e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:02/Issue:05/May-2020 www.irjmets.com

# BANK LOAN APPROVAL PREDICTION USING DATA MINING TECHNIQUE

**Varun S Kumar [*1], Abhishek Rokade [*2], Srinath MS[*3]**

[*1,2,3] Department of Industrial Engineering Management, R V College of Engineering, Bengaluru, Karnataka, India.

## ABSTRACT

Machine Learning is a developing method for building scientific models for machines to "learn" from information and have the option to do prescient investigation. The capacity of machines to "learn" and do prescient examination is significant in this time of huge information and it has a wide scope of use regions. For example, banks and money related foundations are once in a while confronted with the test of what hazard variables to consider when propelling credit/advances to clients. For a few highlights/properties of the clients are regularly mulled over, however a large portion of these highlights have minimal prescient impact on the credit value or in any case of the client. Besides, a vigorous and powerful mechanized bank credit hazard score that can help in the forecast of client credit value precisely is as yet a significant test confronting numerous banks. In this paper, we inspect a genuine bank credit information and direct a few AI algorithms on the information for near investigation and to pick which algorithms are the best fit for learning bank credit information. The calculations gave over 80% precision in forecast. Besides, the most significant highlights that decide if a client will default or in any case in paying his/her credit the following month are separated from an aggregate everything being equal. We at that point applied these most significant highlights on some chose AI calculations and contrast their prescient precision and different calculations that pre-owned all the highlights. The outcomes show no noteworthy distinction, implying that these highlights can precisely decide the credit value of the clients. At long last, we detail a prescient model utilizing the most significant highlights to foresee the credit value of a given client.

**KEYWORDS:** Bank Credit, Classification, Confusion Matrix, Predictive Analysis.

## I. INTRODUCTION

The growing volumes, varieties and velocity of data due to the emergence of the Internet in particular and the cheaper data sharing and storage facilities coupled with the cheaper but more powerful computational tools have opened a new frontier in the field of data science. And thus, there is currently an active ongoing research within the fields of data mining (discovering patterns in data) and machine learnings (building analytical models using algorithms for machine to “learn” from data), both aim at using algorithms and concepts to extract knowledge and pattern from data. One of the major reasons for establishing banks is to advance loans to customers. But in order to stay in business, banks advance these loans to people who have the ability to pay back the money, thereby minimizing the risk of the non payments of loans. However, risk management; knowing who is credit worthy is still an on-going challenge within the banking sector. The ability to identify a risk score of a customer base on some features such as occupation, age, marital status, salary range/amount of equity, credit history, etc. is an important step that banks go through before giving credit to customers. For the credit risk score helps the banks to decide on how much interest to charge on the loan, etc. However, these risk factors sometimes do not give an inform decision on the credit worthiness of customers. Moreover, many banks lack a central well integrated, automated finance and risk management system due to the inability to develop a robust and scalable risk management system to forecast risk score of customers.

Another nightmare faced by many banks these days is frauds. And the machine learning approach is seen and considered as the right tool that can be leveraged on in order to understand the banking transaction pattern of customers, by identifying pattern in customer data, so as to be able to distinguish between fraudulent activity from that of a normal one. Therefore, we leveraged it on the bank credit dataset in order to understand the key factors that influence the payment of bank loans. The dataset is obtained from the UCL machine repository. We perform analysis and applied machine learning algorithms on the bank credit data, firstly to understand the nature of the data and the best algorithms suitable for learning bank credit data. We formulated a predictive

model to determine the credit worthiness or otherwise of a given bank customer using a linear regression method.

The rest of the paper is organized into sections as follows: Section II gives the background information on the machine learning algorithms used in studying on the bank's credit dataset. In Section III, we outlined our methodology, in Section IV the observation and result is outlined. The paper is concluded in Section V.

## II. BACKGROUND

The most important background information on machine learning algorithms and their theoretical formulation are outlined in this section. These algorithms are used in analyzing the bank credit data.

### A. Machine Learning Algorithms

Machine learning techniques can be grouped broadly into two main categories. They include:

(i) Supervised Learning: The main feature of this algorithm consists of target or outcome variable (or dependent variable). The target variable is used to predict other features from a given set of predictors (independent variables). Furthermore, using the target variable, a function is generated that maps input to desired outputs. The training process then continues until the model achieves the desired level of accuracy on the training data. Supervised learning techniques are achieved using regression and classification algorithms or approaches that range from non-linear regression, generalized linear regression, discriminant analysis, Support Vector Machines (SVMs) to decision trees and ensemble methods.

(ii) Unsupervised Learning: In unsupervised learning, there is no target or outcome variable to predict or estimate. This algorithm is used mainly for segmenting or clustering entities in different groups for specific intervention

The dataset used in this paper is a labeled data and is, therefore, suitable for doing classification analysis. And thus, we employed various classification algorithms described comprehensively in Section II-B. Some of the algorithms are implemented in Python scikit-learn package to predict the creditworthiness of bank customers with regards to their ability to pay their credit or otherwise within a given time frame.

### B. Classification Algorithms

Classification algorithms work by predicting the best group to which a data point belongs to by "learning" from labeled observations. It uses a set of input features for the "learning" process. Classification algorithms are good for grouping data that are never seen before into their various groupings and are therefore extensively used in machine learning tasks. Some of the well-known classification algorithms used in this paper are briefly discussed below:

1) Decision Trees: There are two kinds of decision trees; classification trees and regression trees. A decision tree can be described as a flow-chart like structure in which internal node represents test on an attribute, each branch represents outcome of the test and each leaf node represent decision taken after computing all attributes or a response after computing all given attributes

2) Naive Bayes: This classification technique is based on Bayes' theorem that assumes independence between predictors, thus, the presence of a particular feature in a class is independence of another feature in a another class. Naive Bayes classification is therefore, based on estimating $P(X|Y)$, the probability or probability density of features X given class Y.

3) Linear Regression: It is used to estimate real values based on continuous variable(s). In linear regression, a relationship is established between independent and dependent variables by fitting the best line. This best fit line is known as regression line and represented by a linear equation $Y = a * X + b$, where Y is the dependent variable, a is the slope, X is the independent variable and b is the intercept. The coefficients a and b are derived based on minimizing the sum of squared difference of distance between data points and regression line.

## III. METHODOLOGY

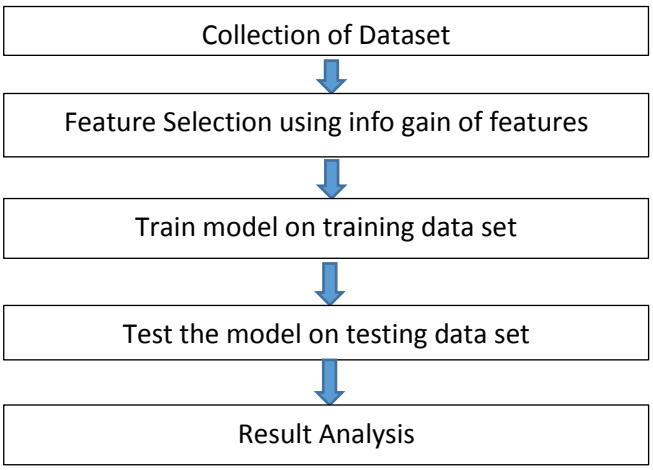


**Fig-1**

Data preprocessing is an information mining strategy that includes changing crude information into a reasonable organization. True information is regularly inadequate, conflicting, as well as ailing in specific practices or drifts, and is probably going to contain numerous blunders. Information preprocessing is a demonstrated technique for settling such issues. Data preprocessing gets ready crude information for additional handling.. Data preprocessing has the following steps

### A. Data Collection

Data collection is the way toward social occasion and estimating data on focused factors in a built up framework, which at that point empowers one to address applicable inquiries and assess results. Information assortment is a part of research in all fields of study including physical and social sciences, humanities and business. While strategies fluctuate by discipline, the accentuation on guaranteeing exact and genuine collection of data continues as before. The objective for all information assortment is to catch quality proof that permits investigation to prompt the definition of persuading and solid responses to the inquiries that have been presented

### B. Data analyzing and data cleaning

Data cleansing or information cleaning is the way toward recognizing and amending degenerate or off base records from a record set, table, or database and alludes to distinguishing deficient, erroneous, mistaken or unimportant pieces of the information and afterward supplanting, changing, or erasing the grimy or coarse information.

### C. Missing Data handling

Missing data are defined as values that are not available and that would be meaningful if they are observed. Missing data can be anything from missing sequence, incomplete feature, files missing, information incomplete, data entry error etc. In this attribution strategy objective is to supplant missing information with factual appraisals of the missing qualities. Mean, Median or Mode can be utilized as ascription esteem. The hypothetical foundation of the mean replacement is that the mean is a sensible gauge for an arbitrarily chosen perception from an ordinary dispersion.

### D. Categorical Data handling

Managing numeric information is frequently simpler than absolute information given that we don't need to manage extra complexities of the semantics relating to every class an incentive in any information trait which is of a straight out sort. Any information trait which is downright in nature speaks to discrete qualities which have a place with a particular limited arrangement of classifications or classes. These are likewise regularly known as classes or names with regards to characteristics or factors which are to be anticipated by a model. These discrete

qualities can be content or numeric in nature. There are two significant classes of categorical data, nominal and ordinal

Nominal attributes comprise of discrete clear cut qualities with no thought or feeling of request among them. The thought here is to change these characteristics into a progressively agent numerical organization which can be effectively comprehended by downstream code and pipelines.

**E. Train the algorithm**

The various classification algorithms are trained using a different set of data. The dataset is further split into 70% for training and 30% for testing the algorithms. Furthermore, in order to obtain a representative sample, each class in the full dataset is represented in about the right proportion in both the training and testing datasets.

**F. Test the algorithm**

The 3 algorithms are used to predict the effectiveness of the algorithm on the test dataset. A cross-validation was done on the algorithms to estimate how accurately these algorithms will perform in practice on a different set of data and their corresponding confusion matrices and classification accuracies measured. In evaluating the performance of the classification algorithms, we adopt the commonly used metrics in the literature. They include accuracy, precision, recall, specificity. These values are calculated using the Python scikit-learn tool with input values as the entities of the confusion matrix. In this paper, a 'negative' instance refers to no(signifying there will be a default in the payment of the loan) whereas the 'positive' instance refers to yes (signifying there will not be a default in the payment of the loan).

## IV. OBSERVATION AND RESULT

**A. Performance of the algorithms**

In this study, three classifier models based on naive bayes, Decision tree and logistic regression are developed. To evaluate these models, 70% of the dataset is used for training while 30% is set aside for validating and testing. Accuracy is used to evaluate the performance of the three classifiers. On the basis of this train data set, system analyze rest of 30 percent data and predict the results in term of loan status either accepted or rejected. Results with loan status by applying the logistic regression, decision tree and Naïve bayes are show in the Figure below.

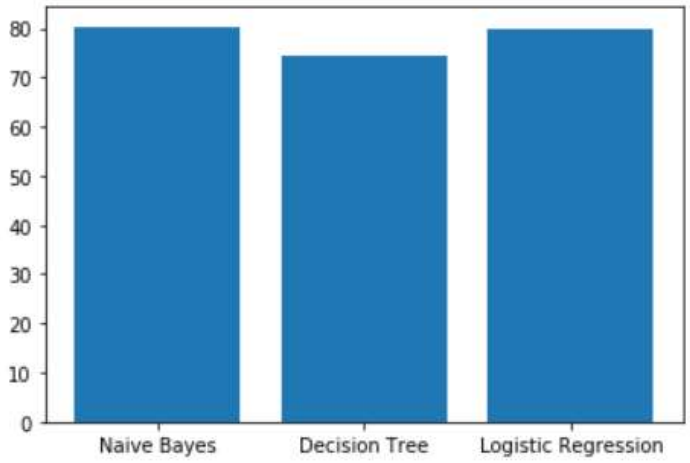


**Fig-2**

This model applied machine learning in prediction of loan approval. Three Machine Learning algorithms are used to predict the loan approval status of customers for bank loans. The results shown that the prediction accuracy is 74%, 80% and 79% for Decision Tree, Naïve Bayes algorithm and Logistic Regression algorithms respectively. Among three the accuracy of Naïve Bayes algorithm is best for prediction of loans. In future the Naïve Bayes algorithm can be applied on other data sets available for loan approvals to further

investigate its accuracy. A rigorous analysis of other machine learning algorithms other than these three can also be done in future to investigate the power of machine learning algorithms for loan approval prediction.

### B. Discussion of Results

In this paper, we employed machine learning approaches to study on the bank credit dataset. The algorithms performed relatively well. Apart from the Decision tree, the rest performed credibly well on the dataset with a prediction accuracy ranging between 74% and 81%. Thus, these algorithms are very suitable for bank data analytics and prediction of credit non-defaulters in particular.

This model can further be improved with the addition of more algorithms into it. However, the output of these algorithms needs to be in the same format as the others. Once that condition is satisfied, the modules are easy to add as done in the code. This provides a great degree of modularity and versatility to the project.

More room for improvement can be found in the dataset. As demonstrated before, the precision of the algorithms increases when the size of dataset is increased. Hence, more data will surely make the model more accurate in detecting frauds and reduce the number of false positives. However, this requires official support from the banks themselves.

## V. CONCLUSION

In this paper, we applied machine learning approach to study bank credit dataset in order to predict customers' credit worthiness. We employed different machine learning algorithms on the dataset in order to determine which algorithms are the best fit for studying bank credit dataset. The experiment revealed that, apart from the Gaussian Naive Bayes, the rest of the algorithms perform credibly well in term of their accuracy and other performance evaluation metrics. Each of these algorithms achieved an accuracy rate between 74% to over 8%. We also determined the most important features that influence the credit worthiness of customers. These most important features are then used on some selected algorithms and their performance accuracy compared with the instance of using all the features. The experimental results showed no significance difference in their predictive accuracy and other metrics. These findings have a lot of implications. The model can be used as a tool to advise banks as which factors are important in determining the credit worthiness of customers. Furthermore, the result showed which machine learning algorithms are not suitable for studying bank credit dataset. We intend to develop a hybrid machine learning system that will incorporate the most important features that determine credit worthiness of customers in order to formulate banks' risk automated system.

## VI. REFERENCES

[1] I. H. Witten and E. Frank, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.

[2] Anuruddha Thennakoon, Chee Bhagyani , Sasitha Premadasa, Shalitha Mihiranga, Nuwan Kuruwitaarachchi,: Real-time Credit Card Fraud Detection Using Machine Learning

[3] Okokpujie Kennedy O, F. Olajide,Covenant University, Ota, Ogun State. Nigeria. : Realtime fraud detection in the banking sector using data mining techniques/algorithm

[4] John O. Awoyemi, Adebayo O. Adetunmbi, Samuel A. Oluwadare ,

Federal University of Technology Akure Akure, Nigeria : Credit card fraud detection using Machine Learning Techniques

[5] Yashvi Jain, NamrataTiwari, ShripriyaDubey,Sarika Jain: A Comparative Analysis of Various Credit Card Fraud Detection Techniques

[6] Anand Motwani, Goldi Bajaj, Sushila Mohane: Predictive Modelling for Credit Risk Detection using Ensemble Method

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E.

Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[8] J. Li, H. Wei, and W. Hao, "Weight-selected attribute bagging for credit scoring," Mathematical Problems in Engineering, vol. 2013, 2013.

[9] 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.