AN IMPROVED BANK CUSTOMER CHURN AND LOAN PREDICTION MODEL USING SUPERVISED MACHINE LEARNING APPROACH

Dr. Dahlia Sam¹, Dr. K. C. Suresh², Dr. N. Kanya³, C. Tamilselvi⁴, M. V. S. L. Tejasria⁵
¹Assistant Professor, Department of IT, Dr. M. G. R. Educational and Research Institute, Chennai-95
²Associate Professor, Department of CSE, Sri Sairam Institute of Technology, Chennai
³Associate Professor, Department of IT, Dr. M. G. R. Educational and Research Institute, Chennai-95
⁴Assistant Professor, Department of IT, Dr. M. G. R. Educational and Research Institute, Chennai-95
⁵Assistant Professor, Department of IT, Dr. M. G. R. Educational and Research Institute, Chennai-95
¹kcsuresh84@gmail.com, ²tamilselvi.cse@drmgrdu.ac.in, ³kanya.v@drmgrdu.ac.in, ⁴tejasri.it@drmgrdu.ac.in

ABSTRACT

The bank’s revenue is solely based upon the income obtained through the loans given to the people across our country. So it becomes essential for the banks to give their loan to the candidate who is more likely to repay it within a limited period of time. This is where the help of data analytics and machine learning techniques are needed. We propose a model by which the bank officials can analyze and predict whether the concerned candidate will be able to repay. The result of the evaluation will be in binary classification. Customer retention is the technique that is followed by leading organizations to increases their revenue, so they are focusing on retaining the customers. It not only focuses on being a cost effective and profitable strategy, but today’s world is significant. It should oversee customer loyalty and the ways by which it can retain them. The company should introduce various schemes, products and services that should help us to retain the customers.

Keywords: Loan Analysis, Client Analysis, Machine Learning, Customer Churn, Classification.

I. INTRODUCTION

The Main aim of banking is to receive and protect the money which is coming from either the individual customer or leading organization. It also serves by lending money to people which will be repaid within the limited time. It determines the finances of each and every country. The Economic revenue of each country is based on the banking sector. The loan sanctioned is based on the customer requirements. Every bank supports both open-ended and closed-ended loans. open-ended loans are granted only after the approval of specific amounts which include credit card and home equity line of credit. the closed-ended loans cannot be modified by the customer and decreases every payment. This includes personal loans, instalment loans, mortgage loans, etc. loan can be secured by mortgaging homes, vehicles, properties etc. The Another name for Unsecured loan is personal or signature loans. The lender should sanction the loan based borrower’s financial resources. The bank focuses on giving loans in safer hands. The bank follows the procedure which verify and validates the documents provided by the Borrower to approve their loan request. Yet this does not guarantee the loan being repaid. This paper classifies and predicts the customer based on some criteria. This classification is based on Random forest and Neural networks to analyse the data set and summarizes into analysis and prediction phase. Random forest is an approach in machine learning that runs efficiently on large datasets and maintains higher accuracy. It can also fix errors in Unbalanced data sets. The backbone of any business and organization is the customer who significantly contributes success and revenue. so the companies are more aware in gaining their loyalty. Customer relationship management is the strategy that selects customers and strives in creating various cost effective relationships with them. It understands and learns the behavior of the customer to support various organizations to improve customer retention and loyalty. It makes use of business intelligence and analytical models to increase customer retention rate. The personal, demographic and behavioral data are studied in customer relationship management. The dimension of relationship in retention includes loyalty benefits, assistance in brand choice, service beyond expectation, selling various protection items, emotional commitment, building company image and maintaining regular interaction with customers. The
objective is to identify the factors that influence customer behavior and retention and to predict their future action and behaviour. This paper identifies the most profitable and loyal customer to obtain importance in customer relationship management. It also finds the way to increase the organization.

Machine learning is the leading emerging technology trend that transforms every known domain and technological innovation in the phase several industry and enterprises. It has become a Technology of choice in the banking and financial industry. machine learning is the advanced domain in computer science that studies and understands the ability to learn, perform and predict without need of set of instruction or commands. Based upon the pattern and inference, It is able to perform various task, facial recognition, search engine result and speech recognition are some of the example. machine learning can perform tasks without human assistance. The banking industry adopting to these technology to predict customer experience and behaviour. Supervised and unsupervised learning are some types of machine learning, supervised learning use instructions and dataset to help train model. unsupervised learning is not liable for prediction learning. It concludes to a pattern and inference from a cluster dataset, reinforcement learning uses software to interact with the environment. It is not a very feasible method of machine learning. The history of machine learning can be dated to Arthur Samuel who coined the word “Machine Learning” in 1956. However, Machine learning gained popularity in 1990s. This led to using data driven approach.

In 2006 “Deep learning” the ability of the computer to understand and distinguishes was introduced. During the 21st century, Microsoft and Google contributed their part in machine learning growth. In 2015, Amazon also joined hands in machine learning. Machine learning can easily study through large amount of data and analysis trends and pattern not visible to humans machine learning is fully automated and can be accessed without human assistance. It can handle diverse data it gives quick and high accuracy results. It can perform high analysis and prediction. In Banking sector, machine learning tools can optimize customer selection and strengthen relationship. It allows us to identify better targets for new customers and design new products and services. It provides a new dimension for enhancing customer experience. It can perform market research, optimize loan approval process and enhance call center processes. It deals with fraud detection and prevention. It can provide solution for risk concerns. In this paper we discuss leading three machine technique-Random forest, Neural networks and support vector machine.

II. LITERATURE REVIEW

Aboobyda Jafar Hamid and Tarig Mohammed Ahmed [1] gave a loan risk prediction model based on the data mining techniques, such as Decision Tree (J48), Naive Bayes (NB) and BayseNet approaches. The procedure involves training set preparation, building the model, applying the model and evaluating the accuracy. Vimala and Sharmili [2] proposed a loan prediction model using NB and Support Vector Machines methods. Naive Bayes, an independent speculation approach, encompasses probability theory regarding the data classification.

Goyal and Kaur [3] proposed a technique based loan prediction procedure for the customers. M.A.H. Farquad [4] proposed a hybrid approach to overcome the drawbacks of the Support Vector Machine model. This approach involves a SVM-recursive feature elimination, extraction of SVM model and support vectors and a Naive Bayes Tree rule generation. The dataset used here is bank credit card customer dataset from Business Intelligence Cup 2004. But the model does not scalable to large datasets.

Chih-Fong Tsai [5] introduced the neural networks techniques to predict the customer churners in a dataset provided by American telecom companies. The model is built by combining two other neural network techniques for churn prediction with artificial neural networks and self organizing maps. Loan Approval Prediction based on Machine Learning Approach Kumar Arun, Garg Ishan, Kaur Sanmeet have suggested an idea using various algorithms such as Decision tree, Random Forest, Support vector Machine and Adaboost. A study by Aboobyda Jafar Hamid and Tarig Mohammed Ahmed has conducted a case study for a personal loan evaluation using Machine Learning Techniques. This technique is carried out in the Netherlands, ABN AMRO bank. Customer retention and Customer Loyalty in Public Sector Banks. Dr. Preeti Yadav, Manoj Sharma has proposed a study which was motivated by the need to establish the factors which prevent and affect customer retention in banks in Kenya using a case study of Barclays Bank of Kenya. To achieve this objective, the study used a case study design approach, but was not able to achieve with greater accuracy and multiple coordinated modules used in their dataset. Customer Retention Strategies,
A Study of Ghanaian Banks Kate Agudze-Tordzro1 : Samuel Buame, Bedman Narteh has proposed an idea of retaining customers in the bank by giving them some amount of extensions so that the customer maintains a good bond with the bank. David Cohen, Christopher Gan, Hua Hwa Au Yong and Esther Choong has proposed the method for customer retention with the help of their age group, gender, education level and income. The Journal of Credit Risk focuses on the measurement of reducing the credit risk theory and practice in the Excellence Research Australia to transfer the techniques of credit risk management models. A consumer credit risk structural model based on affordability: balance at risk Marcelo Perlin, Marcelo Bruttii Righi and Tiago P. Filomena formulates how the financial dynamics varies according to the borrower's cash flows which illustrates how the risk is estimated risk trend using simulated data sets.

III. ANALYTICAL MODEL FOR LOAN PREDICTION

In this paper, we aim to solve the major complexity faced by the banking sector. We are taking the best algorithm of supervised machine learning based on the accuracy and correctness of the output. Based upon the procedure and techniques of random forest, neural network and support vector machine, the highest accuracy yielding technique is considered the best approach for this problem. The output obtained is of binary classification which involves both analysis and prediction phase. This model is only used by the bank official to enhance their work environment.

3.1 Data Collection

The data set for bank loan prediction is obtained from kaggle. This data set consists of over 1000 values and from which the prediction states is performed. The data model is created using the algorithm discussed about which is applied on the training set and based on test result accuracy the prediction is done. The customer retention module of this paper also follows binary classification. The technique of higher accuracy is applied upon the dataset to yield prediction output this module also consist of analysis and prediction state. The analysis state gives the data comparison of all the attribute in dataset. The prediction phase predict whether the customer will exit a particular bank. The following table represent the attribute

3.2 Data Pre-processing

This step is one of the essential steps in machine learning to be performed before prediction. This step makes sure that the data is balanced and correlated. The data obtained during data collection from the dataset can be loosely controlled, out of range values, impossible combinations and missing values etc. Analysing these unbalanced data may lead to error and degrades the accuracy of the approach. The data processing steps may include data cleaning, data integration, instance selection, normalisation, data transformation, feature selection and addition, data reduction and data discretization etc. It has the ability to add missing values, aggregate information and add label. This step is mainly used in customer retention module and in neural network approach.

3.3 Methodology and Architecture

The flowchart of the proposed model consists of various steps that are performed in any basic supervised machine learning approach. These models and their processes are briefly written. They include data collection, data extraction, data preprocessing, testing and training of data and finally the application of the algorithm. The final output is the predicted output. The below flowchart gives a detailed explanation of the underlying process used in our paper.

3.3.1 Random forest

Random forest comes under supervised learning algorithm ,is a method for which makes use of classification regression that operates by constructing a very large amount of decision trees in training period and the output is the prediction of every single tree.

\[ \hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x) \]  

(1)

A decision can either be binary or non binary tree structure. Its non leaf nodes corresponds to a test features, each branch contributing to the output features over a range and each leaf node representing a category its starts with the root node, tests and classifies corresponding features and output branches are concluded until all leaves are reached. In short a collection decision tree were each tree is unrelated contributes to a random forest.
\[ \sigma = \sqrt{\frac{\sum_{b=1}^{B} (f_b(x')-\hat{f})^2}{B-1}} \]  

(2)

It has advantage of maintaining high accuracy, balancing errors in unbalanced data sets, maintains missing data and runs smoothly an larger datasets. Random forest also work on the begging on unpruned decision tree with randomized features. The Implanced prediction can be overcome by cost sensitive learning and oversampling method. The cost sensitive learning adds class weight into the classifier whereas the oversampling method with minority class or down sampling with majority class aims in balancing original data.

\[ \hat{y} = \frac{1}{M} \sum_{j=1}^{m} \sum_{i=1}^{n} W_j(x_i, x') y_i = \sum_{i=1}^{n} \left( \frac{1}{m} \sum_{j=1}^{m} W_j(x_i, x') \right) y_i \]  

(3)

This results in weighed as well as balanced random forest .The Classifier assigns a class probability based on the fraction of the votes .If the votes is higher than the cut off threshold the instance is named positive else negative. In order to crate balanced random forest ,parameter called sampsize is introduced to do sampling which inturn creates a balanced random forest. We can also use wrapper sampling algorithms such as, smote to create a balanced tree. The smote generates new instance minority class using K-nearest neighbour. It grows many classification trees in training phase into a single classification tree .Each tree generates its own response but the model counts the class with more words .

\[ RF_i = \frac{\sum_{j=\text{all features, } k=\text{all trees}} \text{norm} f_{ij}} {\sum_{j=\text{all features, } k=\text{all trees}} \text{norm} f_{jk}}. \]  

(4)

3.3.2 Neural Network

Neural network is the advanced learning technique that is derived from their biological structure and functionality of human brain. They are highly adaptive and fault tolerance. They consist of set of nodes in layers. The input layer has more Number of hidden layers which provides input to output layer. They are in form of the real numbers which acts as single. The output is the function of the sum of all weighted inputs, they able to address churn prediction problem. The neural network is the connection between the pair of neurons ,node property and rules. The Rules control the weight of each node in the processing. The algorithm uses gradient to determine the weights to minimize their performance.

\[ x_{k+1} = x_k - \alpha_k g_k. \]  

(5)

The gradient is determined by using back-propagation, which involves backward computation. The Conjugate gradient algorithm produces generally faster convergence to determines step size. It is designed to avoid time consuming .The function of gradient update which is computed by one step secant. It is the method that bridges gap between complexity and storage computation of QuasiNewton algorithm. They interpret sensory data, patterns they recognize can be numerical or vectors. They group unlabeled data based on their similarities. They extract features that are fed to other algorithms for clustering and classification. They range from short term behaviour models of dynamics of neural circuitry arising from interaction between neurons. The aim is to create models in order to understand how system works. To understand this a link between data, biological neural network and theory a link has to be maintained. They are ultimately related to cognitive processes and behavior. Applications include non-linear system identification, vehicle control, game playing and decision making, pattern recognition, medical diagnosis and email spam filtering, etc.

The LM algorithm uses this approximation to the Hessian matrix in the following Newton like update:

\[ x_{k+1} = x_k - [J^T J + \mu I]^{-1} J^T e... \]  

(6)

When the scalar \( \mu \) is zero, this is just Newtons’s method, using the approximate Hessian matrix.

When \( \mu \) is large, this becomes gradient descent with a small step size.
3.3.3 Support Vector Machine

Support vector machines are advanced supervised machine learning approaches which identify pattern. The points in high dimensional space are its observations. Support vector machines identify these points in separating hyperplanes. These are classified in proximity to the separating gap in new instances.

\[ \hat{z} \rightarrow \text{sgn}(\vec{w} \cdot \varphi(\vec{z}) - b) = \text{sgn}(\sum_{i=1}^{n} c_i y_i k(\vec{x}_i, \vec{z}) - b) \]  

Each data item is plotted in a n-dimensional space where n is the number of features. The classes are distinguished with the use of hyperplane, this approach is administered using an R algorithm which analyses patterns. They perform regression using kernel trick, mapping their inputs into high dimensional space. They can efficiently perform non-linear classification. When data are unlabeled, unsupervised learning is needed which clusters data into group and map new data. This applies statistics of support vectors to categorize data and its most widely used. SVM is the representation of points in a stage mapped in a way so that they are divided by a clear graph.

They are predicted to belong to category based on the side of the gap in which resided, support clustering algorithm applies the support of vector is used to categorize data which is unbalanced. The original problem in finite dimensional state happens to set that are not linearly separable so that they are mapped to a much higher dimensional state. The hyper play represent the set of point whose dot product with vector is constant such that it is orthogonal, minimal that defines a hyper scale the vector in the hyper plan are the linear combination of images of feature vector in the data set.

IV. RESULTS AND DISCUSSIONS

The analysis phase of bank loan prediction is the data collection and summarization of the attributes in the dataset. This phase started with data collection and data pre-processing. The leading techniques are then applied to the dataset and the higher yielding accuracy is considered as the best approach for this model. The main variations in the above given boosting algorithms are based on the weighting the trained data points. The Random Forest algorithm gives an accuracy of 84%. The accuracies of the remaining algorithms are also shown in Fig 1. In Fig 2, the term precision is defined as the ratio of the true observations which are predicted exactly to the total of the true observations. Recall is defined as the division of the true observations predicted exactly for all other observations in the exact classes. In Fig 3, The scores of the Recall and the scores of the predictions both are represented as F1, where the scores are the mean weights of Precision and Recall. The evaluation is based on the evaluation of the false positives and the negatives are taken into account. When the accuracy is considered the second one proves to be better. When the false negative’s cost and positive’s cost are not similar then the Recall and Precision are considered. In Fig 4, The above graph proves that the XGB and gradient boosting techniques gives the highest accuracy of 85% when compared with the other algorithms. Therefore the ensemble model gives better accuracy with high performance.

![Fig 1: Accuracy of the Various Classification Algorithms](image-url)
Customer churn is a major problem in the banking sector as the bank loses the interest that should be gained from the customers. Furthermore, cost to acquire new customer is higher than retaining the existing one. Hence, customer retention campaign is important to retain the customers from leaving. Successful retention campaign is achievable by targeting the right customers which are the potential churners. Based upon the accuracy of all the three models, Random forest proves to be highly effective. So using Random forest a model for Bank Loan Prediction and Customer churn is developed which proves to be useful for the bank officials.

REFERENCES: