IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 8, Issue 1, 2020

Prediction analysis of risky credit using Data mining classification models

Mrs. Lakshmi Lavanya Tumu, Ms. Noore Ilahi, Mr. Guntur Suresh

Abstract—

Customers benefit in a variety of ways from having a high credit score, while banks benefit from being able to evaluate their customers and provide credit properly thanks to this metric. This case In this work, we explore whether data mining approaches can accurately forecast and categories a customer's credit score (good/bad) in order to mitigate the potential future risks associated with lending money to borrowers who may not be able to pay back their loans. Our general models (predictive models) are built using a bank's historical information, and banks may utilize them to improve the results of their credit operations. If a consumer is given a poor credit score by one of these predictive categorization models, for instance, the bank would likely forbid further credit extensions to that person and conduct a thorough evaluation of any other potentially hazardous loans.

I. INTRODUCTION

Credit allows people and businesses to make purchases before they have the financial resources or motivation to pay for them. People in the agricultural, industrial, and commercial sectors may get the financing they need from banks. Markets, businesses, and trades. And when individuals utilize their smarts and entrepreneurial spirit to take advantage of loans, the economy as a whole benefits. Banking institutions face uncertainty when extending loans to consumers in light of the rapid economic growth seen in many nations in recent years.

The same might be said about modern banks; if clients fail to repay credit loans on time, the accumulated loss may amount to an enormous sum, perhaps leading to insolvency. To mitigate this potential for disaster, we have presented a number of data mining methods. Through the use of credit scoring, these models will assist in determining whether or not a customer is likely to be able to make timely payments on a credit loan, categorizing them as either "Good credit" (those with a high credit score and no history of defaulting on loans) or "Bad credit" (those with a low score and potentially problematic loan repayment histories). It will benefit banks financially since they will be able to provide excellent credit, which will increase their earnings each year. The section 2 research section includes Relevant studies. In Section 3, we describe the dataset in detail. Section 4 briefly describes the numerous models used sequentially and their findings, and discusses data analysis via graphical representation,

Outlining important elements determining final conclusions. The last section of the paper serves as a wrap-up.

II. RELATED WORK

The data mining framework in the field of banking and insurance analytics has been the subject of several research discussing relevant topics. Data-driven methods were employed, for instance, by Jin et al. using a 10-fold cross-validation technique and a high value of average percent hit ratio to demonstrate the superior prediction, we examined three data mining models (decision trees, support vector machines, and neural networks) for their ability to forecast loan risk. Quantitative study of the lift curve is performed cumulatively. Best results were achieved using the Support Vector Machine [1].

1,2,3 Assistant Professor
1,2,3 Department of CSE
1,2,3 Global Institute of Engineering and Technology Moinabad, Ranga Reddy District, Telangana State.

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 8, Issue 1, 2020

In order to help insurance company's better forecast loan applicants, Wang et al. suggested a mining model that included a rule generator and a recommendation mechanism. They were successful in part [3] due to the fact that policyholders having access to a more attractive interest rate were more inclined to apply for a loan.

Using a dataset of 20 variables from German bank credits, Hassan et al. developed supervised neural network models for loan prediction. Models were tested, and results were compiled for:

Proportions of accuracy [4] that have been calculated for each. Based on the data provided by the Iranian bank, Jafarpour et al. honed down on customer relationship management. Using data collected from a variety of sources, a customer relationship management model (CRM) creates a formula for estimating future loan clients that may be used by financial institutions [5].

Hsu et al. [6] used a support vector machine (SVM) to classify a bank credit dataset, and they found that SVM performance improves with additional data samples or the addition of other selection factors, making it a better tool for credit rating. Applying both supervised and unsupervised machine learning algorithms to a bank credit dataset, Turk son et al. [7] found an accuracy of up to 80% in their credit score predictions.

Moro et al. looked at data from Portuguese retail banks to see how well neural network models might predict the performance of telemarketing campaigns [8].

III. DATA DESCRIPTION

We retrieved the dataset from the UCI machine learning data repository [10]. A. Data Origin As presented by Hans Hofmann. This Set of Numbers includes fields for Credit, Balance credit acc, Duration, Rate, Age, Profession, and more.

3. B. A Description of the Data

There are 1000 different cases and 18 different parameters in this dataset.

The dataset was split 70%-30%, with 700 records in the Training Dataset and 300 records in the Validation Dataset, respectively. The variables in each instance are given decimal values (0, 1, 2, 3...) for evaluation.

IV. EXPERIMENTS AND RESULTS

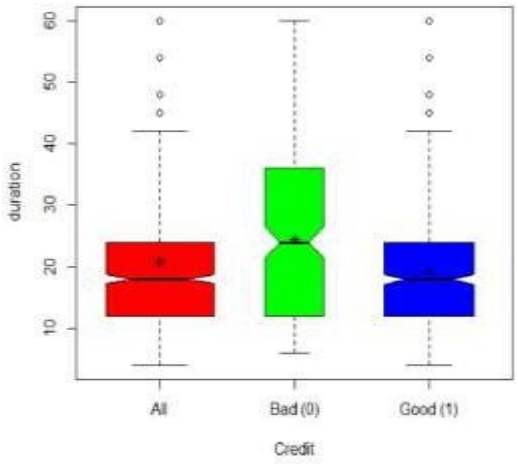
A. Major Factors

In this section, we will detail the actions of all the components that influence the ultimate outcome; in other words, we will outline the primary elements that contribute to the probability of default. (Poor credit) by means of visual depictions of these characteristics and features from our training dataset. After examining the influence of each variable on our dependent variable, we determined that just three factors had a negative effect on the credit, all of which have numeric values that are different from the others. The other variables all have category values that are repetitious.

TABLE I. MAJOR FACTORS AFFECTING CREDIT

S.No	Factors	Description
1	Duration	Number of Months to Repay Credit
2	hoehe	Amount credited in deutsche mark
3	alter	age of person taking credit

We'll go through each factor in detail and explain how it's influencing our dependent variable of interest. To begin, we will analyze the time period in which t the following is a box plot of the training dataset demonstrating the influence of duration on credit for which credit was granted.



Graph 1: Credit-based Time Distribution

Box and whiskers are shown above in a multicolored fashion for each box. The box's median (50th percentile) is shown by its centre higher and lower bands represent the 75th and 25th percentiles (the upper and lower quartiles, respectively) of each box's distribution. The average of all the values in that box or category is shown by the asterisk. All the entries in

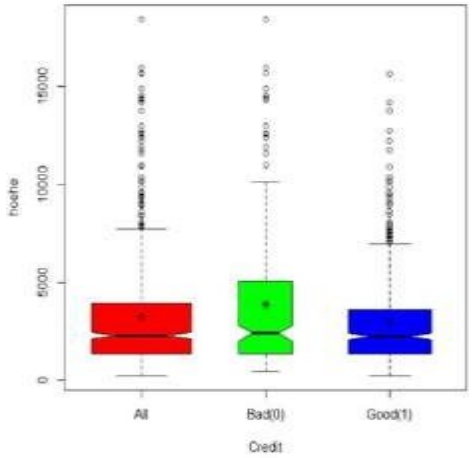
International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 8, Issue 1, 2020

the dataset where the mean is determined to be between 18 and 20 months are shown by the red box. The mean is quite close to 20. Accordingly, we may deduce that the typical duration of a credit loan is between 18 and 20 months. The good credit distribution (1) is shown in blue, and it again closely matches the inference made from the red box, namely that applicants whose credit applications are approved have taken out loans for 18-20 months that they have been able to repay on time. The green box, however, demonstrates that the risk of credit default (poor credit) grows with length since its median and mean are larger than those of the other two boxes. Therefore, we may conclude that the length plays a crucial role in determining whether or not a credit goes into default.

Second, there will be certain effects due to the factor amount ('deutsche mark') as shown below:



Here we see how the hoe he (Amount) is split up across the several credit although there is not a large discrepancy in the median (Amount) of the boxes, the mean for the green box (poor) greater credit) or variation. While the actual amount doesn't make much of a difference to credit, the graph nonetheless implies that poor credit is tied to a certain dollar number. The poor credit shown as a percentage of the total is substantially greater than average since it fluctuates. At last, we used the rattling [2] tool to analyze the age variable, and we discovered that a person's age has a significant effect on their credit.

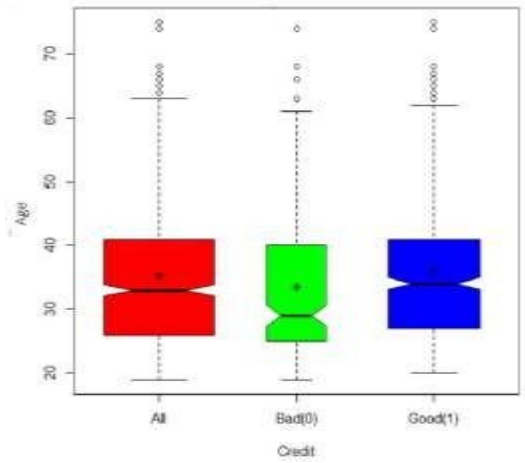


Figure 3: Credit Alter (Age) Distribution

The majority of applicants are between the ages of 33 and 35, as seen above by the red box. Years. Credit scores averaged approximately 35 for those who didn't fall into default, suggesting that individuals over the age of 30 who made timely payments were more likely to have excellent credit. If we compare this to the green box (poor credit), we can see that candidates younger than 30 are more likely to be defaulters. Therefore, it is clear from this graph that millennial have a far higher default rate than their elder counterparts.

Use of Predictive Models

Here, we provide a brief description of the several data mining categorization models we used in our experiment, along with the outcomes we observed while applying each model in turn.

The models' respective accuracy graphs are shown.

In all, the following models were considered:

A. Decision Tree Model
B. Random Forest Model
C. Adaptive Boosting Model
D. Support Vector Machine Model
E. Linear Regression model
F. Neural Network Model

C. Decision Tree Model

One of the most well-known techniques for mining data is the decision tree. A recursive method of allocation is used in this computation.

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 8, Issue 1, 2020

Typically used because to its straightforward interpretation, a decision tree is the gold standard of information mining tools. It consists of a single root node that is divided into two branches by another variable. So, the two new branches are now nodes that may further divide on their own distinct variable. This continues until further splitting the model into smaller pieces won't improve its runtime. Decision trees may be used with either numerical or categorized information.

A decision tree based on our data, constructed with the help of the Rattle tool [2] is shown below.

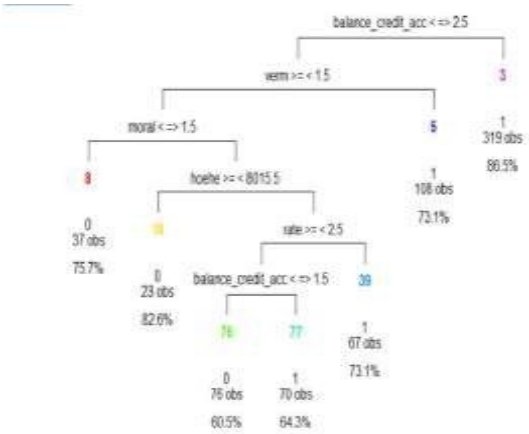


Fig. 4. Decision Tree

D. Random Forest model

To do its analysis, the Random Forest method constructs a forest of decision trees, each one based on a distinct subset of the dataset. At each node in the tree, factors are taken into account in order to split the data. In the case of classification, a simple majority vote is then used for prediction (the average is used in the case of regression). When it comes to resisting over fitting, Random Forests shine. To combine unprimed decision trees into one model, Random Forest uses an ensemble approach. Commonly used for dealing with large datasets and an unusually large number of information components Random Forest (hundreds or even a thousands of info factors).

As a result of repeatedly subletting the available factors, the method works well for a large variety of them. Standard implementations of Random Forest models include tens or hundreds of decision trees. A large forest was employed (500 trees) to collect our data.

E. Adaptive Boosting Model

Boasting's primary goal is to assign a weight to each data point in the dataset. If a model performs poorly, the weights are increased across the board to compensate. Organizes the data in a meaningful way. With the use of boosting, many models are combined to make a single decision in a binary classification task. Single-branch decision trees (decision stumps) may be used as models. When a model is created, any training items that it incorrectly categorizes are "boosted," or given more importance, before the next model is created. The ensemble of models constructed is then summed and given weighted importance to arrive at the final model.

F. Support Vector Machine Model

When analyzing data, a Support Vector Machine (SVM) seeks for support vectors, or items of data, that are located at the limit of a certain area in space. One set of data and moving on to another. In a data classification system, the margin between classes is the empty space between regions that store data items belonging to different classes. An isolating hyper plane (a line in multi-dimensional data or a plane in two-dimensional data) is found with the use of support vector machines. Our dataset made use of 443 support vector machines.

G. Linear Regression Model

When trying to fit a statistical model to data, the standard approach is to use a linear regression model. When the dependent variable of interest is a numerical continuous, this method is suitable. Hypothesis testing with linear regression models fit iteratively to the data when the target variable is reconstructed to a continuous numerical form.

The target variable's distribution and a link function that maps the target mean to the inputs serve as parameters for the extended method. Typically, we use these two characteristics to define a family, which may include distributions like the Poisson, Logistic, etc.

Using a logistic or profit function, the goal is rebuilt if there are only two alternative outcomes. The coefficients in a profit regression are often less than those in a logistic regression, but the outcomes are comparable.

H. Neural Network Model

A paradigm whereby interconnected neurons in various layers process and output numerical data inside the network as a whole. A use of neural networks in modeling dates back many decades. The underlying structure of the model is based on the same principles as the human nervous system. Rather of producing electrical signals, the network of neurons and synapses would produce numerical values.

Initial Graphed Outcomes

Now, we can see how well each model performs by comparing the predicted and actual values in the validation dataset on accuracy graphs (Pr vs., Ob charts).

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 8, Issue 1, 2020

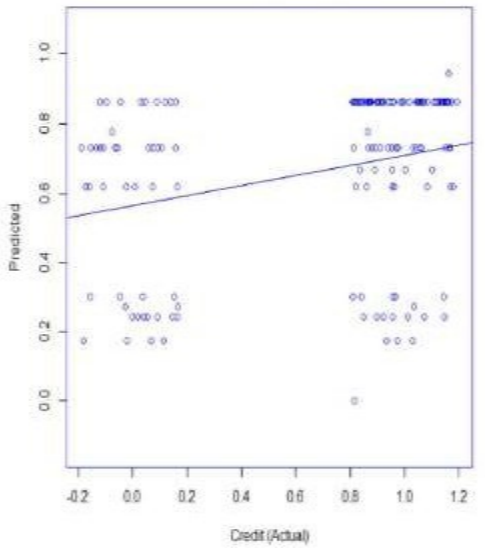


Fig. 5. Pr vs. Ob Graph for Decision Tree

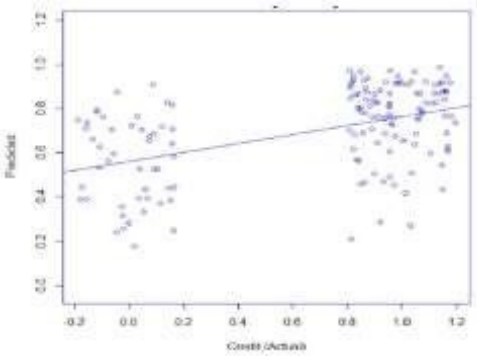


Fig. 6. Pr vs. Ob Graph for Random Forest

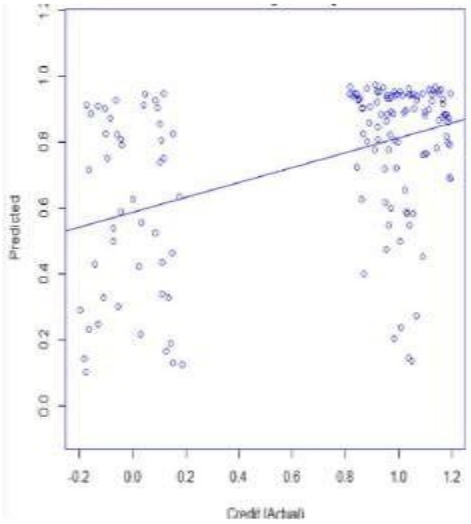


Fig. 7. Pr vs. Ob Graph for Adaptive Boosting

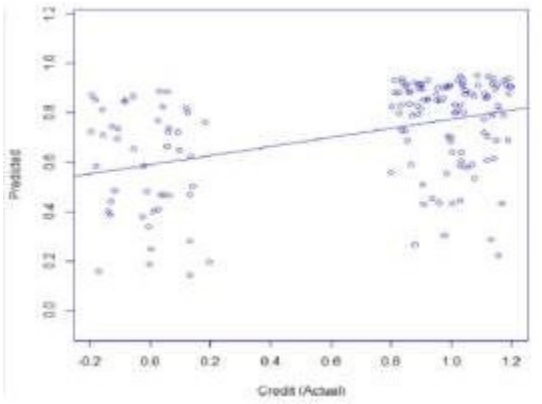


Fig. 8. Pr vs. Ob Graph for Support Vector Machine

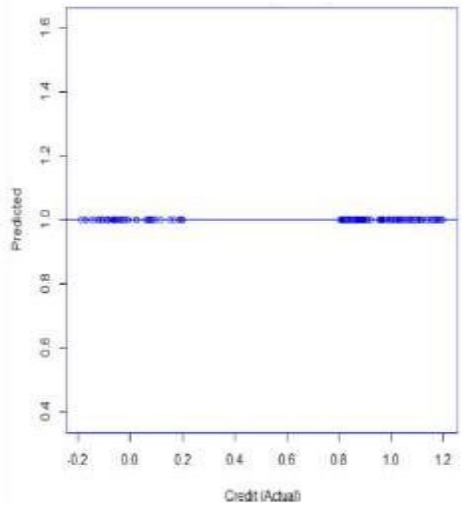


Fig. 9 Pr vs. Ob Graph for Neural Network

The graphs make it clear that the Decision Tree model and the neural network are not very good at predicting the values.

V. RESULTS

Here, we undertake a comprehensive model comparison using the validation and training dataset's ROC curves.

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 8, Issue 1, 2020

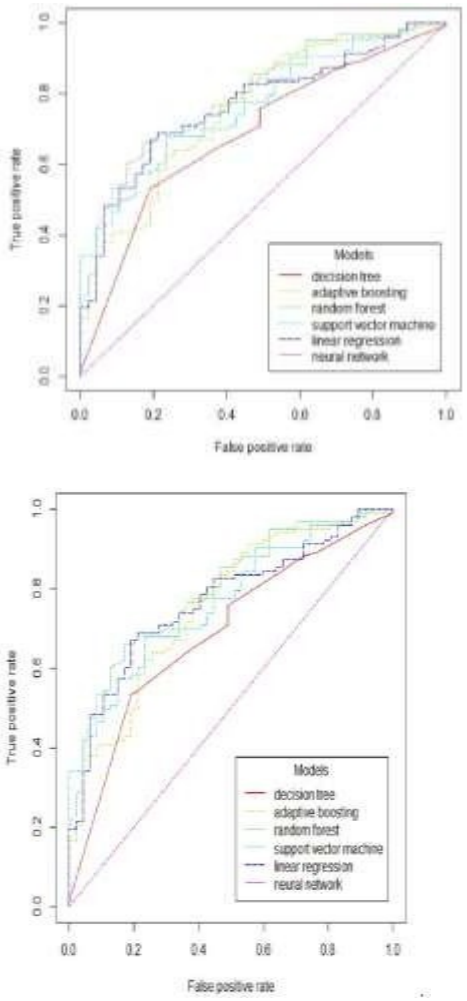


Fig. 12. ROC curve for Training dataset

Dataset, B. To be used for testing the graph shows that the diminishing areas under the curve correspond to the following:

The order of importance for these methods is as follows: Random Forest > Adaptive Boosting > Support Vector Machine > Linear Regression > Decision Tree > Neural Network. In light of this, Random forest is the most promising approach for constructing a strong predictive classification model, whereas the neural network is unable to do so for any of the datasets.

All of our dataset's values may be seen in the table below.

Models Used (Algorithms)	Area under the ROC curve (values)		Time required (secs) To build the model	Rank dependin g on Predictive Power
	Validation Dataset	Training Dataset		
Random Forest	0.7968	1.0000	1.33	1
Linear Regression	0.7686	0.7978	0.12	2
Support Vector M.	0.7672	0.8891	0.28	3
Adaptive Boosting	0.7569	0.9668	2.08	4
Decision Tree	0.6885	0.7702	0.06	5
Neural Network	0.5000	0.5000	0.04	6

TABLE II. PERFORMANCE EVALUATION OF MODELS

Although low-rank algorithms are being constructed more quickly, adaptive boosting is slow since its area under the roc curve for the training dataset is larger. Poor results as compared to the validation dataset. As a result, its ability to foretell the future is thought to be weak.

As a result, Random forest and linear regression are the best and most practical algorithms for categorical data categorization.

VI. CONCLUSION

In this study, we describe a technique for predicting a borrower's creditworthiness that might aid financial institutions in deciding whether to provide a loan. Client loan application considering his or her history, profession, marital situation, and other personal details. We found that age, length, and quantity were the most significant determinants in determining a person's financial well-being. The banking sector may benefit from our suggested analytical work when making credit choices for customers. In order to forecast and categories the application of loan as good or poor, this study uses Algorithms such Decision Tree, Support Vector Machine, Adaptive Boosting Model, Linear Regression, Random Forest, and Neural Network to develop predictive models. Rattle [2] is used to implement the models. We used classification data mining methods to determine that the Random Forest algorithm is superior to other candidates for classifying potentially problematic credit.

ISSN 2347–3657

Volume 8, Issue 1, 2020

REFERENCES

[1] Yu Jin and Yuan Zhu, "A data-driven approach to predict default risk of loan for online Peer-to-Peer (P2P)
Lending," School of Information, Zhejiang University of Finance and Economics, 310018 Hangzhou, China.
[2] Chen-Shu Wang and Yeu-Ruey Tzeng, "Prediction Model for Policy Loans of Insurance Company,"
Department Management Information Systems, National Cheng-Chi University.
[3] Williams, Graham J. "Rattle: a data mining GUI for R." The R Journal 1.2 (2009): 45-55.
[4] Amiga Kamala Ibrahim Hassan and Amity Abraham, "Modeling Consumer Loan Default Prediction Using Ensemble Neural Networks," Department of computer science Sudan University of Science and Technology,
Sudan Khartoum, Sudan, and Machine Intelligence Research Labs (MIR Labs), WA, USA IT4Innovations, VSB - Technical University of Ostrava, Czech Republic
[5] H.Jafarpour and H. Sheikholeslami Garvandani, "New Model of Customer Relationship Management in Iranian Banks," icbme.yasar.edu.tr, pp. 1–12, 2012.
[6] C. F. Hsu and H. F. Hung, "Classification Methods of Credit Rating - A Comparative Analysis on SVM, MDA
And RST," 2009 International Conference on Computational Intelligence and Software Engineering, pp. 1–4, Dec. 2009.
[7] R. E. Turk son, E. Y. Baggier and G. E. Kenya, "A machine learning approach for predicting bank credit
Worthiness," 2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR), Lodz, 2016, pp. 1-7.doi: 10.1109/ICAIPR.2016.7585216
[8] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," Decision
Support Systems, vol. 62, pp. 22–31, 2014.
[9] K. H. Kim, C. S. Lee, S. M. Jo and S. B. Cho, "Predicting the success of bank telemarketing using deep convolutional neural network," 2015 7th International Conference of Soft Computing and Pattern Recognition (Oscar), Fukuoka, 2015, pp. 314-317.
[10] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.