for this dataset, as it closely aligns with the actual distribution of values, indicating a better fit and predictive accuracy.

Classification Report

Table 1: Classification Report

	Precision	*Recall*	*f1-Score*	*Support*
0	0.98	0.97	0.98	318
1	0.98	0.99	0.99	536
accuracy			0.98	854
macro average	0.98	0.98	0.98	854
weighted average	0.98	0.98	0.98	854
	Precision	*Recall*	*f1-Score*	*Support*
0	0.98	0.96	0.97	318
1	0.98	0.99	0.98	536
accuracy			0.98	854
macro average	0.98	0.97	0.98	854
weighted average	0.98	0.98	0.98	854

Decision Tree Classifier Accuracy

R2 score: 0.9299

The coefficient of determination, often known as the R2 score, quantifies the percentage of the dependent variable's variation that can be predicted from the model's independent variables. A score of 0.9299 means that a significant portion of the variance in the data is explained by the model, which is 92.99%, which is a high percentage. This may indicate that the model accurately represents the data (Table 1).

Mean Squared Error (MSE): 0.0164

The average squared difference between expected and actual values is measured by MSE. An MSE of 0.0164 in this instance indicates that the squared difference between the anticipated and actual values is, on average, relatively tiny.

Mean Absolute Error (MAE): 0.0164

The average absolute difference between expected and actual values is measured by MAE. The model's predictions on average agree with the actual values, as shown by the MAE of 0.0164.

Random Forest Classifier

R2 score**:** 0.9098

The R2 score for this model is 0.9098, indicating that it explains approximately 90.98% of the variance in the data. While slightly lower than the R2 score of Decision Tree, it's still a good fit for the data.

Mean Squared Error (MSE): 0.0211

The MSE for random forest is 0.0211, suggesting that, on average, the squared difference between predicted and actual values is slightly larger than in Decision Tree.

Mean Absolute Error (MAE): 0.0211

In comparison to Decision Tree, the MAE for Random Forest is 0.0211, suggesting that, on average, the model's predictions have a somewhat higher absolute deviation from the actual values.

In conclusion, both models seem to work well, with the Decision Tree having a marginally higher R2 score and marginally lower MSE and MAE, suggesting it could be a slightly better fit for the data. However, the precise objectives of the research and any trade-offs between model complexity and interpretability should be considered while deciding between the two models.

Results and Discussion

In conclusion, the exploratory data analysis has shed light on key factors influencing loan approval decisions. Based on our findings:

CIBIL Score: Individuals with higher CIBIL scores are more likely to have their loan applications approved.

Number of Dependents: Those with a greater number of dependents face reduced chances of loan approval.

Assets: Applicants with a higher quantity of assets, including both movable and immovable assets, have an increased likelihood of loan approval.

Loan Amount and Tenure: Those requesting higher loan amounts with shorter tenures tend to have greater odds of loan approval.

Regarding the machine learning models employed, we utilised both the Decision Tree Classifier and Random Forest Classifier. Both models exhibited excellent performance, achieving accuracies of 92.99% and 90.98%, respectively. However, it is noteworthy that the Decision Tree Classifier has yielded better results than the Random Forest Classifier in this context. These insights and model results provide valuable guidance to lending institutions, enabling them to enhance their loan approval processes and prioritise services for applicants with a higher likelihood of approval.

References

Agarwal, S., & Tufano, P. (2018). Credit market consequences of credit scores. *Journal of Financial Economics, 130*(3), 537-558.

Atisattapong, W., Samaimai, C., Kaewdulduk, S., & Duangdum, R. (2018). Data mining for automated assessment of home loan approval. In *Communications in Computer and Information*

Science (pp. 11-21). Cham: Springer International Publishing.

Barr, M. S. (2020). Banking the poor: Overdraft versus asset-based credit. *Journal of Consumer Affairs, 54*(1), 118-148.

Bansal, J., & Anil, K. (2018). Qualitative research on issues and trends in the bancassurance model in India: An interview report. *International Journal of Banking, Risk and Insurance, 6*(2), 67-78. Retrieved from http://publishingindia.com/ijbri/

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. CRC Press.

Blessie, E. C., & Rekha, R. (2019). Exploring the machine learning algorithm for prediction the loan sanctioning process. *International Journal of Innovative Technology and Exploring Engineering, 9*(1), 2714-2719. doi:https://doi.org/10.35940/ijitee.a4881.119119

Jangra, K. (2020). Operating efficiency of Indian public sector banks in light of Basel III norms. *International Journal of Banking, Risk and Insurance, 8*(1), 15-25. http://publishingindia.com/ijbri/

Jiang, B., & Wei, J. (2018). Personal loan risk prediction. *In 2018 IEEE International Conference on Data Mining* (ICDM) (pp. 867-872). IEEE.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction.* Springer.

Huang, Y., Lee, C., & Huang, W. (2018). Loan approval prediction using random forest. In *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication* (p. 26). ACM.

Mangala, D., & Kumari, P. (2015). Corporate fraud prevention and detection: Revisiting the literature. *Journal of Commerce and Accounting Research*, *4*(1), 35-45, Available at SSRN: https://ssrn.com/abstract=2678909

Mester, L. J., & Nakamura, L. I. (2018). Supervisory actions on banks and the severity of the financial crisis: Do capital requirements matter? *Journal of Financial Intermediation, 35,* 45-59.

Mishra, A., Nigam, A., & Chawla, K. (2017). Predictive modeling in banking: Predicting loan approval for home loans using various machine learning techniques. *Procedia Computer Science, 132,* 1075-1082.

Mridha, K., Barua, D., Shorna, M. M., Nouman, H. N., Kabir, M. H., & Singh, A. V. (2022). *Credit approval decision using machine learning algorithms.* 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). IEEE.

Ndanusa, I. H., Adepoju, S. A., & Ojerinde, A. O. (2022). Consensus based bank loan prediction model using aggregated decision making and cross fold validation techniques. In *2022 5th Information Technology for Education and Development* (ITED). IEEE.

Ramesh, K. (2019). Determinants of bank performance: Evidence from the Indian commercial banks. *Journal of Commerce and Accounting Research, 8*(2), 66-71. http://publishingindia.com/jcar/

Thomas, K., & Indirubin, J. (2019). *Predicting loan defaulters using machine learning techniques.* In 2019 International Conference on Advances in Computing and Communication Engineering (ICACCE) (pp. 1-6). IEEE.

Tumuluru, P., Burra, L. R., Loukya, M., Bhavana, S., CSaiBaba, H. M. H., & Sunanda, N. (2022). *Comparative analysis of customer loan approval prediction using machine learning algorithms.* In 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS). IEEE.