Predicting the chance of getting a loan?

Joseph Ejoh, *CE880, 2201124*

Abstract—The financial ecosystem relies on lenders being adept at predicting loans for creditworthy applicants. In this study the aim is to use machine learning models to predict loan approval based on applicants data. The dataset is pre-processed by dealing with missing values, encoding categorical features and scaling numerical features. The study uses classification models including Support Vector Machine (SVM), Random Forest, K-Nearest Neighbour (KNN, gradient Boost and Naïve Bayes. Each model was evaluated and in the results, Random forest was shown to be the best performing model. The feature importance analysis highlights that 'Credit history was the most important feature followed by 'Loan amount' and 'Applicant income'. The study contributes to the use of machine learning algorithms to effectively predict loan approvals..

Index Terms—Machine Learning, Classification, Report .

1 Introduction

LOANS are an important parameter in the world's financial landscape as it assist many individual and business with purchasing homes, further education and starting business ventures. For this reason loan approvals are an important function for a financial institutions to optimize, the process of approving loans for these lender shave inherent risks attached, so it's in the best interest for the economy that the task of correctly classifying loan approvals to individual and businesses is done optimally. This project revolves around the prediction of loan status, in particular whether an applicant will be approved or rejected for a loan based on various features. The objective is to develop a machine learning model that can accurately classify loan applicants as either eligible (approved or ineligible (rejected).

August 20, 2023

1.1 Significance of loan prediction for financial institutions and loan applicants

The lending ecosystem relies on the proportion of correct loan assessments to be favourable, and in doing this can provide benefits to both the financial lenders and the loan applicants. Using machine learning models can give financial lenders a data-driven approach to decision making and mitigate risk and for loan applicants can results in accurate loan predictions which will provide fair treatment, access to credit and for financial stability and growth. This results in the financial ecosystem being healthy for all parties involved

1.2 Objectives

The primary goals is develop and asses the performances in binary classification models to classify loan approval . The assessments of these models will be a comparison of performance on specific metrics then identify the most accurate and reliable model for predicting loan approvals. By considering these machine learning models and evaluation metrics, the aim to identify the best-performing model for loan status prediction, providing valuable insights for financial institutions and loan applicants. Overall, the goal is to contribute to the field of loan prediction by employing advanced machine learning techniques and providing valuable insights for financial institutions and loan applicants.

- *M. Shell was with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332. E-mail: see http://www.michaelshell.org/contact.html*
- *J. Doe and J. Doe are with Anonymous University.*

Manuscript received April 19, 2005; revised August 26, 2015.

2 Background Studies

Previous studies have explores the uses of machine learning models to predict bank loans and similar criteria. These studies have key insights into the makeup of machine learning in predicting loan approval and garner discussions on how they can be optimized. Gupta et al. (2020) [1] explored the utilization of machine learning in predicting loan approval within the context of banking systems. They highlighted the importance of unsupervised learning when dealing with unlabelled data, underscoring the need for algorithms that can effectively handle such scenarios. Notably, the study emphasized the versatility of logistic regression, a technique widely used across various research fields. The authors also underscored the role of predictive analytics, combining methodologies like data mining, statistics, modeling, machine learning, and artificial intelligence to forecast outcomes based on current data. Furthermore, the study recognized the effectiveness of random forest, a robust algorithm, in the context of loan approval prediction. Singh et al. (2021) [2] Singh and co-authors delved into the modernized loan approval system, focusing on the significance of loan repayment for financial institutions and the accessibility of loans for education and business purposes within the middle-class segment. The study identified crucial factors such as credit score, loan duration, loan amount, age, and income as determinants of loan eligibility. The authors stressed the role of machine learning algorithms in predicting loan outcomes and highlighted the potential benefits for both banks and clients. Additionally, the study advocated for stringent verification and background checks to ensure loan repayment and minimize risks for banks. Sheikh et al. (2020) [3] Sheikh, Goel, and Kumar proposed

an approach for predicting loan approval using a machine learning algorithm. They advocated for the use of logistic regression in identifying suitable customers for loan approval based on their likelihood of defaulting. The authors worked with a dataset comprising over 600 rows and 13 columns for training data and 300 rows and 12 columns for test data. The study involved feature engineering and exploratory data analysis to comprehend the relationships between dependent and independent variables. The utilization of logistic regression and the sigmoid function facilitated the classification and prediction of loan defaulters. The proposed model aimed to streamline the loan approval process, leveraging machine learning algorithms to enhance customer satisfaction and operational efficiency.

3 METHODOLOGY

3.1 Data-set

The data-set used consisted of various information on loan applicants, these provided insights into an applicant's financial status and other characteristics, which is often valuable for a financial institution in assessing loan application. The data consisted of 'Gender' of the applicant (e.g.. Male or Female), whether they were 'Married' or not (e.g., Yes or no), the number of 'Dependents' the applicant has. Their 'Education' status (e.g., Graduate or Not Graduate). If the applicant is 'Self-Employed' (e.g., Yes or No). The 'Applicants Income'. The 'Co-applicant Income' if there is any. The 'Loan Amount', The loan 'Term'. Their 'Credit History' (e.g., 1 for good and 0 for bad). The 'Area' the applicant lives. The 'Status' which is the target variable and depicts the loan approval (e.g., Y is for approved and N is for not approved).

3.2 Pre-processing

To ensure the proper handling of the data, various pre-processing steps was used to prepare for the data for the machine learning models.

3.2.1 Data Cleaning

The data had missing values for 'Gender', 'Married', 'Dependents, 'Self Employed' and 'Credit History' which are categorical columns, therefore the missing values were filled with the most frequent value (Mode). The 'Term' value is a numerical column so it's missing values were filled with the average value (Mean). Converting 'Dependents' Column to Strings: The 'Dependents columns were converted to strings to insure a consistent data type before performing one-hot encoding. One-Hot Encoding: One-hot encoding was performed on the categorical columns. It is a technique used to convert categorical variables into a numerical format. By dealing with the missing values and performing one-hot encoding, the data is prepared for further analysis and be used by the machine learning algorithm.

3.2.2 Standardisation

The data was also scaled using the subtraction of the mean and dividing by the standard deviation for each feature, this standardisation was to ensure that any model algorithm that is sensitive to scale of input features will be able to analyses the data effectively and ensures that all data has a mean of 0 and a standard deviation of 1 which can improve the convergence and performance of the model.

3.2.3 Data Split

The data was split into a training set (80%) a validation set (10%) and test set (10%) . The training set which constituted 80% of the of the data which used to train the machine learning model due to ML models having majority of the data to train and allows the model to capture complex patterns and variations in the data. The validation set is used to tune hyper-parameters and evaluate the models performance during the training phases, which is key to prevent over-fitting and also provides an primary assessment of the model. The test set is used as a final evaluation of the models performance after hyper-parameter tuning and validation, it designed to provide an unbiased estimate of the models generalization capabilities on unseen data.

3.3 Classification models

3.3.1 Support Vector Machines (SVM)

This model is a supervised machine learning algorithm typically used for classification tasks. Its works by looking for optimal hyperplane that best separate different classes of data. SVM is capable of handling both linearly separable and non-linearly separable data via different kernels (e.g., linear, polynomial, radial basis function). It's very useful in high dimensional space and is popular due to being less prone to over-fitting.

3.3.2 Random Forest Classifiers

This model is an ensemble learning method based on decision tree classifiers. It works by creating multiple decision tress during the training phase and combines their prediction though majority voting during a classification task. It's robust to over-fitting and performs well on bigger data-sets and is able to handle categorical and numerical data sets.

3.3.3 K- Nearest Neighbour

KNN works as a classification model by classifying data points on the class labels of their k-nearest neighbour in the training phase. KNN is a non-parametric learning algorithm meaning that it does not explicitly learn during the models training phase. However it works well enough to adapt to any number of classes and complex decision boundaries.

3.3.4 Gradient Boost

Gradient boost is an ensemble learning method that build multiple learning sequentially. Each weak learner is trained to correct the error made by the previous ones leading to improved predictive performance sequentially. It is effective in handling categorical and numerical data and is less prone to over-fitting.

3.3.5 Naive Bayes

Naïve Bayes is a probabilistic classification model based on Bayes Theorem. It's assumption are that features are conditionally independent of on each other given the class labels, which can simplify the calculations of class probabilities. It is computationally effective classification model and performs well in high dimensional feature space

3.4 Hyper-parameter Tuning

For each model, the parameter grids were defined for grid search to tune the hyper-parameters of each classification model by looking at varying combinations of hyper-parameters, which helps to systematically search through various hyper-parameters for each model, which will then identify the best combination of hyper-parameters for each classification model

3.4.1 Cross-validation

Cross-validation was to assess the performance of a model and to find the best hyper-parameters. It involves partitioning the data-set into multiple subsets and then training and evaluating each model multiple times. This ensures that there is a better estimate of the models performance and how well it can generalize to unseen data, and prevents over-fitting to the training data.

3.5 Model Evaluation

We assessed the best tuned models using various evaluation metrics on the validation set to better understand each models performance. 'Accuracy' which is the proportion of correct predictions out of all predictions made by the model. It designed to give an overall view of the models performance. 'Precision' is the proportion of true positive predictions out of all the positive prediction made. It's designed to measure the models ability to correctly identify positive samples. 'Recall' is the proportion of true positive predictions out of all actual positive samples in the data-set. It's designed to measure the model ability to find all positive samples. 'F1 Score' is the harmonic mean of precision and recall and provides a balance between precision and recall, making it useful if the event of uneven class distributions. 'ROC AUC Score' (Receiver Operating Characteristic Area Under the Curve) measures the models ability to discriminate between positive and negative classes and plots the true positive rate against the false positive rate at varying threshold. The higher the score the higher the models ability to discriminate. These evaluation metrics and using cross-validation will help assess how well the models are performing and in turn select the best performing model for the task.

3.6 Best Model Selection

The selection of the best model was based on the performance metrics calculated on the validation set. The model that demonstrated the highest overall performance across these metrics was considered the best model. The selection of the best model was a trade-off between quantitative performance metrics and the model's ability to generalize well to unseen data.

4 Results

4.1 Model Performance On Validation Set

4.1.1 Support Vector Machine

ROC-AUC Score: 0.784

The SVM model achieved an accuracy of 0.82 on the validation set. It shows a good recall for class 'Y' (approved), indicating that it can identify most of the approved cases. However, the recall for class 'N' (not approved) is lower, suggesting that the model struggles to identify unapproved cases. The F1-score and ROC AUC score also highlight this trade-off between precision and recall

TABLE 1
Classification Report for SVM Model

Class	Precision	Recall	F1-Score	Support
N	0.90	0.47	0.62	19
Y	0.80	0.98	0.88	42
Accuracy	0.82			
Macro Avg	Precision 0.85	Recall 0.72	F1-Score 0.75	61
Weighted Avg	Precision 0.83	Recall 0.82	F1-Score 0.80	61

4.1.2 Random Forest Model

ROC-AUC Score: 0.792

TABLE 2
Classification Report for Random Forest Model

Class	Precision	Recall	F1-Score	Support
N	0.82	0.47	0.60	19
Y	0.80	0.95	0.87	42
Accuracy	0.80			
Macro Avg	Precision 0.81	Recall 0.71	F1-Score 0.73	61
Weighted Avg	Precision 0.81	Recall 0.80	F1-Score 0.79	61

The Random Forest model shows balanced performance with an accuracy of 0.80. It achieves a higher recall for class 'Y' compared to the recall for class 'N'. This indicates that the model is better at identifying approved cases. The F1-score is higher for class 'Y', showing the model's ability to capture true positives.

4.1.3 KNN Model

ROC AUC Score: 0.678

TABLE 3
Classification Report for KNN Model

Class	Precision	Recall	F1-Score	Support
N	0.58	0.37	0.45	19
Y	0.76	0.88	0.81	42
Accuracy	0.72			
Macro Avg	Precision 0.67	Recall 0.62	F1-Score 0.63	61
Weighted Avg	Precision 0.70	Recall 0.72	F1-Score 0.70	61

The KNN model achieved an accuracy of 0.72. It has a relatively lower precision for class 'N', meaning that it tends to make more false positive predictions for unapproved cases. However, it has a high recall for class 'Y', indicating its ability to identify most of the approved cases.

4.1.4 Gradient Boost Model

ROC AUC Score: 0.754

The Gradient Boosting model achieved an accuracy of 0.80. It has a balanced performance with good precision and

TABLE 4
Classification Report for Gradient Boosting Model

Class	Precision	Recall	F1-Score	Support
N	0.82	0.47	0.60	19
Y	0.80	0.95	0.87	42
Accuracy	0.80			
Macro Avg	Precision 0.81	Recall 0.71	F1-Score 0.73	61
Weighted Avg	Precision 0.81	Recall 0.80	F1-Score 0.79	61

recall for both classes. The F1-score and ROC-AUC score indicate a trade-off between precision and recall, but overall, the model performs reasonably well on both fronts.

4.1.5 *Naive Bayes Model*

ROC AUC Score: 0.756

TABLE 5
Classification Report for Naive Bayes Model

Class	Precision	Recall	F1-Score	Support
N	0.82	0.47	0.60	19
Y	0.80	0.95	0.87	42
Accuracy	0.80			
Macro Avg	Precision 0.81	Recall 0.71	F1-Score 0.73	61
Weighted Avg	Precision 0.81	Recall 0.80	F1-Score 0.79	61

The Naive Bayes model achieved an accuracy of 0.80. It performs similarly to the other models with good recall for class 'Y'. However, its recall for class 'N' is relatively lower, indicating potential difficulty in identifying unapproved cases.

4.1.6 *Model Comparison*

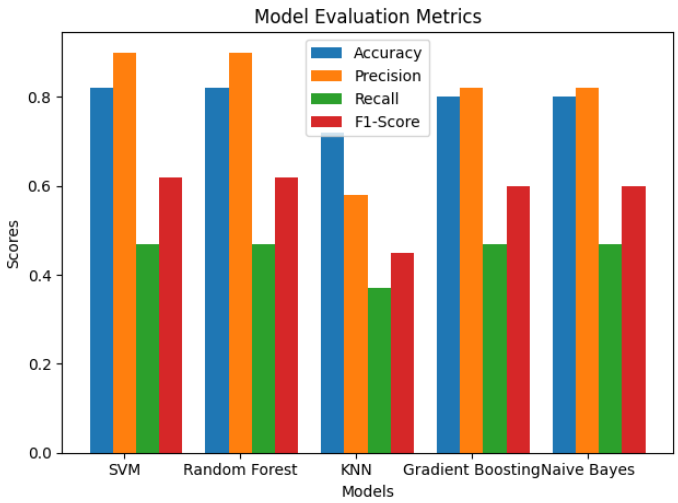

Fig. 1. Model Comparison: Evaluation Metrics

The SVM and Gradient Boosting models show higher precision and recall for class 'Y', while the KNN model excels in identifying approved cases. The Random Forest and Naive Bayes models strike a balance between the two

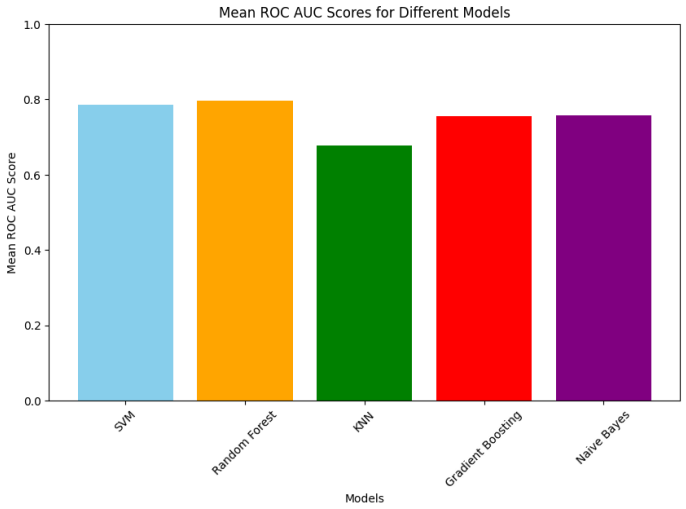

Fig. 2. Model Comparison: ROC AUC Scores

classes. The ROC-AUC scores provide insights into the overall discriminatory power of each model.

Additionally, Figure 1 presents a visual comparison of the models' evaluation metrics, including accuracy, precision, recall, and F1-score. This bar plot reinforces our observations, illustrating the differences in performance metrics among the models.

Figure 2 provides a graphical representation of the ROC AUC scores for each model. This plot offers a clear perspective on the models' ability to distinguish between the positive and negative classes, further supporting our analysis of discriminatory power.

In both the evaluation metrics comparison and ROC AUC plots, we can visually confirm the patterns we observed earlier in terms of model performance, enhancing our understanding of their strengths and weaknesses.

4.2 Best Model Evaluation on Test Set

The classification report for the best-performing Random Forest model on the test set is shown in Table 6.

TABLE 6
Classification Report for Best Random Forest Model (Test Set)

Class	Precision	Recall	F1-Score	Support
N	0.91	0.42	0.57	24
Y	0.73	0.97	0.83	38
Accuracy	0.76			
Macro Avg	Precision 0.82	Recall 0.70	F1-Score 0.70	62
Weighted Avg	Precision 0.80	Recall 0.76	F1-Score 0.73	62

The Random Forest model has good precision and recall for predicting the 'Y' class (loan approval), suggesting that it's able to accurately identify applicants who are likely to get approved for loans. The precision for class 'Y' is higher than class 'N', indicating that when the model predicts an applicant will get approved, it's more likely to be accurate. The recall for class 'Y' is high as well, meaning that the model is able to capture a significant proportion of actual

loan approvals. The F1-score balances both precision and recall, providing a holistic view of the model's effectiveness.

The model's generalization performance on the test set is consistent with its performance on the validation set. It is able to maintain its ability to identify approved cases well, but not so much for unapproved cases. This could mean that further analysis is needed to understand the important features in the model.

4.3 Feature Importance Analysis

The feature analysis shown in Figure 3 highlights that 'Credit history' serves as the most important feature in the model of loan approval. 'Loan amount' is the second most important feature, showing that it is essential because it directly affects the risk associated with the loan. Higher loan amounts could pose higher repayment burdens and potential risks, leading to careful evaluation.

The third most important feature is 'Applicant income', which complements credit history and loan amount in assessing an applicant's financial stability. A higher income not only suggests the ability to repay the loan but also reflects the overall financial health of the applicant.

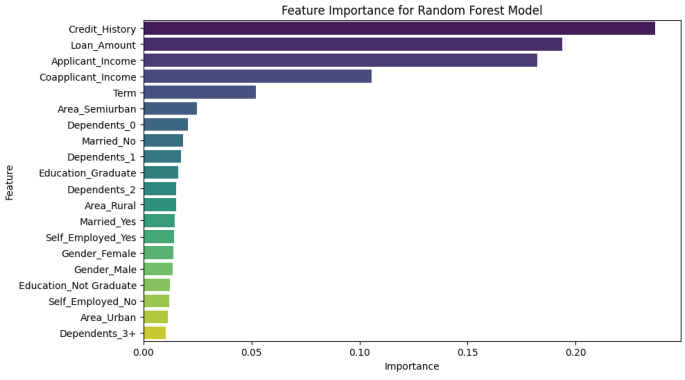

Fig. 3. Feature Importance Analysis for Random Forest Model

5 Discussion

The results and analysis showed important information regarding the models' performances in predicting loan approvals. The five metrics used (accuracy, precision, recall, F1-score and ROC-AUC) provided a holistic understanding of the models capabilities. These metric gave us an indication on how effective it can identify potential loan defaults and approvals. From this, the evidence in the results section shows that Random Forest model with the validation and test set offers the most comprehensive performance in predicting loan approvals. The trade-off between identifying genuine loan applicants (recall) and minimizing the risk of incorrect approval (precision) is important for an financial institution in offering the deserving applicants a granted loan while minimising the risk, and the Random Forest model seems to balance these trade-offs the most comprehensively. While the SVM model demonstrated a high recall, indicating its ability to capture actual positive cases, it lagged in precision. This means that it might identify a larger number of loan applicants as potentially eligible, leading to more false positives. The Random Forest model's strong performance can be attributed to its ability to handle non-linear relationships and interactions among features. The KNN model showed decent recall but struggled with precision. Its performance might be influenced by its sensitivity to the choice of the number of neighbours. Both Naive Bayes and Gradient Boost models exhibited competitive performance, with the latter excelling in recall and the former performing well in precision, but overall didn't perform as comprehensively as the other models. Models like Random Forest and Gradient boost leverage ensemble techniques that combine multiple models to improve predictive power, while models like Naïve Bayes performs well when specific assumptions align well with the data distribution, which will produce differences in performance across the models. With loan predictions the trade-offs between precision and recall are important things to consider. Higher recall indicates that a model is adept at identifying actual positive cases which can be important to avoid as much missed opportunities as possible. A higher precision indicates that a model is more cautious about false positives which can in turn reduce the risk of granting loans to potentially unfit applicants. Future work could be steered towards the potential of additional features to improve the models performance. A great way to further improve the models is to explore algorithms that excel with imbalances data-sets. Also it's important to highlight the role of domain-specific knowledge and how I can affect the direct of model training.

6 Conclusion

In conclusion the significance of developed models for loan predictions are that the finding underpin the importance of leveraging machine learning algorithms in the lending industry. The classification models offer institutions a quantitative approach to assess the creditworthiness of a loan applicant. So, by accurately identifying potential loan defaults and approving viable candidates, these algorithmic approaches can enhance lending practices while simultaneously reducing risk. The overall success of the study is that it embodies an important step for machine learning for loan prediction. The success evaluation of multiple models and identification of the most effective provide an in depth analysis into the models predictions tasks and success. As this is an important area in banking and lending it's important to highlight that domain specific knowledge and external data sources might be a critical step in furthering a models accuracy. Ultimately, the analysis provides a big enough step into the machine leanings applicability in lending domains.

References

[1] Anshika Gupta, et al. "Bank Loan Prediction System using Machine Learning." In *2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)*, IEEE, 2020.

[2] Vishal Singh, et al. "Prediction of modernized loan approval system based on machine learning approach." In *2021 International Conference on Intelligent Technologies (CONIT)*, IEEE, 2021.

[3] Mohammad Ahmad Sheikh, Amit Kumar Goel, and Tapas Kumar. "An approach for prediction of loan approval using machine learning algorithm." In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, IEEE, 2020.