Research on loan approval and credit risk based on the comparison of Machine learning models

Chunyu Yang [1, *]

[1]University of Melbourne, Melbourne Business School, 200 Leicester Street, Australia

Abstract. Nowadays, home loan is a frequently accessed component of people's financing activities. Homeowners wants to increase the probability of loan acceptance, however banks seek to borrow money to low risk customers. This paper compared and examined the machine learning models to select when loan applicants evaluating their probability of success. This paper introduced the recommended models for the problem, explanations on how to use the selected model. 6 candidate models, including Logistic regression, Decision tree, Random Forest, support vector machine (SVM), Ada Boost and Neural Network are selected. The model selection process would focus on the model's accuracy on test data as well as the interpretability of these models. The models' result was interpreted to derive optimal strategies to be undertaken by both debtors and creditors. Throughout comparison between these models, logistic regression was the best in terms of interpretability and accuracy. Nonetheless, other models could bolster the decision-making process by examining their confusion matrices and the fitted importance of predictors in each model. This paper revealed practical implications of machine learning theories on home loan approval and credit risk and aimed to help decision making for both debtors and creditors.

1 Introduction

Recent years, as the demand for financing activities increases, home loan, becomes a quintessential part of people's equity allocation. Therefore, getting a successful loan approval becomes the key concern of borrowers. Banks, on the other hand, aims to lend money only to trustworthy clients. The game is even more brought to attention due to the prevalence of machine learning techniques. This paper provides guidelines for decision making process of both parties regarding to loan and aims to propose a model selection process on loan prediction and potential credit risk.

Several research papers are used as guidelines to this research. Arun and Tejaswini's proposed 6 potential models, linear models such as linear regression, non-linear models such as random forest and neural networks are referred for this paper [2,3]. Since the response variable is binary, logistic regression which constrain the response variable to be either 0 or 1, would be more appropriate compare to linear regression as indicated by Sheikh and Singh's selection of logistic model for analysis [4,5]. Other highly praised models such as the flexible support vector machine is also selected to fit the dataset[6]. Leaving out the final candidate models of linear models which stress the logical and marginal effect of each individual predictors: logistic regression, support vector machine (SVM); nonlinear models which aims to capture a more complex relationship between these variables: decision tree, Random Forest, Ada Boost and deep learning to provide insights and prediction in a most complex situation between elements: Neural Network [7-12].

There are several studies arguing for different machine learning techniques to approach the loan prediction problem, however on each paper, the candidate models are either limited to a small amount or exemplified with limited elaboration on the model selection process, therefore this paper will focus on explaining features of different models and their use in predicting loan approval. This research will compare between those candidate models and yield a best model in terms of test accuracy rate and interpretability. In previous research, there is also a lack of discussion about the interpretation of model results in a business situation and how machine learning models could be adopted by end users, therefore this paper will discuss the potential strategies for creditors and debtors when they receive a loan approval dataset or they are given the fitted models to maximize their benefit. Therefore, this paper aims to use machine learning techniques to clarify these the ambiguity in the area of loan approval and credit risk, both in theory and application.

2 Data and Method

* Corresponding author: chunyy@student.Unimelb.edu.au

2.1 Data

The study is grounded in a comprehensive dataset composed of mortgage application data from 1988, collected from major banks in Australia and provided through 'Wooldridge' [1]. This collection encompasses both applicant-specific attributes and loan details to uncover potential factors that may significantly influence the loan approval process.

Initially, the dataset comprised 1988 entries and 59 variables. Nevertheless, 25 predictors were deemed irrelevant to the current analysis due to reasons such as minimal positive observations or a lack of logical correlation with the response variable 'APPROVE'. One such variable, 'gift as down payment', was removed due to its insignificant role in loan approval.

To enhance the validity of the model comparison and maintain balance within the dataset, oversampling was employed. Instances with 'APPROVE' equal to 0 were replicated six times using Excel. Simultaneously, entries with missing values were systematically removed to ensure data integrity.

As a result, the refined dataset contained 3413 entries across 24 variables （shown in Table 1）. To facilitate effective training and validation of the machine learning models, the dataset was partitioned into an 80:20 split for training and testing purposes, respectively.

Throughout this research, the random seed was set to 42 to ensure consistency and reproducibility in all computational processes. The aim of this study is to deliver a comprehensive comparison of various machine learning models for predicting home loan approvals, offering valuable insights for both applicants and financial institutions in their decision-making process.

Table 1. Selected variables for model fitting

Variable	Explanation	Variable	Explanation
APPROVE	=1 if the mortgage application is approved	OTHER	amount of other financing in thousands of dollars
APPINC	applicant's annual income in thousands of dollars	ATOTINC	total monthly income
HRAT	monthly housing expenditures to monthly income ratio	DEP	number of dependents
OBRAT	total monthly obligations to monthly income ratio	MARRIED	=1 if applicant is married
EMP	years employed in current line of work	MALE	=1 if male is used as applicant's identification
SELF	=1 if applicant employed herself	PUBHIST	=1 if there is record of default in public
LIQ	amount of liquid assets in thousands of dollars	SCH	(years of education) =1 if the applicant has more than 12 years in school
NETW	net worth in thousands of dollars	LOANAMT	loan amount in thousands of dollars
PRICE	purchase price of the property	LOANPR	loan amount to purchase price ratio
APPR	appraised value of the property	THICK	=1 the application file is identified to be thick if more than two credit reports observed
BLACK	=1 if applicant is black	MULTI	=1 if the property is a multi-bedroom family
ASIAN	=1 if applicant is Asian	WHITE	=1 if applicant is white

2.2 Methods

2.2.1 Logistic Regression

Logistic Regression is a similar model to linear regression model, specialised in capturing the linear relationship between the response variable and the explanatory variables [13]. When it comes to measuring the predicted outcome of logistic regression, logistic regression would give a predicted outcome of anywhere between 0 and 1, thus a cut-off point equal to the median of the predicted data could be used to classify the object in the case of equal observations for approval and rejection.

2.2.2 Decision Tree

A decision tree operates by splitting the dataset into subsets based on different conditions, progressively narrowing down the possibilities until a decision is reached. At each node of the tree, a decision is made based on these features, leading to a specific outcome: loan approval or denial. This method provides clear interpretability of the model, showing the path leading to a particular decision, which is easier to analyse and interpret.

2.2.3 Random Forest

Random Forest enhances prediction accuracy and mitigates the overfitting issue common in single decision trees. Each decision tree in the forest considers

a random subset of features at each split, introducing randomness into the model building. The final prediction is made based on the majority vote (for classification) or average (for regression) across all trees The algorithm is specialised in handling a large number of independent variables, identify the most significant ones, and model complex interactions between them.

2.2.4 Support Vector Machine (SVM)

Support Vector Machine can be employed as a binary classifier in the problem of predicting loan approval. SVM's principle is to locate a hyperplane in the multidimensional feature space that distinctly classifies the data points (loan applicants) into these two groups. The ideal hyperplane (support vectors) broadens the gap between the closest points to its maximum from the two classes, thereby facilitating the most effective partitioning. Linear, and radical kernel functions will both be used in this research to compare and verify the intrinsic relationship between variables. This flexibility and robustness make SVM an effective choice for predicting loan approval, which often involves complex, high-dimensional data.

2.2.5 Adaptive Boosting

Adaptive Boosting (AdaBoost) seeks to construct a strong classifier which combines several weak classifiers. Within the scope of predicting loan approval, AdaBoost's algorithm is to sequentially educating a range of basic models, typically decision trees, on a segment of the data. Each model in this sequence is designed to amend the errors committed by its predecessor. It does this by assigning higher weights to the misclassified instances, thereby focusing more on the difficult cases in subsequent iterations. This adaptive adjustment allows the algorithm to give more importance to challenging observations. The final fitted value can be calculated using the weighted majority vote or predictions' summation from all the models in the ensemble. AdaBoost's ability to learn from the mistakes and adapt makes it a highly effective algorithm for complex classification tasks such as loan approval prediction, where decision boundaries can be intricate and data may be imbalanced.

2.2.6 Neural Network

A neural network is a type of deep learning model designed to manage intricate and high-dimension datasets. Its structure is composed of layers of interconnected nodes or "neurons", encompassing an input layer, one or several hidden layers, and an output layer. Each node within a layer is linked to all nodes in the succeeding layer, and these links possess weights that are progressively fine-tuned during the training process. The learning procedure entails the use of a technique like backpropagation for determining the gradient of a loss function, and an optimization procedure such as stochastic gradient descent for modifying the weights to reduce the loss. Neural networks can capture non-linear relationships between variables, making them a highly versatile tool for loan approval prediction, capable of modelling complex decision boundaries that other algorithms might miss. However, the neural network would be hardest to interpret comparing to other candidate models because it only gives a 'black box' prediction rather than explaining the relationship between features and the response variables.

3 Results

This paper utilizes the above-mentioned models to fit the cleaned dataset, the models' tuning parameters are determined by which is giving the highest accuracy rate while avoiding underfitting and overfitting. Parameter settings for each method is shown in the table below.

Table 2. Parameter settings for final fitted models

Model	Parameter Setting
Logistic regression	Selected significant predictors: 'APPINC', 'OBRAT', 'EMP', 'SELF', 'OTHER', 'ATOTINC', 'MARRIED', 'MALE', 'PUBHIST', 'LOANAMT', 'PRICE', 'APPR', 'LOANPR', 'MULTI', 'WHITE'
Decision Trees	Minimum sample split=20, Max depth=3, Minimum sample leaf=5
Random Forest	Max_features=16, Max_depth=3
Support Vector Machine	Linear Basis, Cost parameter=10, Gamma=0.0001
Neural Network	Hidden layers =10
Ada Boost	Number of estimators=16, Learning rate=1

3.1 Logistic Regression

This paper analyses the relationship between predictors and the response variable approve using logistic regression, the formula is shown in figure 1.

$$P(y=1)=1 / (1 + \exp(2.7027 -0.0015\times APPINC -0.0450\times OBRAT -0.1157\times EMP 0.6633\times SELF -0.0034\times OTHER -3.2882e\text{-}05\times ATOTINC + 0.3201\times MARRIED -0.2507\times MALE -1.7515\times PUBHIST -0.0048\times LOANAMT -0.0033\times PRICE+ 0.0088\times APPR -2.0479\times LOANPR + -0.4444\times MULTI + 1.0527\times \text{WHITE})) \quad (1)$$

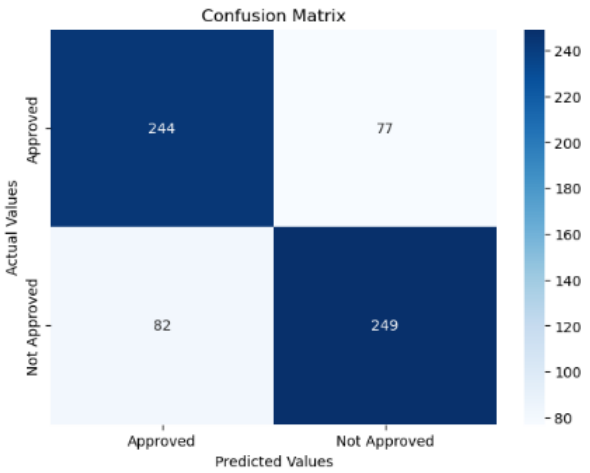

Fig. 1. Confusion Matrix for Logistic Regression.

As indicated by the confusion matrix, the model is relatively balanced with similar false positive and false negative rates and it showcases a good fit accuracy of 0.76 with all predictors significant. All predictors with a positive slope coefficient will exert a positive impact on the approval probability as the variable increase and vice versa. For example, as the appraised value of the property (APPR) increase, the probability of being accepted for mortgage will increase.

3.2 Decision Tree

This paper analyses the relationship between predictors and the response variable approve using decision tree, the tree plot is shown Figure 2. In Figure 2, each decision nodes' class=A represents rejection, and class=P represents approval. The root node is OBRAT<=41.1, branches include WHITE, LOANPR and etc. For example, if the applicant's ratio of total monthly obligations to monthly income (OBRAT) is less than 41.1 and is not white (WHITE=0), and has a record of default (PUBHIST=1), the application is likely to be rejected. Figure 3 shows the decision tree is biased towards application denial, leading to high false negative rate. Figure 4 indicates OBRAT, WHITE and PUBHIST are the most important features.

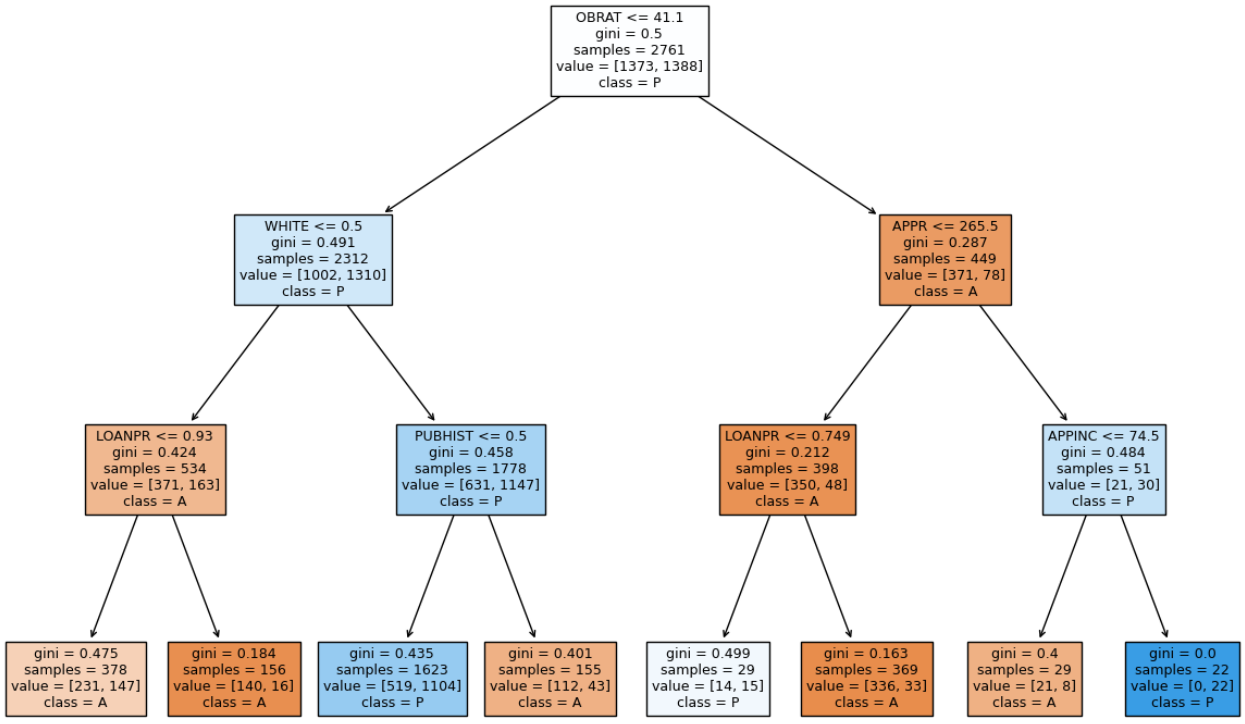

Fig. 2. Fitted decision tree.

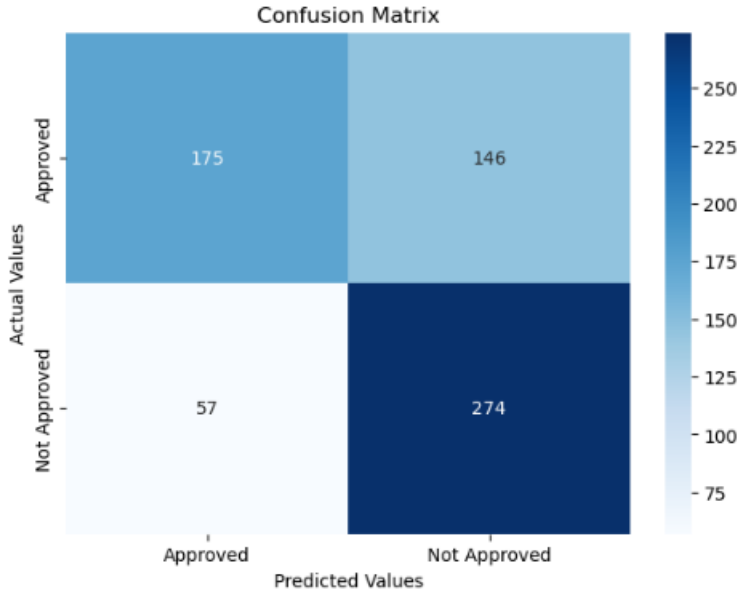

Fig. 3. Confusion matrix for decision tree classifier.

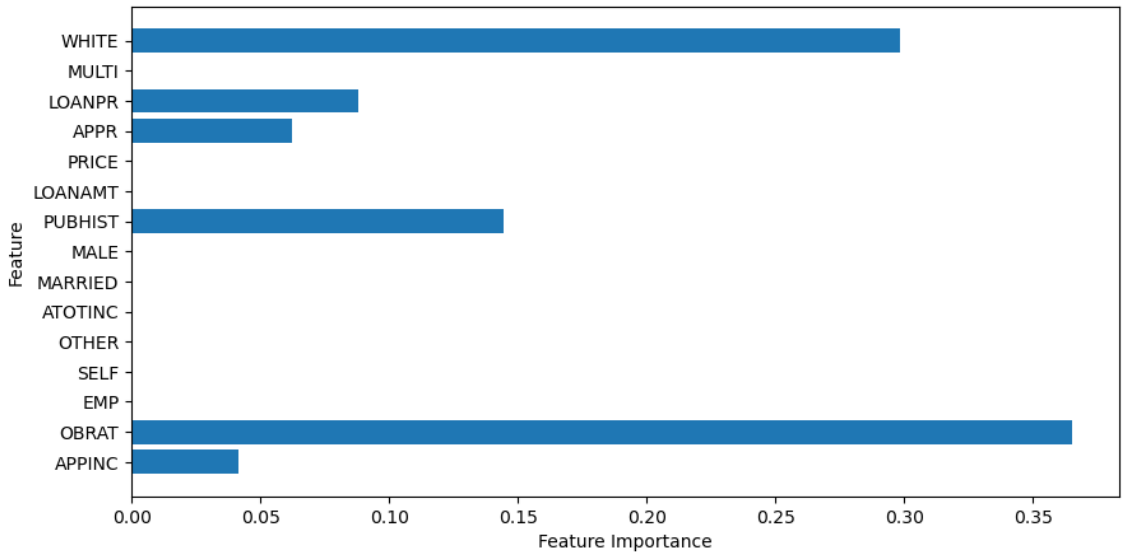

Fig. 4. Feature importance for each explanatory variable.

3.3 Random Forest

The result of random forest is shown in Figure 5 and Figure 6. The fitted result is biased towards application denial, leading to high false negative rate, nonetheless, less biased than the single decision tree. As a 'Black box model', one can seek for business interpretation by computing the feature importance [14]. The factors of most importance are OBRAT, WHITE and PUBHIST, suggesting debtors could lead to different application outcomes by changes those predictors' values.

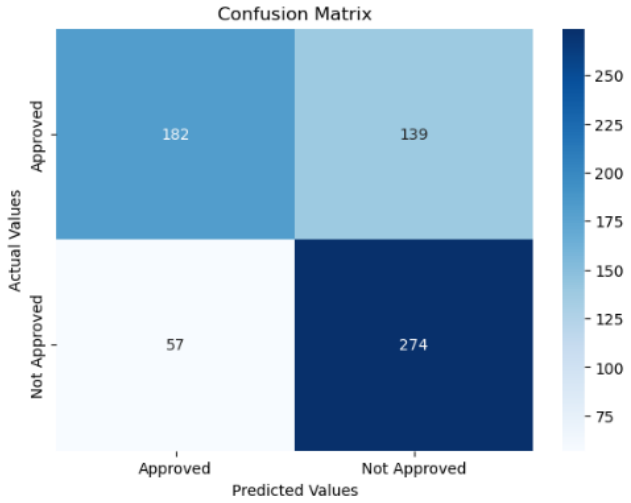

Fig. 5. Feature importance for each explanatory variable.

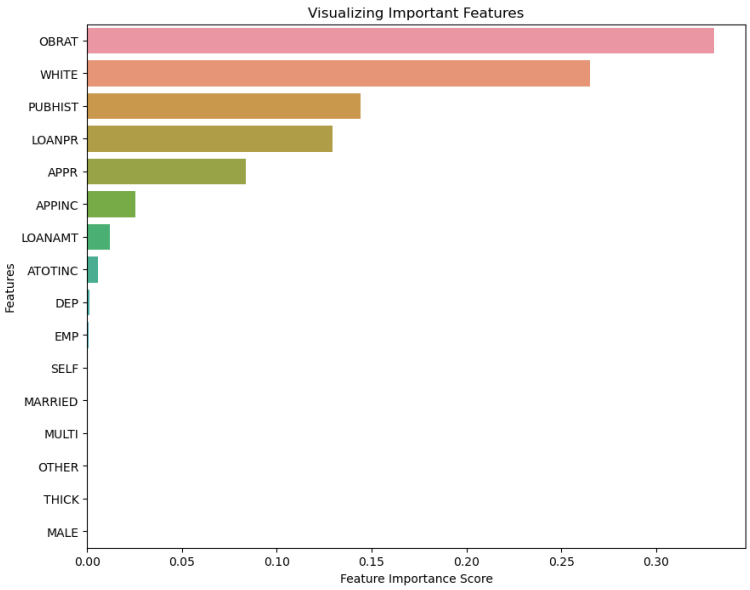

Fig. 6. Feature importance bar plot for each explanatory variable.

3.4 Support Vector Machine (SVM)

3.4.1 Non-Linear kernel

As shown in Figure 7. The Radial basis function kernel (RBF) gives very complicated decision boundaries.

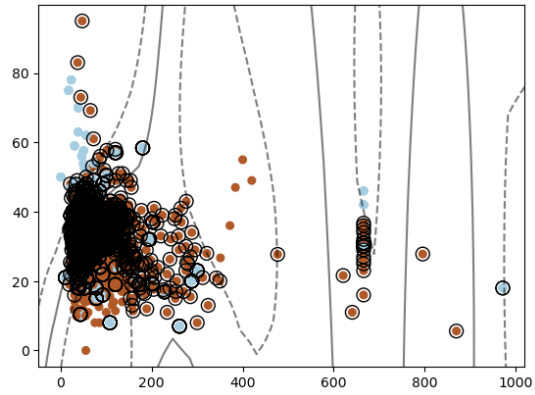

Fig. 7. This figure plots the non-linear decision boundary between LOANAMT (x) and OBRAT (y)

3.4.2 Linear kernel

As shown in Figure 8. The linear kernel fit seems more straightforward, although this plot indicates the existence of higher dimensional relationships, application with lower OBRAT and LOANAMT are more likely to be accepted as indicated by orange observations. As shown in Figure 9, the confusion matrix showcases slighted higher false negative rate than false positive rate. PUBHIST, LOANAMT and WHITE are the most important factors (Figure 10).

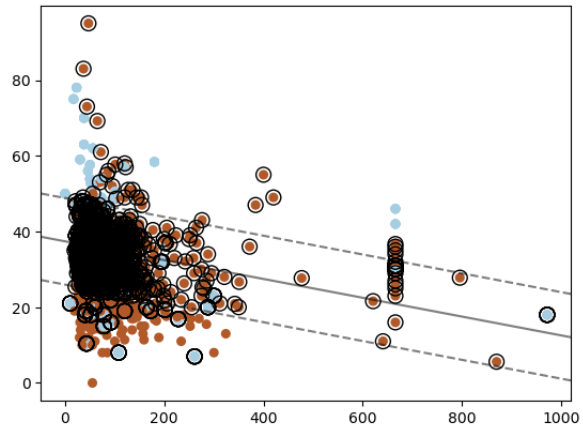

Fig. 8. This figure plots the linear decision boundary between LOANAMT (x) and OBRAT (y)

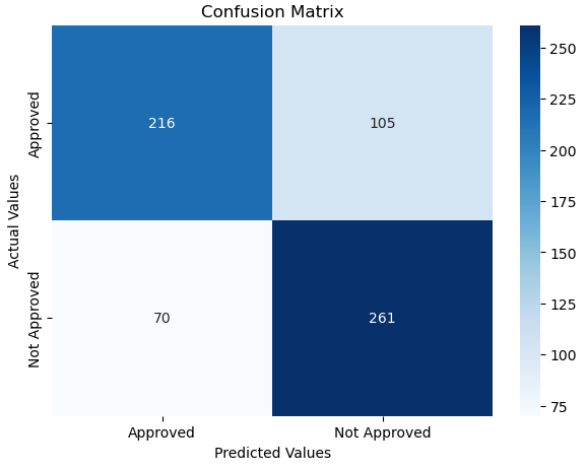

Fig. 9. The confusion Matrix of linear SVM.

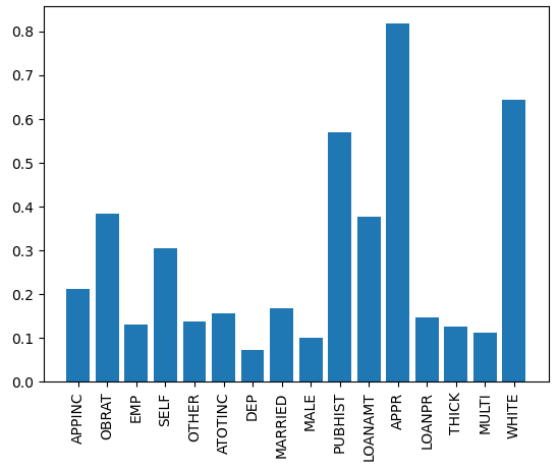

Fig. 10. Feature importance of the fitted SVM.

3.5 Adaptive Boosting

For applicants with total monthly income lower than 3416.5, there are higher chance of being accepted than being rejected, while an ATOTINC greater than 3416.5 will lead to the opposite (Figure 11).

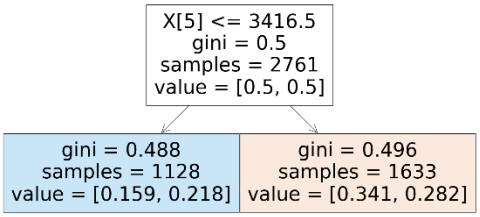

Fig. 11. Single decision tree with respect to ATOTINC.

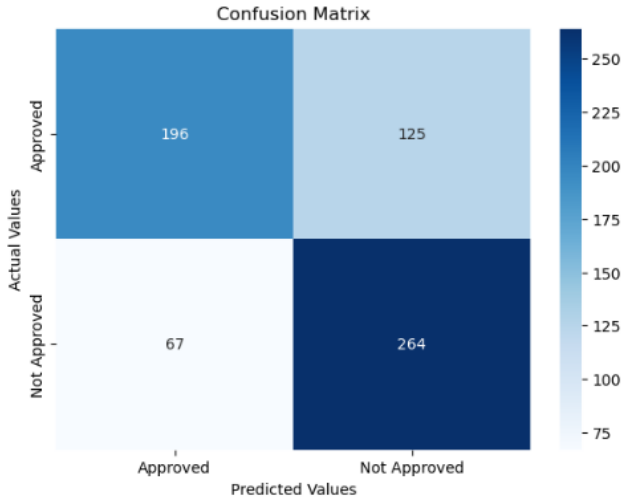

Fig. 12. Confusion Matrix of AdaBoost.

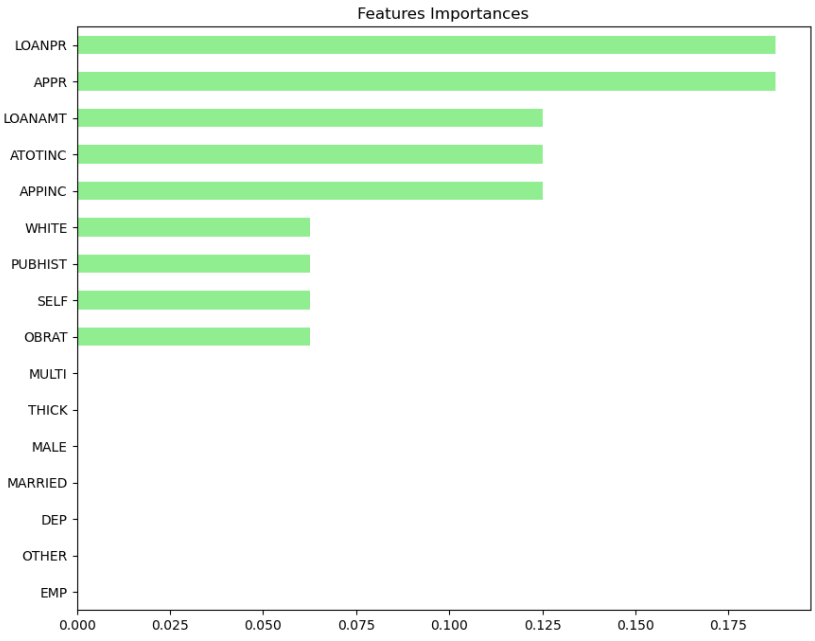

Fig. 13. Feature importance of AdaBoost.

The confusion matrix showcases higher false negative rate than false positive rate (Figure 12). When looking at the overall importance of predictors, LOANPR and APPR are most significant (Figure 13).

3.6 Neural Network

As shown in Figure 14, the confusion matrix demonstrates very high bias towards rejection. Feature importance is normalised by permutation to avoid bias also for the ease of computation [16]. WHITE, LOANAMT, APPR and OBRAT are observed to the factors of the most importance (Figure 15).

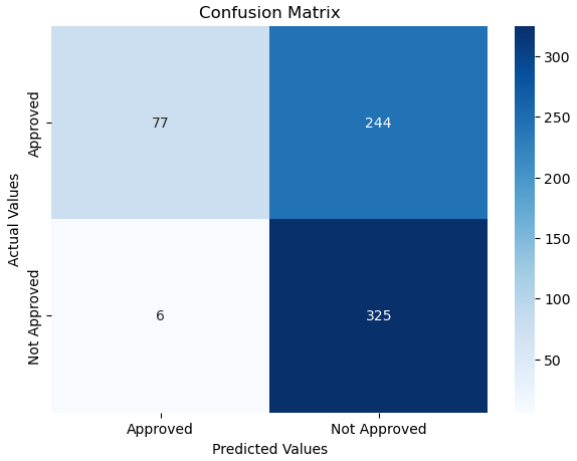

Fig. 14. Confusion Matrix of Neural network.

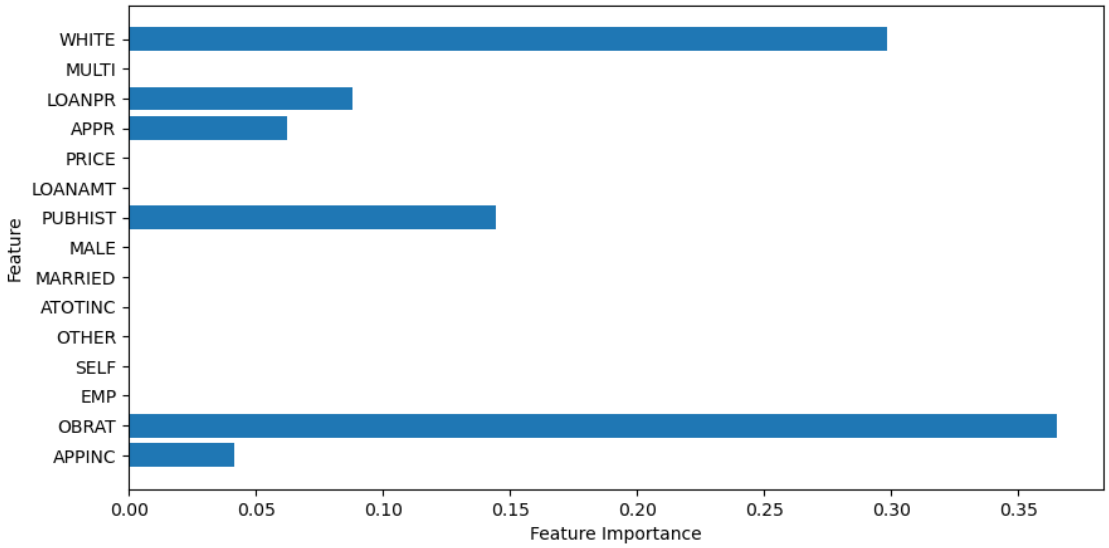

Fig. 15. Permutation importance for neural network model.

3.7 Model Comparison

As demonstrated in Table 3, Logistic regression has the highest test accuracy rate of 0.76, with SVM being the second highest of 0.73. The worst performing model in terms of accuracy is neural network, which has an accuracy rate of 0.63.

Table 3. Accuracy rate of each model.

Model	Accuracy
Logistic Regression	0.76
Decision Tree	0.69
Random Forest	0.71
SVM (linear)	0.73
Ada Boost	0.70
Neural Network	0.63

In terms of interpretability, 'black-box' models give debtor limited information about area of improvement with respect to their application, it only gives an estimate of if the application would succeed. Yet, in this study, 'black-box' models such as neural network, random forest performs worse than intuitive models such as logistic regression and SVM with respect to both accuracy and accessibility.

Therefore, compare to other candidate models，Logistic regression is the best selected model both in terms of testing accuracy and interpretability to predict loan approval.

However, it is worth mentioning that in each model, the variables of highest importance are different, nonetheless, WHITE, LOANPR, APPR and OBRAT appears to be important generally, which leads to the conclusion that debtors should review these variables especially carefully to maximize the probability of being accepted.

The predictors with positive coefficient in the logit regression model means by manipulating and increasing these variables, the debtor will have a higher chance of being approval for home loan. Explanatory variables with negative coefficient are to keep as low as possible since it will likely to decrease the chance of receiving an approval. Except Price, all other predictors in the logistic model are significant. The most variables that has the most effect are PUBHIST, LOANPR and

WHITE which people should be aware of and try to follow the direction of marginal effect of these factors by for example avoiding defaulting a loan at any time and decrease the ratio of loan to purchase price to a suitable extent to have a fair chance of being accepted and also plan an affordable onetime payment.

4 Conclusion

Logistic regression is the best model in predicting for loan approval in the given dataset. Logically, logistic regression captures a linear relationship between predictors and approval and is easier to be interpret and used. The result of other models can be used to give indication after noticing the major field of error for example a 'black box' model with low false positive rate and high false negative rate returning a positive result will indicate that the applicant has a strong application.

WHITE (if the applicant is white), LOANPR (loan to purchase price ratio), APPR (appraised value of the property) and OBRAT (total monthly obligations to monthly income ratio) most significantly determines if the loan would be approved.

There are two main implications in this paper. Firstly, debtors are interested in finding out the coefficients of the explanatory variable in the bank's system. They aim to Increase chance of getting approved (by increasing the magnitude of variables with positive marginal effect and vice versa in the logistic model, such as reporting adjusted income level, or report a mixed race rather than reporting a race that could disadvantage them). In addition, debtors can apply for multiple banks to increase their overall chance of being accepted for home loan if they are predicted to be given approval. Secondly, bank is assumed to approve the loan if there is low level of credit risk, the bank can fit the logistic model to determine if their algorithm is out of date or exploited by potential debtors. They can adjust the weights of respective variables in their credit system. For example, the weight of PUBHIST– whether the applicant has defaulted before, can be reduced in areas with high probability of faking the history, while increase in areas with strict online record of default history.

However, there are limitations in the study. Firstly, although non 'black-box' models tend to perform better generally, different dataset with different predictors might yield different preference of model, thus this research method could be in future tested on different loan approval datasets to validify. Secondly, there are unconsidered predictors which might be very significant but is hard to collect data for: the region of the applicant, mental state. Thirdly, Logistic regression parameters also need to be chosen reasonably for the model to be useful base on significance of predictors and test accuracy.

References

1. Wooldridge: 115 Data Sets from "Introductory Econometrics: A Modern Approach, 7e" by Jeffrey M. Wooldridge version 1.4-3 from CRAN (rdrr.io)
2. K. Arun, G. Ishan, K. Sanmeet, Loan approval prediction based on machine learning approach, IOSR J. Comput. Eng. **18**, 18-21 (2016).
3. J. Tejaswini, T.M. Kavya, R.D.N. Ramya, P.S. Triveni, V.R. Maddumala, Accurate loan approval prediction based on machine learning approach, J. Eng. Sci. **11**, 523-532 (2020).
4. M.A. Sheikh, A.K. Goel, T. Kumar, An approach for prediction of loan approval using machine learning algorithm, in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), IEEE, 490-494 (2020).
5. A. Vaidya, Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval, in 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE, 1-6 (2017).
6. A.S. Kadam, S.R. Nikam, A.A. Aher, G.V. Shelke, A.S. Chandgude, Prediction for loan approval using machine learning algorithm, Int. Res. J. Eng. Technol. **8** (2021).
7. P. Tumuluru, L.R. Burra, M. Loukya, S. Bhavana, H.M.H. CSaiBaba, N. Sunanda, Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms, in 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), IEEE, 349-353 (2022).
8. U. Aslam, H.I. Tariq Aziz, A. Sohail, N.K. Batcha, An empirical study on loan default prediction models, J. Comput. Theor. Nanosci. **16**, 3483-3488 (2019).
9. T. Ndayisenga, Bank loan approval prediction using machine learning techniques, Doctoral dissertation (2021).
10. P.S. Murthy, G.S. Shekar, P. Rohith, G.V.V. Reddy, Loan Approval Prediction System Using Machine Learning, J. Innov. Inf. Technol. 21-24 (2020).
11. P.S. Murthy, G.S. Shekar, P. Rohith, G.V.V. Reddy, Loan Approval Prediction System Using Machine Learning, J. Innov. Inf. Technol. 21-24 (2020).
12. Y. Diwate, P. Rana, P. Chavan, Loan Approval Prediction Using Machine Learning, Int. Res. J. Eng. Technol. **8**, 1741-1745 (2021).
13. D.W. Hosmer Jr, S. Lemeshow, R.X. Sturdivant, Applied logistic regression, John Wiley & Sons (2013).
14. A. Palczewska, J. Palczewski, R.M. Robinson, D. Neagu, Interpreting random forest models using a feature contribution method, in 2013 IEEE 14th International Conference on Information Reuse & Integration (IRI), IEEE, 112-119 (2013).
15. A. Sánchez, V.D. Advanced support vector machines and kernel methods, Neurocomputing **55**, 5-20 (2003).
16. A. Altmann, L. Toloşi, O. Sander, T. Lengauer, Permutation importance: a corrected feature

importance measure, Bioinformatics **26**, 1340-1347 (2010).