Tree-Based Methods for Loan Approval

Mohamed Alaradi
MSc. in Big Data Science and Analytics
University of Bahrain
mohamed.alaradi@aspentech.com

Sawsan Hilal
College of Science
University of Bahrain
shilal@uob.edu.bh

2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI) | 978-1-7281-9675-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICDABI51230.2020.9325614

***Abstract*—** Loan approval is one of the most important processes that any banking organization owns. The acceptance or rejection of any loan application has a direct impact on the bank revenue and the profitability in quarterly issued financial statements. Though loan approval is a critical process, the actual decision made is not a straightforward procedure and comes with a lot of uncertainties. Recently, statisticians and data scientists have tried to automate this process to minimize risk and increase profitability by applying different statistical learning methods. In this work we explore a framework with an application by applying tree-based methods on publicly available dataset. This work aimed at developing a high performance predictive model for loan approval prediction using decision trees. Experiments were made in different varieties of tree methods ranging from the most simplified and comprehensible decision tree reaching up to the most complex random forests. Results yielded inadequate performance with respect to simplified decision trees due to the highlight correlated and complex feature space, majority of critical parameters affecting loan approval was not reflected upon and yielded an impractically over-simplified tree. However, boosting came in superior in terms of performance, relevance and interpretation via the importance chart scoring accuracy on testing dataset [98.75%] specificity [100%], Minority class prediction accuracy [92.85%], and classification efficiency of [97.0%]. Therefore, boosting-based decision-tree predictive model was recommended to facilitate decision making regarding the eligibility of loan applicants based on their characteristics.

Keywords—decision making, loan eligibility, predictive model

I. INTRODUCTION

The banking and financial sector is one of the richest sectors in data availability. In recent years the amount of data has surged up significantly even as more technologies emerged to facilitate automated and cost-effective data collecting methods [1]. Many banks are investing heavily in resources and technologies for data collection. The recent emphasis has been on the selection of data collection approaches that are relatively "cost effective" and could satisfy certain requirements. Preferred approaches should be able to significantly lower the overall reporting burden on respondents without compromising too much the level of detail on data received. Despite this ongoing trend, even the institutions within the financial sector that has not gone towards this direction, data is readily generated from the nature of the operational process of these institutions. This leaves a great opportunity to leverage data for automated, risk free decisions making for day to day operations that minimizes cost and increase profitability [2].

Distribution and approval of loans is one of the core business parts of any operational retail bank. It resembles a major income for the bank's capital for its assets. With the enhancement of the technologies in the banking sector and the daily processes becoming much more streamlined and efficient, this caused a larger influx of applicants for loans seeking capital from banks which readily expands any bank reservoir of data for this type of decision making [3,4]. The process of selecting an eligible candidate is a typical process and each financial institution is somewhat unique and each follows a specific standardized workflow. This type of decision making has an inherent risk associated with it. Moreover, selecting and developing the proper resources to make that judgment also possess another operational risk and cost.

The current work aims at automating the process of loan approval by using statistical learning methods based on a bank's historical data that outlines certain characteristics of the applicants. The rationale behind this work is obviously minimizing risk and enabling bank personnel to take better decision and judgment on selecting eligible candidate for loan approval [5,6,7]. The rest of the paper is organized as follows. Section II details the exploratory data analysis including data summary and data processing as well as investigating response-predictor relationships. Section II is devoted to the description of the implemented tree-based methods, while the study results along with the relevant discussion are collected in Section III. The adopted criterion for building the developed predictive model is outlined in Section IV. Section V concludes the study.

II. EXPLORATORY DATA ANALYSIS

A. Data Summary

The dataset that has been analysed in this work was obtained from Analytics Vidya competition [8]. It consists of 981 observations for different applicants with various attributes that are usually collected for the loan approval process. These attributes are represented by 9 variables along with the corresponding decision as the target variable. A summary statistics of the variables under study are collected in TABLE I.

TABLE I. VARIABLES SUMMARY

Variable	*Counts (%)*
Gender	Male: 80.94% Female: 19.06%
Marital Status	Married: 64.53% Not Married: 35.47%
Education	Graduate: 77.87% Non-Graduate: 22.22%
Job	Employer: 87.36% Self- Employed: 12.64%
Credit History	Meeting Standards: 84.00% Off Standards: 16.00%
Location	Rural: 29.56% Semi-Urban: 35.58% Urban: 34.86 %
Dependent	0: 57.06 % 1: 16.75% 2: 16.65% 3: 9.53%
Income (in dollars)	Median = 5,314 $, IQR = 3,142
Loan Amont (in $1000s)	Median = 126.00, IQR =
Loan Status	Approved: 74.41% Not Approved: 25.59%

B. *Data Processing*

The data processing techniques covered data transformation and missing data imputation following the approach adopted in [9]. There was a seemingly extreme values present in both loan amount applied for and the income of certain individual. Accordingly, different forms of data-transformation have been implemented to normalize the variables of interest, but the most appropriate form was found to be the log-transformation as the distributions of the transformed variables were found to be close to symmetric and evidently right skewness was corrected.

A separate case was made to scale the same numerical variables to have a mean of 0 and a variance of 1. This is done for a special case that will be addressed later when discussing the methods implemented. Moreover, missing values in the data were handled by the imputation method that filled data gaps according to Synthetic Minority Over-sampling Technique.

C. *Response-Predictor Relationship*

To investigate the relationship between income and loan status, the income has been categorized as shown in Fig. 1 and the difference between these income-based categories in relation with loan status was tested statistically using ANOVA where the corresponding p-value (0.568) indicates insignificant relationship between income and loan status.

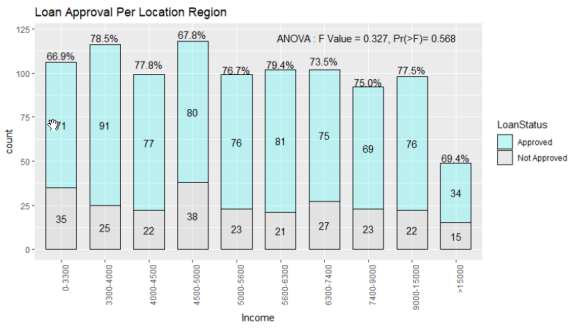

Fig. 1. Loan Status and Income relationship

There is a slight decrease of approval rate by approximately 1% for self-employed applicants as shown in Fig. 2. However, Chi-Square test (p-value = 0.624) shows no significant association between job class membership and the loan status. The latter realization applies to the association between applicant's gender and the loan status (Fig. 3) since their association was insignificant (p-value = 0.559).

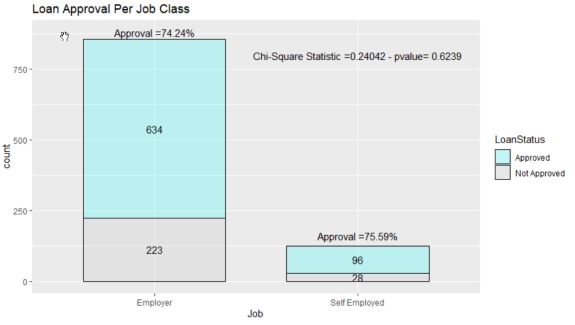

Fig. 2. Loan Status and Job Class Relationship

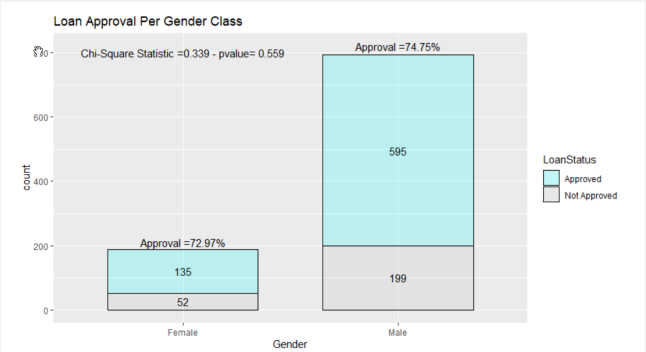

Fig. 3. Loan Status and Gender Relationship

On the contrary, the level of applicant's education had a significant influence on the approval rate from the bank such that non-graduates were less likely to get their loans approved by roughly 8% as shown in Fig. 4, though the influence was moderate as indicated by the reported p-value (p-value = 0.039).

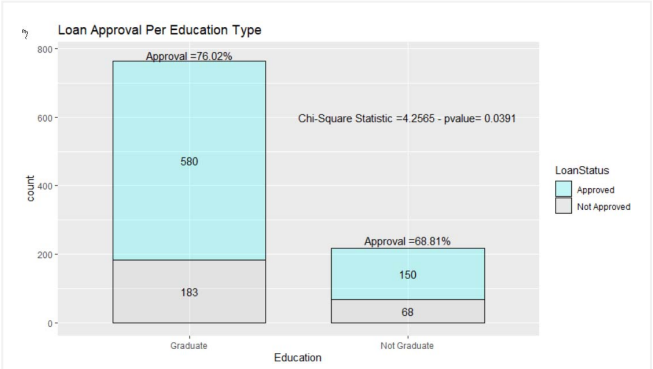

Fig. 4. Loan Status and Education relationship

On the other hand, the credit history seemed to be the most influencing attribute to the loan approval which is an indication that the applicant's credit history acts as an initial screening to the loan applications. The loan approval for non-conforming applicants was 5.66% while the credit regulation conforming applicant's approval rate was 87.71% as shown in Fig. 5.

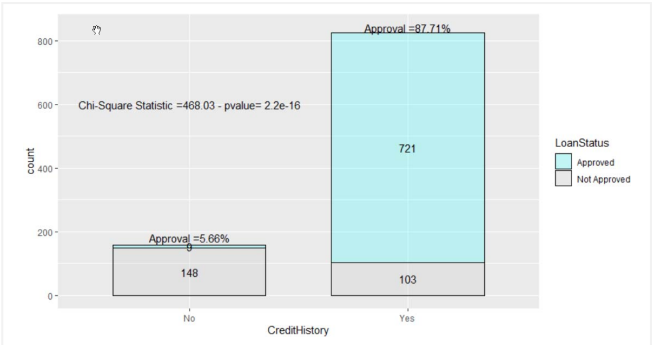

Fig. 5. Loan Status and Credit History Relationship

Moreover, the applicants residing in semiurban areas had slight increase of approval rates as depicted by Fig. 6 and indicated by the corresponding p-value = 0.029. However, the viewpoint on such attributes might have a deeper causality. For example, it is possible that people residing in

such area have better social and financial status enabling them to feature better in the approval criterion by the bank.

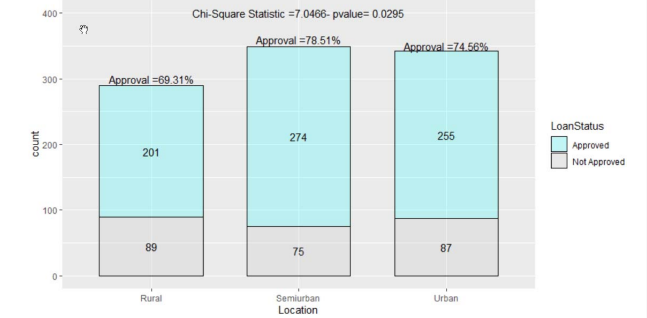

Fig. 6. Loan Status and Location Relationship

Rationally speaking, the number of dependents on its own as an attribute should cause lower approval rates as the family would be under higher financial obligations as reflected by Fig. 7. However, this viewpoint isn't enough as it is possible that people with higher number of dependents would most likely have higher income or better background in terms of education, and hence, the correlation analysis from the univariate viewpoint seems to be insufficient to anticipate a clear conclusion about the dependents-target relationship.

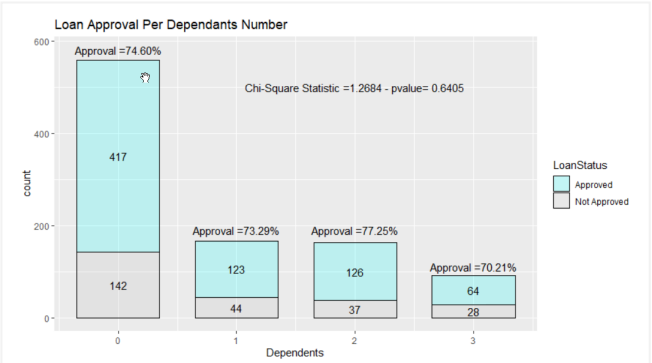

Fig. 7. Loan Status and Number of Dependents Relationship

Similarly, no influence of the loan amount on the approval decision was apparent as shown in Fig. 8 and also confirmed by the statistical test (p-value = 0.568).

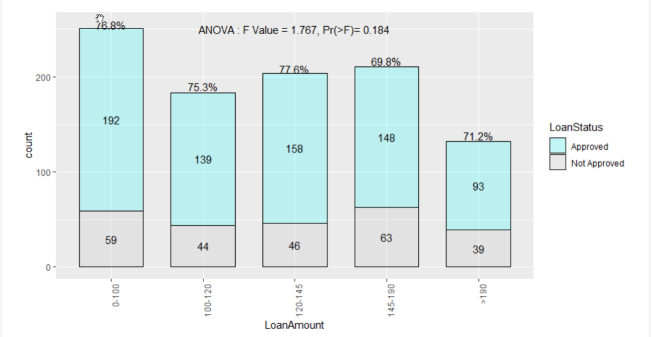

Fig. 8. Loan Status and Loan Amount Relationship

The exploratory data analysis concluded that only a few variables are highly correlated with the approval status of the loans due to the fact that most of those variable comes from an individual and the possibility of attribute collinearity is high. This early analysis inferred that parametric methods are likely to suffer in predicting the loan status accurately, and hence, a method that segments the observation space very finely will be better suited for the data at hand.

III. Tree-Based Methods

This section covers a brief introduction about the implemented methods to predict loan status (approval/disapproval). It mainly outlines the theoretical background behind each method, its underlying assumptions, limitations and the model selection criterion followed in the scope of this work [10,11].

A. Method 1: Decision Tree

A decision tree is a type of supervised learning algorithms that can be used in both regression and classification problems. It works for both categorical and continuous input and output variables. It is a powerful tool for facilitating decision making in sequential decision problems. The attribute space is segmented using a recursive binary greedy algorithm where the optimal decision is made step-by-step by splitting the larger attribute space into two further subspaces to minimize the classification error. Here are some properties of decision trees.

- They are graphically represented with a simple interpretation as they mimic human decision-making.
- They make no assumption on the distribution of the attributes/predictors and they can handle both numerical and categorical variables.
- They can be aggregated by using methods like bagging, random forests and boosting which will be addressed in the next sections.

A classification tree is used to predict a qualitative variable based on certain attributes. In the context of this work, the interest lies in the prediction of the loan status based on applicant's characteristics. More specifically, with classification trees, the prediction made based on the majority vote in the sense that each observation belongs to the most commonly occurring class of the training observations in the region to which it belongs.

Apart from the usefulness of classification trees, they tend to be complex and hence suffer from overfitting. Pruning is a process used to reduce complexity of the tree bottom up by restricting the error term such that a proper balance is made between model variance and bias.

B. Method 2: Bagging

Bagging is a special case of ensemble methods. An ensemble method is a machine learning technique that combines several base models in order to produce one optimal predictive model. Reflecting on this definition, then bagging creates multiple full classification trees based on bootstrap sample. When a new observation is in and a prediction is needed, the bagging method calculates the predication of each built tree and aggregates the results in the form of a majority vote. The class that the trees predicted highest in vote is the overall model prediction. This procedure of majority vote reduces the variance of decision trees inherently and this is the core reason why the trees could grow without any restriction or pruning in the building

phase with no worry on an increase of variance. Bagging is considered a special case of random forests which will be addressed up next.

C. Method 3: Random Forests

The random forests adopt the same procedure as bagging by building several decision trees on bootstrap samples. However, the difference is that each time a split in a tree is considered, a random selection of predictors is chosen as split candidates from the full set of predictors. This technique has the advantage of de-correlating the constructed decision trees and hence reducing the variance when averaging the trees.

D. Method 4: Boosting

Following the same methodology as bagging and random forests by creating multiple trees but in the boosting case the trees are constructed sequentially so that each tree is grown using information from the previous tree. The transfer of information is conducted by calculating the residuals (classification errors) at each tree building step and using it to initiate the other one accordingly. The mathematical procedures conducted are not of primary concern in the context of this work. The most important thing is to understand is the functionality behind these multiple tree methods. Essentially, bagging, random forest, and boosting all improves prediction accuracy for classification tree by decreasing variance, yet this comes at the cost of harder interpretability because the trees are no longer graphically represented due averaging.

E. Performance Metrics

For a binary classification such as the problem under consideration, the performance of the model is assessed by several metric as detailed below [10,12].

- Sensitivity (Sens) is defined as the proportion of positive cases that are correctly identified by the model. In the context of this work, the disapproved loan applications represent the positive cases.
- Specificity (SPEC) is defined as the proportion of negative cases that are correctly identified by the model.
- Accuracy (ACC) is defined as the proportion of all cases that are correctly predicted by the model.
- AUC refers to the area enclosed by the Receiver Operating Characteristic (ROC) where the latter is a curve representing the true positive rate (sensitivity) against false positive rate (1-specificity) are used to illustrate the ability of the classification model when varying the discrimination threshold.

F. Programming Software

The relevant analysis of the studied data was conducted using R statistical programming language via R-Studio.

IV. Model Building Criterion

The data was treated initially as detailed in Section II and was segmented into two portions one for training and the other one for testing and validation. Specifically, cross validation was used to ensure conservative estimates of the testing error. Once the accuracy converges to a satisfactorily result the loop is broken and a final model is presented. This criterion for model building is represented by Fig. 9.

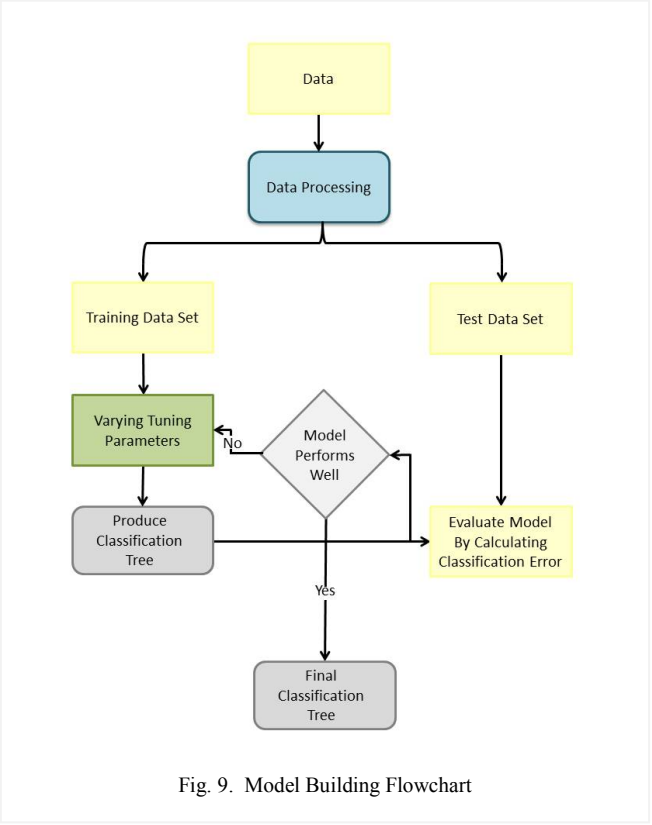

Fig. 9. Model Building Flowchart

V. Results and Discussion

A. Decision Trees: With and Without Pruning

The decision-tree-based model was applied on two different methodologies. The first was restrictive where a pruning procedure was put in place to avoid unnecessary branching and possible high variance, while the second methodology was set free without any restriction. Fig. 10 shows the fully-grown decision tree without any restriction or pruning effects, while the corresponding restricted/pruned decision tree is represented by Fig. 11.

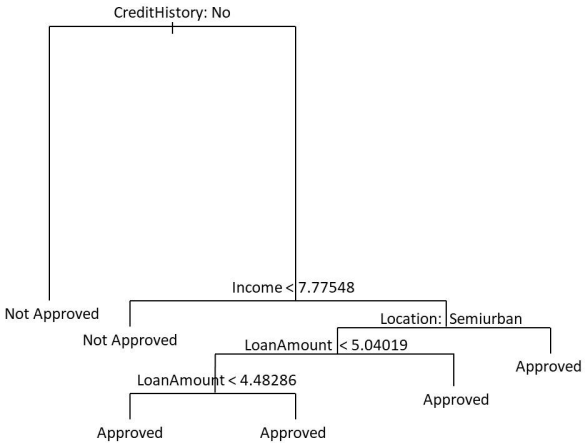

Fig. 10. Decision Tree (Unpruned)

Fig. 11. Decision Tree (Pruned)

The resulting decision trees are quite self-explanatory where the main branch is shown on the top and represents the root of the decision tree and according to the model this is the most important factor (credit history) as it segregates the response space the finest. Indeed, this result was expected as it has been shown by the exploratory data analysis that the credit history was the most highly correlated attribute with the loan status.

TABLE II showcases the performance of both the pruned and unpruned decision trees in terms of several metrics on which it is clear that the two approaches are very closely matching. The only lacking parameter is the sensitivity. The reported results overall show good performance for both models. The models classify the approved category quite well which is represented by the specificity. However, they underperform the not approved category and lay out almost 8% of the not approved applications to the approved category. This very strict criticism comes from the fact that the original data set was imbalanced with the "not approved" being the minority category [13].

TABLE II. DECISION TREE PERFORMANCE METRICS

Model	ACC [%]	SPEC [%]	Sens [%]	AUC [%]
Decision Tree (No Pruning)	97.25	98.18	92.85	96.88
Decision Tree (Pruning k=2)	97.25	98.18	92.85	96.38

Acc : Accuracy , SPEC : Specifcity ,Sens : Sensitvity ,AUC : Area under curve metric

As an inherent feature within the decision tree, it captures all the behavior of the data in a singular representation. The illustrated trees do not represent the actual process followed by the bank to approved or disapprove loans since the majority of the factors that would rationally influence loan approval was not considered in these oversimplified trees. This oversimplification is a limitation of decision trees.

B. Bagging

The performance metrics for bagging (with 500 trees constructed) are collected in TABLE III. The bagging-based model performed worst in overall accuracy as well as specificity. It performed less efferently in classifying the approved category with higher rates for false negatives which translates into the problem statement of this work as lost opportunity to grant a loan.

Since bagging is an algorithm made of multiple trees, graphical representation of how these trees are constructed is impractical. However, variable importance can be plotted as shown in Fig. 12 which clearly highlighted the credit history as the most infleuencial variable on loan status.

TABLE III. BAGGING PERFORMANCE METRICS

Model	ACC [%]	SPEC [%]	Sens [%]	AUC [%]
Bagging [m= p]	92.52	92.44	92.85	97.39

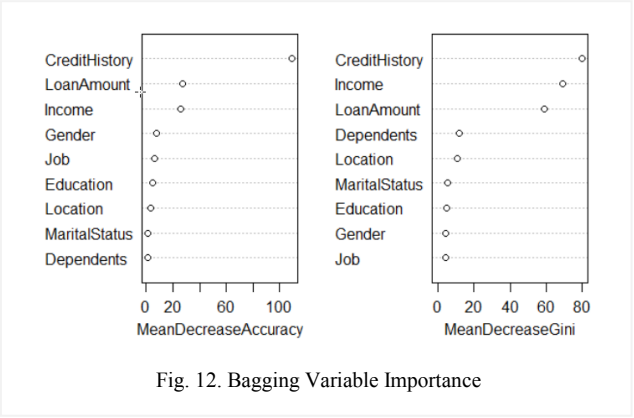

Fig. 12. Bagging Variable Importance

C. Random Forest

Similar to bagging, random forest is an algorithm made of multiple trees (with 500 trees in this case), and hence a graphical representation of the final model is impractical.

Fig. 13 shows variable infleuence on loan staus. The order here could be infered as importance in its raw form, and it could also mean rarity of cases where the factor under consideration is used to approve or disapprove an application decisvily. The rank in this chart does not mean importance only as the rank also is dependenant on case presence within the dataset.

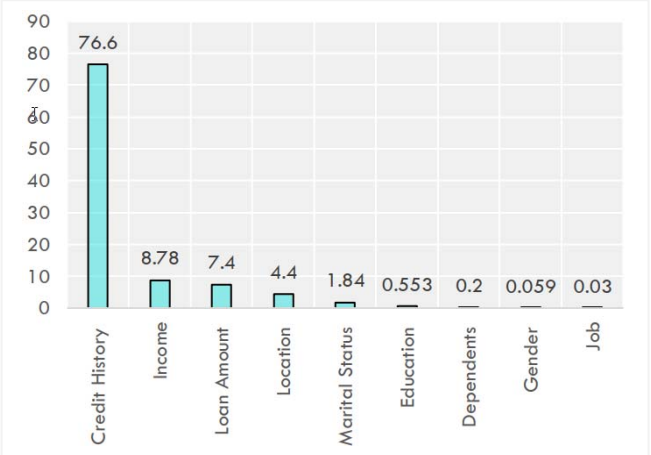

Fig. 13. Random Forest Variable Importance

The relevant results with the key performance metrics are summarized in TABLE IV. Therefore, from the fitted models, it was observed that multiple tree methods

outpefrom standard decision trees as they segment the feature space with more granularity.

TABLE IV. RANDOM FOREST PERFORMANCE METRICS

Model	ACC [%]	SPEC [%]	Sens [%]	AUC [%]
Random Forest [m= sqrt(p)]	96.25	96.97	92.85	96.91

D. Boosting

The pefroemance metics of boosting were showcased in TABLE V using 5000 trees and shrinkage factor of 0.001.

TABLE V. BOSSTING PERFORMANCE METRICS

Model	ACC [%]	SPEC [%]	Sens [%]	AUC [%]
Boosting	98.75	100	92.85	97.00

VI. CONCLUSION

In this work, many statistical learning classification techniques were performed under the pursuit of predicting loan approval status for a bank. The focus was on decision trees, random forests with different variants and boosting. The decision tree approach suffered to establish a thorough and meaningful relation between the attributes and the loan status, but it was not due to some infringement of some of its core assumptions, rather because the apparent true function for deciding loan status is too complex to represent in a singular decision tree. Accordingly, multiple tree techniques proved to be the most representable methods in this scope of work. This included random forests, bagging and boosting. Those methods are effective to be utilized for modelling this type of data as it models multiple decision trees and ultimately computes the common vote. The three methods shared the same ranking of variable importance (credit history, income, loan amount, location, marital status, and education) ordered from highest to lowest. However, amongst the implemented multiple tree methods, boosting came in superior according to selection criteria outlined earlier with accuracy of 98.75%, specificity of 100%, sensitivity of 92.5%, and AUC of 97%.

REFERENCES

[1] C. Agarwal, and Musigchai, "*Background Note on Data Collection Techniques*", IFC Bulletin No 30, 2011. Available: https://www.bis.org/ifc/publ/ifcb30c.pdf

[2] L. Al-Blooshi and H. Nobanee, "Applications of Artificial Intelligence in Financial Management Decisions: A Mini-Review", *SSRN Electronic Journal, 2020.*

[3] W. Kluwer, Bizfillings, "What Banks Look for when Reviwing a Loan Application", 2019. Available: Bizfillings.com: https://www.bizfilings.com/toolkit/research-topics/finance/business-finance/what-banks-look-for-when-reviewing-a-loan-application

[4] Moody's ANALYTICS, "Improve the Loan Approval Process by Implementing the Credit Decision Strategy", n.d. Available: https://www.omega-performance.com/improve-the-loan-approval-process-by-implementing-a-credit-decision-strategy/

[5] V. Nagajyothi, "Loan approval prediction using KNN, decision Tree and Naïve Bayes models", *International Journal of Engineering in Computer Science*, vol. 2, pp. 32-37, 2020.

[6] K. Arun, G. Ishan, and K. Sanmeet, "Loan Approval Prediction based on Machine Learning Approach", *IOSR Journal of Computer Engineering*, pp. 18-21, 2009.

[7] R. Kumar, V. Jain, P.S. Sharma, S. Awasthi, and G. Jha. "Prediction of Loan Approval using Machine Learning", *International Journal of Advanced Science and Technology*, vol. 28, pp. 455-460, 2019.

[8] Analytics Vidhya, "Loan Prediction Practice Problem", 2020. Available: https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/.

[9] X. Francis Jency, V.P. Sumathi, and J. Shiva Sri, "An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients". *International Journal of Recent Technology and Engineering,* vol. 7*,* 2018.

[10] G. James, D. Witten, T. Hastie, and R. Tibshirani, "An Introduction to Statistical Learning: with Applications in R". New York: Springer, 2013.

[11] Institute for Science & Computing - University of Miami, "Decision Tree: Introduction". n.d. Available: http://web.ccs.miami.edu/~hishwaran/papers/decisionTree_intro_IR2009_EMDM.pdf

[12] My R Codes Archive, "*Area Under Curve (AUC) - pROC package*", 2013. Available: http://myrcodes.blogspot.com/2013/12/area-under-curve-auc-proc-package.html

[13] V. Ganganwar, "An Overview of Classification Algorithms for Impalanced Datasets", *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, 2012.