International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-10, August 2020

# The Prediction of Application for Loan using Machine Learning Technique

**Youngkeun Choi, Jae Won Choi**

***Abstract*: Machine learning techniques are used to verify the many kinds of loan prediction problems. This study pursueS two major goals. Firstly, this paper is to understand the role of variables in loan prediction modeling better. Secondly, the study evaluates the predictive performance of the decision trees. The corresponding variable information is drawn from a third-party website, international challenge on the popular internet platform Kaggle (www.kaggle.com), which provides data in the title of 'Loan Prediction' that was uploaded by Amit Parajapet. We used decision tree which is a powerful and popular machine learning algorithm to this date for predicting and classifying big data. Based on these results, first, women seem to be more likely to get to loan than men. credit history, self-employed, property area, and applicant income also show significance with loan prediction. This study contributes to the literature regarding loan prediction by providing a global model summarizing the loan prediction determinants of customers' factors.***

***Keywords: Machine learning, Decision tree, Artificial intelligence, Financial service, Loan prediction.***

## I. INTRODUCTION

Machine learning must be a research area of computer science that learns the theory of computer learning and the identification of patterns of artificial intelligence. Machine learning is a change in systems that typically perform tasks related to artificial intelligence (AI). These tasks include recognition, analysis, planning, robot control and prediction. Explore the research and configuration of algorithms that can predict data. Machine learning is used to build programs with tuning parameters to adapt to early data and improve functionality. Machine learning is a technology that grows rapidly and works with the human mind; it represents multi-level records and effectively resolves the selectivity dilemma.

Machine learning is used in many areas, especially in the financial sector; today's banks play an important role in the market economy; loan distribution is a key business part of almost every bank. The main part of bank assets is directly derived from the profits of loans distributed by banks; the main goal of the financial environment is to invest assets in safe places. The success or failure of the organization depends primarily on the industry's ability to assess credit risk; the bank judges whether the borrower is bad or bad before providing credit to the borrower. Predicting the borrower's condition (e.g., the borrower is likely to be a debtor) is a difficult task for an organization or bank; by default, the default prediction of the loan is a binary classification problem. The history of loan amounts, costumer dominates his credit for loan receipts; the problem is to classify borrowers as basic or non-basic.

Many banks/financial companies today approve loans after the regression process of verification and verification, but it is not clear whether the applicant selected is the correct applicant among all applicants. This system allows you to predict whether a particular applicant is safe and machine learning technology automates the entire functional verification process. The downside to this model is that loans can only be approved for one powerful element that emphasizes different weights for each element, but in reality they are impossible in this system.

The purpose of this paper is to find and analyze loan predictions so that banks/financial firms can use this study to construct possible solutions to risk. The methodology of pre-access and modeling techniques used in this white paper may be considered a roadmap to follow the steps taken in this study by the reader and to apply procedures to identify the causes of many other problems. The purpose of this paper is to provide a way to quickly and quickly select qualified applicants; we can offer a special advantage to the bank. The loan prediction system can automatically calculate the weights of each feature participating in the loan processing, and the same feature is processed in connection with the weights associated with the new test data. You can set a time limit to determine whether an applicant can approve a loan. The loan prediction system allows you to move to a specific application to determine the priority.

## II. RELATED STUDY

Amira and Ajith [1] use a prediction model using three different training algorithms to train the supervised two-tier feedforward network. Results show that training algorithms improve the design of loan-based prediction models. This paper shows that the two models above provide optimal results with fewer errors. Ngai et al. [2] uses classification models to predict future behavior of costumers in CRM; the most popular model in CRM domains is neural networks.

Revised Manuscript Received on August 30, 2020.
* Correspondence Author
**Youngkeun Choi***, Division of Business Administration, Sangmyung University. E-mail: penking1@smu.ac.kr
**Jae Won Choi**, Department of Computer Science, University of Texas at Dallas. E-mail: jxc190057@utdallas.edu

Chitra and Uma [3] introduced an ensemble learning method for time series prediction based on the Radial Basis Function Networks (RBF) and KNN (K-Nest Neighbor). Self-configuration map (SOM); they proposed a PAPEM model that was better than individual models.

Akkoç [4] used hybrid adaptive neurofuzzy inference models, statistical groups, and neurofuzzy network models:Tenfold cross-validation is used to compare better results with other models. Sarwesh and Sadna [5] proposed combining two or more classifiers to create an ensemble model for better prediction. They used the bagging and booting techniques and then used any forest technology; they say the proposed method provides better results and accuracy than a single classifier and other models.

## III. METHODOLOGY

### A. Dataset

Variable information is obtained from third-party websites, an international issue on the popular Internet platform Kagle (www.kaggle.com); this website provides data titled 'Abort Prediction' uploaded by Amit Parajapet. In all industries, the insurance sector uses the most analytical and data scientific methods. This data set provides good taste for the insurance company's data set operations, one-on-one problems, strategies used, variables affecting the outcomes, etc. There are problems identifying customer classifications to automate this process, and there are loan amounts specifically targeted to customers. Kagle asked participants to predict heart disease; the organizers provided data stream types for large individual factors to help develop algorithms. These variables are listed and defined in Table 1.

**< Table 1> the variables in each category**

| Variables | Measurement |
|---|---|
| ID | Identification No. |
| Gender | Male or Female |
| Married | Yes or No |
| Dependents | 0, 1, 2, or 3+ |
| Education | Graduate, or Not graduate |
| Self-employed | Yes or No |
| ApplicantIncome | USD |
| LoanAmount | USD |
| Loan_Amount_Term | USD |
| Credit_History | 0 or 1 |
| Property_Areas | Rural, Semiurban, or Urban |
| Loan Status | Y or N |

### B. Decision Tree

Among the various analytical techniques, decision tree (DT) is a powerful and widely used machine learning algorithm for predicting and classifying big data to date. It's used for both classification and regression problems; now you'll wonder why we're more willing to use DT classifiers than other classifiers. We can give two reasons to answer that question. One is that it is very simple to understand data and to make good conclusions or interpretations because decision trees often try to emulate the way the human brain thinks. The second reason is that the decision tree allows you to see the logic that the data interprets, not the black box algorithms such as SVM and NN. It has a simple and clear expertise, and is one of the generation's favorite programmers. Now we've looked at why the decision tree can look closely at what the decision tree classification flag is.

### C. Equations

Performance measurements can be divided into technical performance measurements and heuristic measurements. The technical performance measurements used in this study show performance results by generating models in training data, processing test data as models, and comparing class labels in the original verification case with predicted class labels. Measurements of technical performance can be divided into instructional and biz learning; instructional learning used in this study is classified and returned. All data used for this learning and testing will have the original class value; the original class value is compared with the expected results and analyzed to achieve performance.

Classification is the most common data analysis problem; various metrics have been developed to measure the performance of the classifier model. Category type classification issues often include accuracy, precision, recall, and f measurements. RapidMiner includes: performance (classification) to measure performance metrics for general classification problems; and performance (differentiation) to provide performance metrics for binary classification problems. Table 2 shows how these indicators are calculated.

**<Table 2> Key performance indicators of binomial classification**

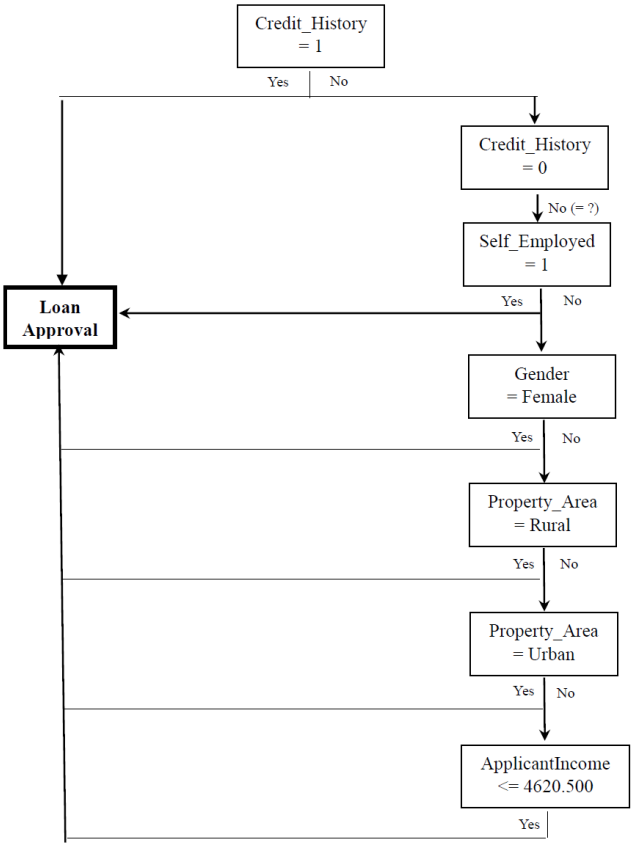
| | | Actual class (as determined by Gold Standard) | |
|---|---|---|---|
| | | True | False |
| Predicted class | Positive | True Positive | False Positive(Type I error) |
| | Negative | False Negative(Type II error) | True Negative |

Precision = TP/(TP+FP), Recall = TP/(TP+FN), True negative rate = TN/(TN+FP), Accuracy = (TP+TN)/(TP+TN+FP+FN), F-measure = 2·((precision·recall)/(precision + recall))

## IV. RESULTS

### A. Decision Tree

Figure 1 shows the classification tree of the entire model after cleaning the tree using cross-validation to avoid overload sums. In the overall model analysis, the main variables are composed of 12 variables according to the criteria set for each variable as follows: Women are more likely to receive loans than men. Credit records, self-employed, real estate and applicant income are also important with loan forecasts.




**<Figure 1> Classification Tree for the Full Model**

Table 3 shows each mixed matrix measurement; in all models, the accuracy is 0.815 and the error rate is 0.185. 79.62 percent of patients who expected no loans were accurate without loans and 92.59 percent were accurate with loans.

**< Table 3> Performance evaluation**

| | True Y | True N | Class precision |
|---|---|---|---|
| Pred. Y | 125 | 32 | 79.62% |
| Pred. N | 2 | 25 | 92.59% |
| Class recall | 98.43% | 43.86% | |

## V. CONCLUSIONS

The purpose of paper is to verity the accuracy of the model and develop a new model that can predict the lending of people in costumes. In summary, this study has essentially two main goals. First, this paper aims to better understand the role of variables in loan prediction modeling. Second, this study attempts to evaluate the predictive performance of decision tree; a series of meanings are derived based on the results reported above. In relation to the first goal, the study suggests that evaluating the role of variables is complex and depends on the classification method in which the impact is used. Decision tree methods emphasize the most important explanation ability for analysis. Therefore, it is impossible for the explanatory variable to draw the most important unanimous conclusion to predict the loan for all the methods used in general. However, the findings provide additional information about customer profiles: banks/financial firms must predict loans on how they are employed. For example, women are more likely to get loans than men first; credit records, self-employed, real estate and applicant income are also important with loan forecasts. Second, the accuracy of all models is 0.815, and the error rate is 0.185. 79.62 percent of patients who expected no loans were accurate without loans and 92.59 percent were accurate with loans.

In fact, this application helps banks/financial companies manage their personal records and make decisions faster if they already have reports with their customers. By default, the prototype is described in a paper that can be used by organizations to make the right decisions to approve or reject customer loan requests. The white paper is also for the management rights of banks/financial companies only, and all prediction processes are personal and stakeholders can be changed. Results for a specific loan ID are sent to various bank departments to take appropriate action against the application; this helps all other departments perform different formats.

The proposed system involves a database that stores customer records, and as the number of customers increases, more data is generated and storage is problematic. This means that future releases will provide cloud functionality that will allow all records to be stored in the cloud. So if you have the privilege to protect and access data, you can search anywhere; smart devices are synchronized with applications in future releases. As a result, we warn the bank/financial company when the customer's real-time financial status is monitored and financial needs are met.

For the future, the machine learning model will use a larger set of training data and more than a million different data points maintained in the electronic financial recording system. While it can be a big leap forward in computing performance and software elaboration, with artificial intelligence and working systems, financial experts can make the best decisions for customers as soon as possible. You can develop software APIs to allow customers to access health websites and apps for free; probabilities predictions are made with little or no processing delays.

## REFERENCES

1. K. Amira, H. Ibrahim and A. Ajith, "Modeling Consumer Loan Default Prediction Using Ensemble Neural Networks," International Conference on Computing, Electrical and Electronics Engineering, 2013, pp. 719 – 724.
2. E. W. T. Ngai, L Xiu and D. C. K. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification," *Expert Systems with Applications*, 2009, Vol. 36, pp. 2592–2602.
3. A. Chitra and S. Uma, "An Ensemble Model of Multiple Classifiers for Time Series Prediction," *International Journal of Computer Theory and Engineering*, 2010, Vol. 2, No. 3, pp. 454–458.
4. S. Akkoç, "An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System ( ANFIS ) model for credit scoring analysis : The case of Turkish credit card data," *Elsevier Europezan Journal of Operational Research*, 2012, Vol. 222, No. 1, pp. 168–178.
5. S. Sarwesh and K. M. Sadhna, "A Review of Ensemble Technique for Improving Majority Voting for Classifier," *International Journal of Advanced Research in Computer Science and Software Engineering*, 2014, Vol. 3, No. 1, pp. 177- 180.

## AUTHORS PROFILE

**Young Keun Choi** Division of Business Administration, School of Business and Economics, Sangmyung University, Seoul, The Republic of Korea. His research areas are business analytics, data science, entrepreneurship, technology management, and etc.

**Jae Won Choi** Department of Computer Science, Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, TX, USA. His research interests are data science, artificial intelligence, blockchain, game, and etc.