Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 8, August-2020

Predict Loan Approval in Banking System Machine Learning Approach for Cooperative Banks Loan Approval

Amruta S. Aphale
Department of Computer Science and Engineering
Savitribai Phule Pune University
Vishwakarma Institute of Technology, Pune

Prof. Dr. Sandeep. R. Shinde
Department of Computer Science and Engineering
Savitribai Phule Pune University
Vishwakarma Institute of Technology, Pune

Abstract - In today's world, taking loans from financial institutions has become a very common phenomenon. Everyday a large number of people make application for loans, for a variety of purposes. But all these applicants are not reliable and everyone cannot be approved. Every year, we read about a number of cases where people do not repay bulk of the loan amount to the banks due to which they suffers huge losses. The risk associated with making a decision on loan approval is immense. So the idea of this project is to gather loan data from multiple data sources and use various machine learning algorithms on this data to extract important information. This model can be used by the organizations in making the right decision to approve or reject the loan request of the customers. In this paper, we examine a real bank credit data and conduct several machine learning algorithms on the data for that determine credit worthiness of customers in order to formulate bank risk automated system.

Keywords— Machine learning, bank credit, classification, confusion matrix, predictive analysis.

I. INTRODUCTION

Bank plays a vital role in market economy. The success or failure of organization largely depends on the industry's ability to evaluate credit risk. Before giving the credit loan to borrowers, bank decides whether the borrower is bad (defaulter) or good (non defaulter).The prediction of borrower status i.e. in future borrower will be defaulter or non defaulter is a challenging task for any organization or bank. Basically the loan defaulter prediction is a binary classification problem Loan amount; costumer's history governs his credit ability for receiving loan. The problem is to classify borrower as defaulter or non defaulter. However developing such a model is a very challenging task due to increasing in demands for loans. Prototypes of the model which can be used by the organizations for making the correct or right decision for approve or reject the request for loan of the customers. This work includes the construction of an ensemble model by combining different machine learning models. Banks struggle a lot to get an upper hand over each other to enhance overall business due to tight competition. Credit Risk assessment is a crucial issue faced by Banks nowadays which helps them to evaluate if a loan applicant can be a defaulter at a later stage so that they can go ahead and grant the loan or not. This helps the banks to minimize the possible losses and can increase the volume of credits.

II. BACKGROUND

The most important background information on machine learning algorithms and their theoretical formulation are out- lined 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 pre- dictors (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. Examples of unsupervised learning algorithms include Apriori and K-means algorithms.

The various machine learning approaches and the algorithms that describe them are shown in Fig. 1

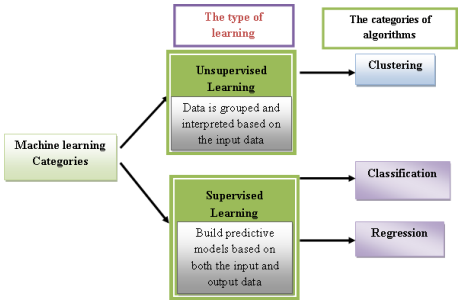

Fig. 1. Machine learning Tasks

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 8, August-2020

Labeled data is known in the literature to be suitable for classification algorithms. 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 MATLAB [§] and some taken from the *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) Neural Networks: The neural network supports both classification and regression algorithms and therefore, is very appropriate for studying the classification problem in this paper.
2) Discriminant Analysis: The discriminant analysis is based on the assumption that different classes of data are generated by using different Gaussian distributions. The main types of discriminant algorithms used for classification are the linear and the quadratic discriminant. We used the quadratic discriminant classifier in this paper.
3) 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 $P(X \mid Y)$, the probability or probability density of features X given class Y.
4) *K*-Nearest Neighbor: The KNN algorithm is used for both classification and regression problems. However, the KNN is more widely used in classification problems in the industry and thus will be used in doing classification and predictive analysis in this paper. The KNN is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. The case being assigned to the class is most common amongst its K nearest neigh- bors measured by a distance function. The common distance functions used are the *Euclidean*, *Manhattan*,*Minkowski* and *Hamming distance*.
5) 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.
6) Ensemble Learning/method: An example of ensemble learning method is the TreeBagger, where the bagging stands for bootstrap aggregation. Every tree in the ensemble is grown on an independently drawn sample of input data. To compute the prediction for the ensemble of trees, TreeBagger takes an average of predictions from individual trees (for regression) or takes votes from individual trees (for classification). Ensemble techniques such as bagging combine many weak learners to produce a strong learner.
7) Decision Trees: There are two kinds of decision trees; classification trees and regression trees. A deci- sion 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.

III. RELATED WORK

The related work n the application of machine learning and data approaches to study financial data are comprehensively described below. Li *et al.* [6] conducted research on using attributes of customers to assess credit risk by using a weighted-selected attribute bagging method. They benchmarked their result experimentally by using two credit databases and reported outstanding performance both in term of prediction accuracy and stability as compared with another state of the art methods. A data mining approach is also pro- posed by Moro *et al.* [7] to predict the success or otherwise of a Portuguese retail bank in telemarketing. They applied various data mining models on the bank telemarketing data and reported that the neural network data mining method was the best for analyzing the data. The role of machine learning techniques in business data mining is outlined by [8]. Their work described the strengths and weaknesses of various machine learning techniques within the context of business data mining approach. Their analysis revealed that Rule Induction Technique was the best approach in mining business data, followed by that of the neural network approach. C. Tsai and M. Chen [9] used a hybrid machine learning approach to study credit rating by comparing four different types of hybrid machine learning techniques. They showed experimentally that 'classification + classification' hybrid model based on a combination of logistic regression and neural networks provides the highest prediction accuracy and also maximize the profit. Bank default data was used by [10] to model bank failure predictions using neural network approach. They compared their result with other machine learning approaches and concluded that the neural network approach is a promising method in terms of pre- dictive accuracy, adaptability, and robustness. [11] proposed. They experiment with the hybrid recommendation algorithms on two sets of data and reported high scalability and better performance in terms of accuracy and coverage. A hybrid online sequential extreme learning machine with the simplified hidden layer is proposed by [12]. The algorithm is a combination of the Online Sequential Extreme Learning Machine and the Minimal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 8, August-2020

Resource Allocation Network. Their experimental results showed that the algorithm has a comparable performance as that of the original online sequential extreme learning machine but with a reduced number of hidden layers.

IV.METHODOLOGY

The proposed model focuses on predicting the credibility of customers for loan repayment by analyzing their behavior. The input to the model is the customer behavior collected. On the output from the classifier, decision on whether to approve or reject the customer request can be made. Using different data analytics tools loan prediction and there severity can be forecasted. In this process it is required to train the data using different algorithms and then compare user data with trained data to predict the nature of loan. To extract patterns from a common loan approved dataset, and then build a model based on these extracted patterns. The training data set is now supplied to machine learning model; on the basis of this data set the model is trained. Every new applicant details filled at the time of application form acts as a test data set. After the operation of testing, model predict whether the new applicant is a fit case for approval of the loan or not based upon the inference it conclude on the basis of the training data sets. To extract important information and predict if a customer would be able to repay his loan or not.

TABLE I
THE COMPOSITION OF THE DATASET
Bank Credit Data

Value	Count	Percentage
No	23364	77.88%
Yes	6636	22.12%
Training Dataset		
No	14092	78.29%
Yes	3908	21.71%
Test dataset		
No	9272	77.27%
Yes	2728	22.73%

If T and N denotes the number of clients that will not default the credit payment and clients that will default in the payment of their credit respectively, then the total number of the dataset is expressed as $T + N$. Furthermore, *TP* (True Positive) and *TN* (True Negative) represent the total positive and negative cases/instances that are rightly classified, respectively. The *FP* and *FN* also denote the number of predicted/classified instances that are incorrectly predicted *yes* when it is actually *no* and the number of instances that are predicted *no* when it actually *yes*, respectively. These constitute the entries to the confusion matrix shown in Table II

TABLE II
LAYOUT of CONFUSION MATRIX

		Predicted class	
		no	*yes*
Actual class	*no*	**TP**	**FP**
	YES	**FN**	**TN**

V. EXPERIMENT SETUP

The major steps we employed in developing the machine learning tasks/algorithms are further discussed below

- Step 1: *Collect the data*: The dataset used in this paper is from cooperative bank .
- Step 2: *Prepare the input data*: This step was done by the original owners of the dataset. And the composition of the dataset is shown in Table I.
- Step 3: *Analyze the input data*: understand the relationship among different features. A plot of the core features and the entire dataset.The dataset is further split into 2/3 for training and 1/3 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. The various proportions of the training and testing datasets used in the paper are shown in Table I.
- Step 4:*Train the algorithm*: The various classification algorithms are trained using a different set of data. The training dataset is shown in Table I
- Step 5: *Test the algorithm*: The various algorithms are used to predict the effectiveness of the algorithm on the test dataset. In evaluating the performance of the classification algorithms, It include accuracy, precision, recall, specificity and F-measure (F1-measure). These values are calculated using the Python scikit-learn tool with input values as the entities of the confusion matrix. The formula for the various evaluating metrics is shown in III, with their definitions. In this paper, a 'positive' instance refers to *no*(signifying there will not be a default in the payment of the loan) whereas the 'negative' instance refers to *yes* (signifying there will be a default in the payment of the loan).

A. *Extracting the Importance Features for Predicting Credit Defaulters*

The total number of features within the bank credit Defaulters dataset. However, not all have significant influence in determining the ability of a given customer in paying his/her loan or not. The designed system is tested with test set and the performance is assured. Evolution analysis refers to the description and model regularities or trends for objects whose behavior changes over time. Common metrics calculated from the confusion matrix are Precision; Accuracy

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 8, August-2020

A. The Predictive Model

The most important features since these features are to develop a predictive model using ordinary linear regression model. This can serve as a tool in determining the credit worthiness of bank clients because these are among the main features taken into consideration by most banks in advancing loans to customers.

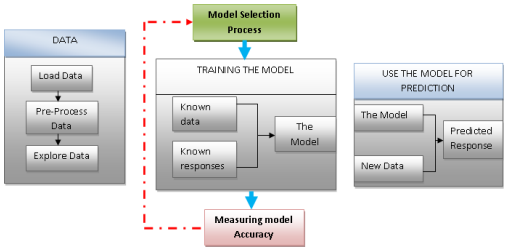

Fig 2. Work Flow in Machine learning

Evaluation Metrics

1 Accuracy:
It measures how often the classifier is correct for both true positives and true negative cases. Mathematically, it is defined as:
Accuracy = (True Positive + True Negative)/Total Predictions
2 Sensitivity or Recall:
measures how many times did the classifier get the true positives correct. Mathematically, it is defined as:
Recall = True Positive/ (True Positive + False Negative)
3 Specificity:
It measure how many times did the classifier get the true negatives correct. Mathematically, it is defined as:
Specificity = (True Negative)/ (True Negative + False Positive)
4 Precision:
Precision measures off the total predicted to be positive how many were actually positive. Mathematically, it is defined as:
Precision = (True Positive)/(True Positive + False Positive)

model prediction

actual loan status		no default (0)	default (1)
	no default (0)	TN	FP
	default (1)	FN	TP

Fig 3:Confusion Matrix

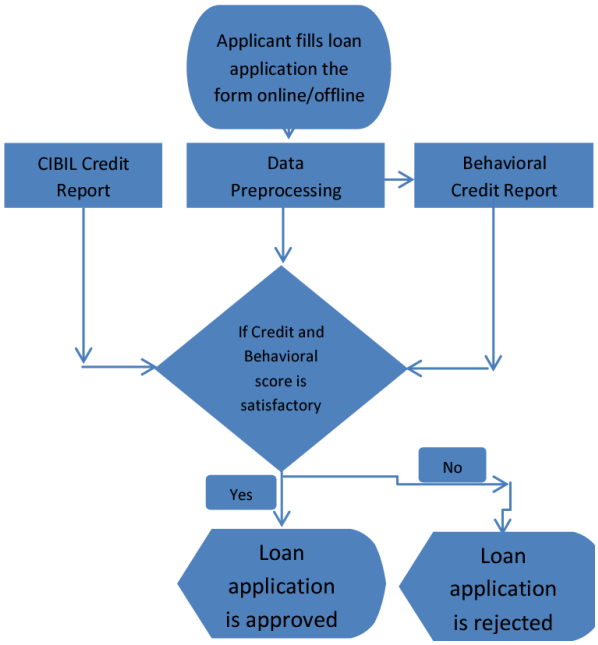

Fig. 4. Predict loan approval flow model

VI.CONCLUSION

In this paper, machine learning approach to study bank credit dataset in order to predict customers' credit worthiness their ability to pay their loan. 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 Nearest Centroid and 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 76% to over 80%. 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 features. The experimental results showed no significance difference in their predictive accuracy and other metrics. We further formulated a predictive model using linear regression, that composed of the most important features, for predicting customers credit worthiness. Predict loan approval in Banking system that will incorporate the most important features that determine credit worthiness of customers in order to formulate bank risk automated system.

REFERENCES

[1] G. McLachlan, K.-A. Do, and C. Ambroise, *Analyzing microarray gene expression data*, vol. 422. John Wiley & Sons, 2005.
[2] E. Ngai, Y. Hu, Y. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decision Support Systems*, vol. 50, no. 3, pp. 559–569, 2011.
[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.
[4] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 8, August-2020

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Van- derplas, 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.

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

[7] 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.

[8] I. Bose and R. K. Mahapatra, "Business data mininga machine learning perspective," *Information & management*, vol. 39, no. 3, pp. 211–225, 2001.

[9] C.-F. Tsai and M.-L. Chen, "Credit rating by hybrid machine learning techniques," *Applied soft computing*, vol. 10, no. 2, pp. 374–380, 2010.

[10] K. Y. Tam and M. Y. Kiang, "Managerial applications of neural networks: the case of bank failure predictions," *Management science*, vol. 38, no. 7, pp. 926–947, 1992.

[11] M. Ghazanfar and A. Prugel-Bennett, "Building switching hybrid rec- ommender system using machine learning classifiers and collaborative filtering," *IAENG International Journal of Computer Science*, vol. 37, no. 3, 2010.

[12] M. Er, L. Zhai, X. Li, and L. San, "A hybrid online sequential extreme learning machine with simplified hidden network," *IAENG International Journal of Computer Science*, vol. 39, no. 1, pp. 1–9, 2012.

[13] M. Lichman, "UCI machine learning repository," 2013.