International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

An Effective Method to Understand Bank Customer Retention System

Tushar Suri, Shailly Singh, Tejkaran Singh, Arul Kumar R, S. Metilda Florence

Abstract - Banking industry is one of those industries where data is generated every day in large amounts. This data can be used for extracting useful information. Hence it is important to store, process, manage and analyze this data. It helps in making business lucrative. This data helps in making prediction which helps in solving problems that are faced by banks these days. People are constantly working on various aspects of Banking System like fraud detection, Risk Analysis etc. Various Machine Learning algorithms like CNN, ANN etc. have been used in order to study the patterns from such datasets. Here, we are focusing on risk analysis, customer retention and customer segmentation. In this paper, we have implemented classification algorithm, namely Decision Tree, for different aspects. Training of model is done on the given data and testing is done on real time data provided by the user. This study might help various banking systems to gain knowledge about their investment scheme for a particular customer. Thus, the banking companies will have a greater control on their customer and can develop policies that will benefit both the parties.

Keywords – Banking Industry, Risk Analysis, Customer Retention, Customer Segmentation, Fraud Detection

I. INTRODUCTION

Banking industry is one of those sectors where huge amount of data is generated on a daily basis. It handles this data using methods like data analysis which extracts hidden information, hidden patterns and undiscovered information that might be useful to the analysts. This data usually contains customer information, customer account information, transaction information, loan related information and the entire financial data.

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.
*Correspondence Author
Tushar Suri*, Department of IT, SRMIST, Chennai, tushar.suri7@gmail.com
Shailly Singh, Department of IT, SRMIST, Chennai, shaillysingh159@gmail.com
Tejkaran Singh, Department of IT, SRMIST, Chennai, tjs6818@gmail.com
Arul Kumar Rajappan, Department of IT, SRMIST, Chennai, arulk3398@gmail.com
Dr. S. Metilda Florence, Assistant Professor (Senior Grade), Department of IT, SRMIST, Chennai, medildam@srmist.edu.in

Various challenges and problems are faced by the banks like risk analysis, customer retention, customer segmentation and fraud detection. Customer retention and risk analysis are effective methods for growth and increasing profits of business.

Churn and risk are the main problems faced by banks and so it is important to identify the customer's behavior and retain them as per the results.

To handle such huge data, there are various Machine Learning based algorithms for which generate useful information for the banks. In this work, we are using

Bank Marketing Data. Classification is performed using decision trees. Marketing selling campaigns constitute a typical strategy to enhance a business. In this project, it is attempted to work on a dataset to predict if a client will potentially subscribe to a long-term deposit at a certain bank after receiving a marketing call. While all the power remains with the customer to continue with the bank or not, it becomes important to decide from a bank's perspective as to whether it is worthwhile to invest time on a particular customer, in their marketing campaign or not. Here we have tried to give a bank an idea as to how their customer would behave in future, particularly with respect to the retaining.

II. RELATED WORK

Literature survey reflects upon analytical studies on banking and data related to finance. The analysis is carried out using different methods and techniques. Analysis and prediction models have been made by various researchers that use different data mining techniques.

Paliwal and Kumar [1] reported that ANNs were widely used in research work that focused on prediction and classification in a varied mix of applications from different fields. They regarded neural networks as competing approaches for model building and traditional statistical techniques.

Angelini, Tollo and Roli [2] noted that Artificial Neural Networks (ANNs) have been effectively emerging in credit scoring because of their ability to develop a non-linear relationship between a particular set of inputs and outputs. They treated ANNs as black boxes, since their internal configurations made it impossible to extort any symbolic details.

Chitra and Subashini [3] found that the solution discovers churn trends within the actions of past churners by using data mining statistical algorithms(CART), and used this information to allocate current customers, a possible churn score rating.

CART algorithm helped them to classify the customers based on various attributes and hence handle the categorical attributes present within the dataset.

Oyeniyi and Adeyemo [4] showed how data mining model used K-means for clustering and JRip, which is rule based algorithm for working on Nigerian dataset. This helped them to develop model that can ultimately give banks with important knowledge regarding customer transactional trends, identify likely churners and hence develop the customer retention modalities.

Zakrzewska and Murlewski [5] worked with the various algorithms mainly used for cluster analysis i.e. k-means, two phase clustering and the DBSCAN for bank customer segmentation. K-means is an algorithm which mainly depends in the choice of input parameter k and works well with large multidimensional datasets.

I. PROPOSED METHOD

A. System Architecture

The system architecture comprises several segments like data gathering, data preprocessing, dividing training and testing dataset, employing Decision Tree Algorithm and outcome examination. Projected system architecture is presented in figure 1.

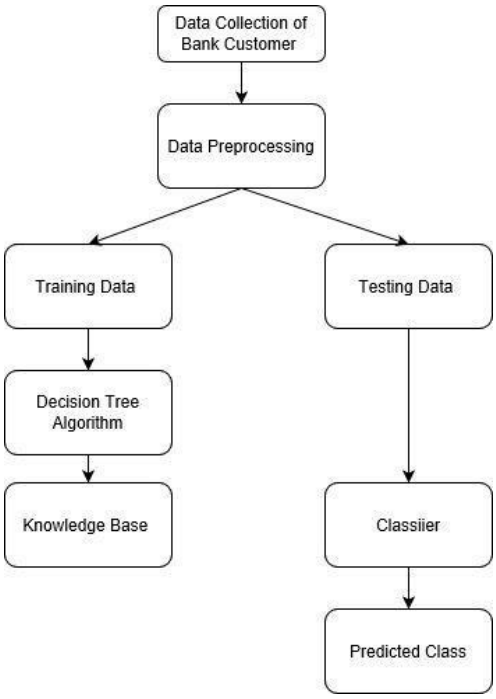

Fig. 1: System Architecture

First customer banking data is organized for processing. Input dataset that is collected has two types of attributes, categorical and numerical. In order to implement Decision Tree algorithm, first of all we need to convert the whole data into numerical values and then use these values as the input to algorithm for processing [6]. Hence first data is preprocessed for further processing. The data is divided into two sets training dataset and testing dataset to understand how the model behaves and hence analyze its performance.

Here we shall be using Decision Tree as a machine learning algorithm for first classification and then prediction.

Decision Tree algorithm can easily process multiple inputs efficiently and effectively. It also handles large and complex datasets very easily.

B. Working of Decision Trees Algorithm

A decision tree is majorly considered as a classification scheme which has the ability to generate a tree and a set of rules, depicting the model of various classes, from a particular dataset. Generally speaking, the collection of instances available for working on a classification method is segregated into two mutually exclusive subsets-a training dataset and a testing dataset. The training one is the classifier used to extract, while the testing one is used to calculate the classifier's accuracy.

In order to determine the accuracy of the classifier, we check the percentage of the test example correctly ranked by the algorithm.

We have categorized the instances into two different types. Numerical characteristics are called attributes, and characteristics whose scope is not numerical are called categorical attributes. There is one distinct feature called the mark of class. The cataloging objective is to create a succinct model which can be used to forecast the class of records whose type mark is unknown.

a. CART

CART is one of the common methods used by the machine learning community to create decision trees. CART generates a binary decision tree by dividing the instances according to a single attribute function at each node. To determine the best split, CART uses the gini index. CART follows the aforementioned decision tree construction theory [3]. The aim of this technique is just for the sake of totality. The preliminary split generates two nodes, both of which we seek to fragment the root node in the same way. Once again we are analyzing all the aspects of data to identify how the splitting of candidates is done. If there is no split that reduces a given node's diversity substantially, we will mark it as a leaf node. Eventually only leaf nodes remain, and we have evolved the tree of complete decision. Generally speaking, the complete tree may not be the tree that does the best job of classifying a new set of data, due to overfitting. At the end of the tree growing process, some leaf of the full decision tree has been allocated to every record of the training set. Every leaf now has a class and error rate to be assigned. A leaf node's error degree is the percentage of improper arrangement at that node. The error rate of a complete decision tree is a weighted totality of the fault raters of all the leaves. Each leafs involvement to the total is the error rate at that leaf increased by the likelihood that an instance will end up in there. Expected information needed to classify a tuple is given by

$$Info(D) = -\sum (p_i * log_2(p_i)) \quad \text{(I)}$$

Where p is the likelihood value that occurs in a given tree node. Additional information that is still needed to arrive at a conclusion is given by

$$Info_A(D) = \sum (|D_j| / |D| * Info(D_j)) \quad \text{(II)}$$

Where $|D_j| / |D|$ acts as weight of jth partition [7].

$$Gain(A) = Info(D) - Info_A(D) \quad \text{(III)}$$

II. IMPLEMENTATION

First and foremost, the understanding of the dataset was done from the UCI dataset website and also from the relevant research papers that have worked on this dataset. All the features (Demographic, Financial behavior, last contact of current campaign, data from previous campaign, macroeconomic) were understood. The potentially important features were identified by looking at the statistics of the data and data visualized as histograms [8].

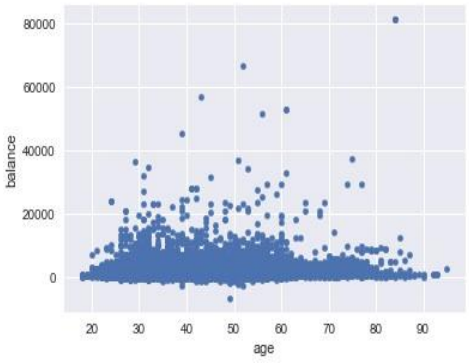

Fig. 2: Scatter Plot of Age-Balance

Some features were re-casted as found appropriate. A data point that was proving to be an outlier was removed. Stratified splitting of data into testing and training was performed. The training dataset was observed to have unknown values. Ways to generate the unknowns were identified keeping in mind the business sense with which banks work on [9]. Further the dataset is read and visualized using different types of plots (box plot, bar plot, etc.). The next step is Data Cleaning. In this process only that much data is kept which is useful in our modules and the remaining data is removed and for further implementation this cleaned dataset is used The Actual implementation starts with building Data Model using cleaned dataset. The Data Model majorly uses Decision Tree algorithm as it is a section of very authoritative Machine Learning model capable of achieving extraordinary precision in many jobs while being extremely interpretable.

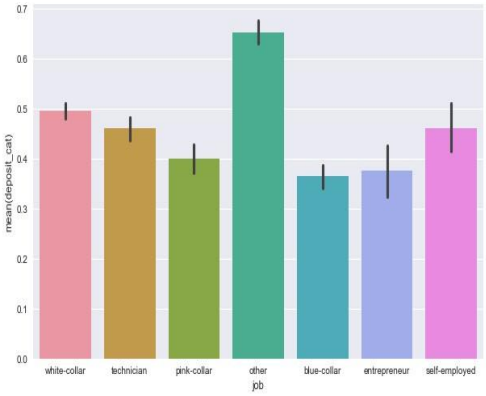

Fig. 3: Segmenting Customers based on Jobs

A. Decision Tree Classifier

Decision Trees (DTs) are non-parametric supervised learning method used for classification and regression. The intention is to generate a design that predicts the value of a target variable by cramming manageable rules of knowledge inferred from the data traits. Decision Tree Classifier is a class proficient in classifying diverse classes on a dataset. As with additional analysis, Decision Tree Classifier grosses as its input, two arrays: an array X, meagre or compact, of size [p_samples, p_features] holding the training samples, and an array Y having integeral values, size [p_samples], holding the class labels for the training samples: *max_depth=int, default=None max_depth=int, default=None*

If None, nodes will be maintained until all leaves are pure or until all leaves comprise less than samples min samples split. The default rates for the considerations that command the structure of the trees (e.g. max depth, min samples leaf, etc.) commence to adequately advanced and unpopulated trees that may theoretically be very deep on some datasets.

The intricacy and extent of the trees should be managed by setting certain bound values to reduce the memory consumption.

The characteristics are always permuted at random on each break. Therefore, even with the identical training data and max features = p features, the best found split will differ if the criteria for improvement is similar to numerous splits reckoned during the pursuit for the finest split. The random state has to be set to achieve a deterministic behavior during fitting.

Hence we have tested our decision tree by iterating its max_depth and found that when max_depth=6, testing and training scores do match with our needs.

It was seen that, higher the depth, training score increased and matched perfect with the data set on which it was trained. However higher the depth the tree goes, it over fitted to the training data set. So it is of no use to keep increasing the tree's depth. According to above observations, tree with a depth of 2 seems more reasonable as both training and test scores are reasonably high.

Now we have to determine which is the most important feature among the dataset which will lead us to our result of customer retention.

We found out the '*duration*' feature is most important via feature_importances implementation [10].

III. EXPERIMENTAL RESULTS

The Bank Marketing dataset is used for customer churning problem. This dataset is freely accessible at UCI machine learning repository. It comprehends customer's information. The data is related with direct marketing campaigns. There are overall 20 inputs and single output. The input has information such as age, job, education, duration of call, outcome of previous campaign etc.

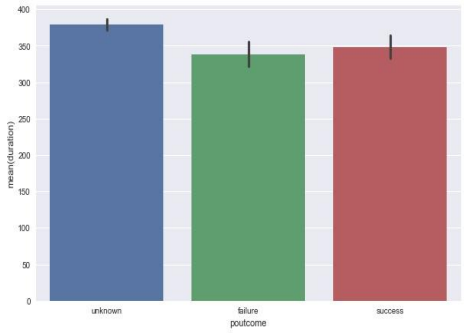

Fig. 4: Visualizing our output variable

The output is the prediction i.e. if the customer is retained or not retained. If the customer is retained, then it has been identified as a low risk proving and if the customer is not retained, then it means it can act as a risk to the bank. With this implementation we recieved the following results as shown in Table 1.

Table-I: Results

Depth	Training Score	Testing Score
2	0.7285	0.7268
3	0.7704	0.7572
4	0.7885	0.7742
6	0.8080	0.7796
max	1.0	0.7330

It could be seen that, higher the depth, training score increases and matches perfects with the training data set. However higher the depth the tree goes, it over fit to the training data set. So there is no point in increasing the tree's depth. According to above observations, tree with a depth of 6 seems more reasonable as both training and test scores are reasonably good. The accuracy score and the area under curve has been achieved as shown below.

Accuracy score:

0.7796

Area Under Curve:

0.8605

IV. CONCLUSION AND FUTURE WORK

In banking industry, large volume of data that is being continuously generated. It is analyzed using data mining and hence used as training data with algorithms. Here the dataset used was in raw form. Firstly, the dataset is visualized and cleaned. After that useful information is extracted from it and used for training algorithm. Classification algorithm i.e. Decision Trees is used for classifying data at every depth with respect to the attributes. Decision Tree is one of the most powerful algorithm which finds its use in various industries like Sports, Medical Science etc. The leaf nodes in decision trees give results indicating whether the customer will be retained or not and hence showing whether it would pose risk to banks or not. That means, if a customer is retained it means it possess less risk, and if the customer is not retained it means it possess more risk than usual customers. We tested our data for various depths, results for which are shown in previous section. It can be seen that depth = 6 provided us with better results, hence that has been used as our model. This work can also have Fraud Detection as one of the aspects along with the other three aspects i.e. Customer Retention, Risk Analysis and Customer Segmentation. Fraud Detection can help in predicting whether a transaction in dispute would be fraudulent or non-fraudulent. Central governments can use and deploy this model with improvements in order to study Bank patterns. Further we can use Generative Adversarial Networks (GANs) for comparing results, with results obtained from Decision Trees Algorithm and analyze which is more efficient. Many regression techniques can also be used so as to have a better understanding about this dataset and therefore grabbing knowledge as to how banks tends to behave w.r.t a particular customer. Moreover, we can deploy the same over a mobile as an application for Android and IOS. This can act as an easy way for anyone to analyze a particular banking data.

REFERENCES

1. Paliwal, M. and Kumar, U.A. (2009). "Neural Networks and statistical techniques: A review of applications". *Expert Systems with Applications, 36(1), 2-17.*
2. Angelini, E., Tollo, G., di. and Roli, A. (2008) "A Neural Network approach for credit risk evaluation". *A Quarterly Review of Economics and Finance*, 48(4), 733-755.
3. Chitra, K. and Subashini, B. (2011). "Customer Retention in Banking Sector using Predictive Data Mining Techniques". *Proceeding of the 5th International Conference on Information technology*, Amma, Jordan, pp.1-4.
4. Oyeniyi, A., O. and Adeyemo, B. (2015) "Customer Churn Analysis in Banking Sector Using Data Mining Techniques". *African journal of Computing and ICT*, 8(3), 165-174.

5. Zakrzewska, D. and Murlewski, J. (2016) "Clustering algorithms for Bank Customer Segmentation". *Proceedings of the 5th International Conference on Intelligent Systems Design and Applications*, Warsaw, Poland, pp.197-202
6. Patil, S., P. and Dharwadkar, V., N. (2017) "Analysis of Banking Data using Machine Learning". Proceedings of the *International Conference on I-SMAC*, Coimbatore, India, pp.876-881.
7. Han, J., Kamber, M., & Pei, J. (2016). *Data Mining: Concepts and Techniques (3rd ed., p.337).* Waltham, USA:Morgan Kaufmann Publishers.
8. Moro, S., Cortez, P. and Rita P., (2014) "A Data-Driven Approach to Predict the Success of Bank Telemarketing". *Decision Support Systems*, 62, 22-31.
9. Abbas S. (2015), "Deposit subscribe Prediction using Data Mining Techniques based Real Marketing Dataset". *International Journal of Computer Applications*, 110(3) 1-7
10. Colaianni, G., Magdangal, J. and Mitchell, M., (2016) "Factors determining term deposit purchases: How a Bank can get Other People's money", Kennesaw State University.

AUTHORS PROFILE

Tushar Suri is currently pursuing B.Tech in the field of I.T. from SRMIST, Chennai. He is a fourth year student having interest in the field of Data Science, Business Intelligence, DBMS and UML. He has worked on Neo4j Database, SQL Server, Visual Paradigm, Micro Strategy etc.

Shailly Singh is currently pursuing B.Tech in Information Technology from SRMIST, Chennai. She likes to read books related to technology and has interest in the field of Web Development and App Development.

Tejkaran Singh is currently pursuing B.Tech from SRMIST in Information Technology. He has working experience in Web Development and Machine learning. He is enthusiast about developing websites and software

Arul Kumar Rajappan is currently pursuing Bachelors in Technology from SRMIST in Information Technology. Machine Learning is his field of interest and he thoroughly enjoys reading about its applications in day to day life.

Dr.S.Metilda Florence, completed PhD in Video Processing area from Bharathiar University, Coimbatore, India. Received M.Tech degree in Computer Science from SRM University, Chennai, India and MCA from Bharathidasan University, Trichy, India. She has 18+ years of teaching and research experience. Currently she is an Assistant Professor (Senior Grade) in Information Technology Department at SRM Institute of Science and Technology, India. She is the author of 10 International journals and 8 International conference papers. Her research interests include Image processing, Data mining, Recommendation system, Machine learning and deep learning.