Bank Data Classification for Customer Acquisition, Retention and Loan Prediction

Mrunal Patil, Aditi Sonawane, Venkatesh Kabra, Akhilesh Jain, Dr. S.K.Pathan

Department of Computer Engineering,
Smt. Kashibai Navale College of Engineering,
Savitribai Phule Pune University, Maharashtra.
mrunalpatil1198@gmail.com, aditi.sonawane2212@gmail.com, kabra.venkatesh27@gmail.com, akhileshjain.me@gmail.com, spathan@sin

***Abstract*—The Banking industry generates a massive volume of data every day. It contains customer account information, transaction information, all financial data etc. Data analysis can be used to analyze large volume data to extract meaningful information from it. It helps to uncover hidden information, hidden patterns and to discover knowledge from the large volume data. Banks are facing various challenges like customer retention, fraud detection, risk management and customer segmentation. It needs to focus on these challenges to increase business profit. In this project, data mining algorithms are used for customer acquisition, customer retention and loan prediction. The model is trained for these three functionalities. This trained model is then used for predicting whether the customer is interested, whether he is about to leave the bank or whether a loan should be granted to the customer based on the input provided. The algorithms used for this purpose are C 4.5 and Random Forest.**

***Index Terms*—Banking, customer acquisition, customer retention, data mining, loan prediction.**

I. Introduction

Today, data mining is used in the solution of problems in many fields such as health, finance and education. Data mining studies are being carried out in the field of health for diagnosis of the disease, in customer-oriented industries such as telecommunication, insurance and banking to work on customer churn and customer acquisition. As a data intensive subject, banking has been a popular implementation field for researchers with DM skills over the past decades of the information science revolution. Banks have acknowledged that knowledge instead of financial resources is the new biggest asset. Moreover, the development and popularization of e-banking and mobile banking adds to the exponential growth of real time banking information[13].

Customer acquisition and retention is effective method for the growth of the banks. In the banking sector acquisition, churn and fraud becomes major problem today[13] . So it is important to identify customer's behavior and retain them. To acquire customers, the customer's interest should be taken under consideration[1]. To retain customers first it is necessary to identify which customers are active and inactive. Through data mining classification algorithm, banks can set up classification models by using the relevant personal information and consumption data of the loan applicants in the past, and find out the characteristics of risk customers. Then, use the classification model making classification prediction for new loan applicants, from which identify the risk customers, so as to reduce the risk of default repayment[4].

Machine learning helps to handle large data in the most intelligent fashion by developing algorithms to generate insights from it. Here bank customer data is used. In this work we are using decision trees to perform prediction in all the three functionalities. Various algorithms and research papers were studied and the algorithms which were most accurate were selected for implementation. These algorithms include C 4.5 decision tree for customer acquisition[1] and random forest for customer retention[2,14] and loan prediction[2,5].

II. Motivation

All over the world, banking industry has a revenue of 315 Billion U.S Dollars and a growth rate of 8.8 percentage making it a vital sector for the development of the economy. The core of developing this financial sector is customer acquisition and retention. The Banking industry generates a massive volume of data every day. Data analytics can be used to analyse large volume data to extract meaningful information from it.

III. Literature Survey

Muhammet Sinan Başarslan and İrem Düzdar Argun in their paper Classification Of A Bank Data Set On Various Data Mining Platforms published elaborated the attributes used for prediction for customer acquisition [1]. An application was carried out with classification algorithms of data mining methods in order to predict customer acquisition using the bank marketing data set in the UCI database [11]. In this paper, Accuracy, Precision and Fmeasure criteria were used to test performances of the k-nearest neighbor algorithm (K-NN), Naive Bayes algorithm, C4.5 Decision Tree Algorithm in data mining programs like RapidMiner (Yale), Knime , Weka and R. Within the scope of the study, the training and test sets were compared with 60-40 percent, 75-25 percent, 80-20 percent and 90-10 percent separations in each data mining program to test the performances of all the models applied. Different results were obtained in the four programs used. However, the algorithm that gives the best result in all programs was the decision tree algorithm. This result suggests that decision tree method gives better performance regardless of the program used. In all three of the performance criteria, the best-performing was the C4.5 decision tree.

Research on bank credit default prediction based on data mining algorithm published by Li Ying aims to compare various algorithms for Loan Prediction [5]. This paper uses the bank credit data set loan model sample in kaggle as the target data set for the study. There are 11017 samples and 199 attribute features. After the data collection is completed, the data is viewed and pre-processed. This paper uses Random Forest method, Logistic Regression method and SVM method to establish classification models. Perform research and analysis on pre-processed bank credit default data sets, and use the GridSearchCV method to search for the best parameter [4]. A series of measures for measuring the performance of learning systems such as Overall Accuracy, Recall, precision and F1-score, can be defined based on the confusion matrix. Comparative analysis Experimental results show that compared to Logistic Regression and SVM classification algorithms, Random Forest algorithm is more suitable for the bank credit default precision model because its high classification effect for Class=0, especially when the dataset is very large or has high dimensions.[10]

Machine-Learning Techniques for Customer Retention: A Comparative Study published by Sahar F. Sabbeh presents a comparative study of the most used algorithms for predicting customer churn [14]. The comparison is held between algorithms from different categories. The main goal is to analyse and benchmark the performance of the models in the literature. The selected models are: 1) Regression analysis: logistic regression. 2) Decision tree–CART. 3) Bayes algorithm: Naïve Bayesian. 4) Support Vector Machine 5) Instance – based learning: K-Nearest Neighbor. 6) Ensemble learning: Ada Boost, Stochastic Gradient Boost and Random Forest. 7) Artificial neural network: Multi-layer Perceptron. 8) Linear Discriminant Analysis. Accuracy is used to evaluate the model performance. Accuracy indicates the ability to differentiate the credible and non-credible cases correctly. It's the proportion of true positive (TP) and true negative (TN) in all evaluated news. This study tries to present a benchmark for the most widely used state of the arts for churn classification. The accuracy of the selected models was evaluated on a public dataset of customers in Telecom Company. Based on the findings of this study, ensemble – based learning techniques are recommended as both Random forest model gave the best accuracy. Random Forest achieved the highest performance with approximately 96 percent. Both MLP and SVM can be recommended as well with 94percent accuracy [11]. DT achieved 90 percent, NB 88 percent and finally LR and LDA with accuracy 86.7 percent.

IV. GAP ANALYSIS

A. Customer Acquisition

TABLE I
ANALYSIS OF ALGORITHMS FOR CUSTOMER ACQUISITION

Algorithm	Accuracy	Precision
NB	0.87	0.93
KNN	0.87	0.91
C 4.5	0.97	0.92

We can observe that C4.5 algorithm has a very high accuracy of 90.6 percentage, therefore we will use C4.5 algorithm for Customer Acquisition[1].

B. Customer Retention

TABLE II
ANALYSIS OF ALGORITHMS FOR CUSTOMER RETENTION

Paper Name	Algorithm	Efficiency (per)
Analysis of Banking Data using ML	ANN	72
Techniques for Customer Retention- A comparative study	Logistic Regression Random Forest KNN Naive Bayes	87 96 91 88

By close analysis of the figures we can say that Random Forest Algorithm can be used for Customer Retention as it has a accuracy of 96 percentage in the experiments carried out by the Researchers[2].

C. Loan Prediction

TABLE III
ANALYSIS OF ALGORITHMS FOR LOAN PREDICTION

Algorithm	Efficiency (per)
Random Forest	88.63
Logistic Regression	71.30
SVC	85.80

Research on bank credit default prediction based on data mining algorithm- This paper established bank credit default prediction models using Random Forest, Logistic Regression and SVM classification algorithms and compared the classification effect. Results show that Random Forest algorithm is more suitable.[5]

V. PROPOSED WORK

A. System Architecture

The system architecture includes various phases like data collection, making training and testing dataset, implementing algorithm and then using this model for predicting the class of input variable provided. Proposed system architecture is shown in figure 1.

First bank customer data is prepared for processing. Input dataset contains both types of values categorical and numerical. To implement decision tree algorithm we require these categorical and numerical variables. The data is divided into

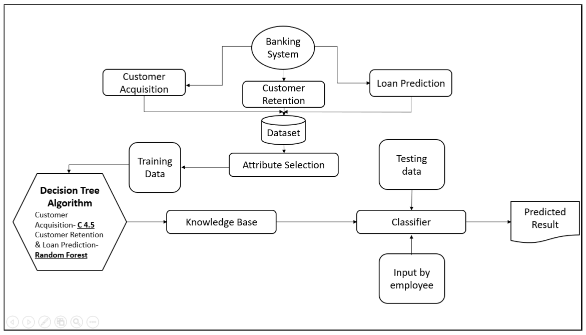

Fig. 1. System Architecture of Proposed System

two parts training data and testing data to check performance of the model[1,2,14]. And then this trained model is used for prediction of the input provided by the bank employee. Here we have used C 4.5 and Random Forest as a machine learning algorithm for classification and prediction. Decision Tree can process multiple inputs efficiently and also gives highest efficiency [1,5]. Hence these algorithms are used.

B. Working of C 4.5

C4.5 builds decision trees from a set of training data in the same way as ID3[1], using the concept of information entropy. The training data is a set of already classified samples. Each sample consists of a p-dimensional vector , where they represent attribute values or features of the sample, as well as the class in which falls. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurses on the partitioned sublists. This algorithm has a few base cases.

- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

The calculation of Entropy is done by:

$$\text{E(S)} = \sum_{i=1}^{n} -P_r(C_i) * log_2 P_r(C_i)$$

The calculation of Information Gain is done by:

$$\text{G(S,A)} = \text{E(S)} - \sum_{i=1}^{m} P_r(A_i) E(S_{Ai})$$

where,

- E(S) – information entropy of S
- G(S,A) – gain of S after a split on attribute A
- n – nr of classes in S
- Pr(Ci) – frequency of class Ci in S
- m – nr of values of attribute A in S
- Pr(Ai) – frequency of cases that have Ai value in S
- E(SAi) – subset of S with items that have Ai value
- The C4.5 algorithm improves the ID3 algorithm by allowing numerical attributes, permitting missing values and performing tree pruning.

C. Working of Random Forest

Random forests or random decision forests are an ensemble learning method for classification that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees [1]. In particular, trees that are grown very deep tend to learn highly irregular patterns: they overfit their training sets. Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance [2]. In this case, we are using Random Forest for determining whether the customer will discontinue his/her relationship with the bank [9]. The same can be used for predicting if the customer will be able to pay his loan within specified time considering attributes like credit score, property, etc[12].

VI. Conclusion

Data mining is a technique used to extract vital information from existing huge amount of data and enable better decision-making for the banking industries. The data is then analyzed and the information that is captured is used throughout the organization to support decision-making. In banking field massive volume of data is continuously generating. This data can be used to extract meaningful information from it.

Data Mining techniques can banking sector for better targeting and acquiring new customers, most valuable customer retention and loan prediction providing segment based products, analysis of the customers, transaction patterns over time for better retention and relationship, risk management and marketing. Finally we conclude that Bank will obtain a massive profit if they implement data mining in their process of data and decisions.

VII. Future Work

- Fraud Detection:
  With the increase in an online transaction, the incidents of fraud have increased too. To avoid such fraud we can extend this system using the big data technology which helps banking industry to understand the financial history and spending pattern of customer and increase security on every unusual transaction[13]. This will help them to mitigate any fraudulent activities before it grows bigger.
- Offering Personalized Services:
  Offering Personalized Services to a customer is nothing but the next level of marketing where they offer product and services to as per customers interest and requirement. Banking industry collects data from e-commerce website

and Big data technology analyze the buying habit, interest and requirements of individual customer by doing sentimental data analysis[13].

- Addressing Compliance Requirement:
  Banks and financial services are required to do regular compliance, audit and maintain certain regulations for their data, finance, privacy and security measures. Banks now have access to billions of customers' needs. They can use Big Data to cater to serve the customers more effectively.

References

[1] Muhammet Sinan Başarslan and İrem Düzdar Argun, ”Classification Of A Bank Data Set On Various Data Mining Platforms”, IEEE, 2018.

[2] Priyanka S. Patil and Nagaraj V. Dharwadkar, ”Analysis of Banking Data Using Machine Learning”, *In International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), (I-SMAC 2017)*, IEEE, 2017.

[3] B. Rajdeepa1 and D. Nandhitha2, ”Fraud Detection in Banking Sector using Data mining”, Vol. 4, IJSR,2015.

[4] X.Francis Jency, V.P.Sumathi and Janani Shiva Sri, ”An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients”, Vol. 7 Issue-4S, IJRTE, 2018.

[5] Li Ying , ”Research on bank credit default prediction based on data mining algorithm”, Vol. 5 , Issue 06, THEIJSSHI, 2018.

[6] Dr. Radhakrishna Rambola, Prateek Varshney and Prashant Vishwakarma , ”Data Mining Techniques for Fraud Detection in Banking Sector”, *In 2018 4th International Conference on Computing Communication and Automation (ICCCA)*, IEEE, 2018.

[7] Utkarsh Srivastava, Santosh Gopalkrishnan, “Impact of Big Data Analytics on Banking Sector: Learning for Indian Banks”, ELSEVIER 2015.

[8] M.S. Başarslan, F. Kayaalp, “Customer churn analysis with classification algorithms in telecommunication sector. ICAT'17, Istanbul, Turkey, 2017

[9] Mei Mei. Application of data mining classification algorithm in credit card risk management [J]. modern computer,2013(19):13-16.

[10] Liaw A, Wiener M. Classification and regression by randomForest[J]. R news, 2002, 2(3): 18-22.

[11] S. Moro, P. Cortez, and P. Rita, “A data-driven approach to predict the success of bank telemarketing,” Decision Support Systems, vol. 62, s. 22-31, 2014

[12] Iain Brown, Christophe Mues,” An experimental comparison of classification algorithms for imbalanced credit scoring data sets”, Expert Systems with Applications 39 (2012) 3446–3453, ELSEVIER.

[13] Hossein Hassani, Xu Huang 2 and Emmanuel Silva,” Digitalisation and Big Data Mining in Banking”, IEEE, 2018.

[14] Sahar F. Sabbeh,” Machine-Learning Techniques for Customer Retention: A Comparative Study”, Vol. 9, No. 2, IJACSA, 2018.