Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/090310

Churn Prediction for Savings Bank Customers: A Machine Learning Approach

Prashant Verma[1, 2,*]

[1]Analytics Department, Global IT Centre, State Bank of India, Navi Mumbai, India
[2]Department of Statistics, Banaras Hindu University, Varanasi, India

Received: 18 Jan. 2019, Revised: 22 May 2019, Accepted: 24 Jul. 2019.
Published online: 1 Nov. 2020.

Abstract: This paper explores churn prediction for savings account customer, based on various statistical & machine learning models and uses under-sampling, to improve the predictive power of these models, considering the imbalance characteristics of customer churn rate in the data. Model Accuracy, Area under the curve (AUC), Gini coefficient, and Receiver Operating Characteristics (ROC) curve have been utilized for model comparison. The results show that out of the various machine learning models, Random Forest which predicts the churn with 78% accuracy, is the most powerful model for the scenario. Customer vintage, customer's age, average balance, occupation code, population type, average debit amount, and an average number of transactions are found to be the variables with high predictive power for the churn prediction model. The commercial banks can deploy the model in order to avoid the customer churn so that they may retain the funds, which are kept by savings bank (SB) customers. The article suggests a customized campaign to be initiated by commercial banks to avoid SB customer churn. Hence, by giving better customer satisfaction and experience, the commercial banks can limit the customer churn and maintain their deposits.

Keywords: Savings Bank, Customer Churn, Customer Retention, Random Forests, Machine Learning, Under-Sampling.

1. Introduction

Currently, Indian commercial banks are undergoing prodigious and tangled changes; also, the entire industry faces many difficulties related to the growth of various products and services. Information technology has been grown rapidly using cloud computing, machine learning technologies in the last few decades. Also, financial regulation specialized in capital regulation is increasingly strengthened, and the process of financial disintermediation and interest rate marketization is gradually accelerating, resulting in a sharply narrowed interest margin for banks [1]. Likewise, the e-Business enterprises represented by the third-party payment, which use both the Internet technology and big data technology, are advancing to the field of conventional business of banks, which subvert the gain of the commercial bank's channels. At the current moment, the financial consumption demand of Indian consumers is continuously expanding. The banking customers pay more attention to their experience, customized services, assortment, and agility, which further intensifies the competitiveness among commercial banks.

Customer retention and expanding the revenues from existing customers by up/cross-selling is a much more profitable strategy for growth in comparison to new customer acquisition. In order to maximize the profit, commercial banks must increase the customer base by incrementing sales while decreasing the number of churners. Furthermore, it is common knowledge that retaining a customer is about five times to six times less expensive than acquiring a new one [2] and [3], while [4] suggests acquiring a new customer is anywhere from five to 25 times more expensive than retaining an existing

*Corresponding author e-mail: prashantvermag@gmail.com

one. In addition, positive word-of-mouth from existing customers leads to low-cost or almost free customer acquisition. Besides it, SB customer retention also help Bank to maintain Current Account & Savings Accounts (CASA) Ratio.

One of the paramount competitions among commercial banks is customer retention, especially for high-value customers. As customers are directly related to profits, commercial banks must avoid the loss of customers while acquiring new customers. [4] believes that by reducing the customer defection rate by 5%, companies can increase profits by 25% to 95%, while Business Week thought the profits would increase by 140%. As can be seen, reducing customer attrition has a significant impact not only on increasing profits for commercial banks but also on enhancing their core competitiveness. Therefore, it is strongly needed for commercial banks to improve the capabilities to predict customer churn, thereby taking timely measures to retain savings bank customers and preventing other clients from churning.

Churn is a critical area in which the banking domain can make or lose their customers. Hence, the business spends a lot of time making statistical predictions, which successively helps to make the necessary business conclusions. The customer churn can be averted by studying the demographic features and history of the customers, especially the transaction patterns.

In pursuance of predicting customer attrition for commercial banks, many scholars carried out the research by using various data mining methods. Many scholars used classification methods to predict customer churn. [5] used CART, Tree-Net, and C5.0 classification method for predicting the loss of customers of commercial banks, the results narrated that the CART algorithm was found to be the best algorithm to predict the customer churn. [6] proposed a Fuzzy C-Means clustering algorithm for retail banks' churn prediction model. [7] applied both CART and C5.0 classification techniques to a commercial bank in India. It is found that the prediction accuracy of the CART algorithm is better than that of the C5.0 classification. [8] discussed commercial bank customer churn prediction based on the Support Vector Machine (SVM) model and uses a random sampling method to improve the SVM model. Their study showed that the technique could effectively enhance the prediction accuracy of the selected model. In 2015, [8] presented a new cost-sensitive framework for customer churn predictive modeling. The results showed that using a cost-sensitive approach yields an increase in cost savings of up to 26.4%.

Many scholars have used the random forests method for churn prediction. For example, [9] used three ways, including sampling techniques, gradient Boosting and weighted random forests for churn prediction, and the result expressed that the weighted random forest method was found as the best algorithm for prediction. [10] proposed improved balanced random forests (IBRF) integrating sampling techniques, and cost-sensitive learning techniques and compared with Artificial Neural Network (ANN), decision trees, and SVM methods, the results displayed IBRF work most effectively. Similarly, a few other scholars utilized an algorithm based on sequence patterns [11]. [12] Established the Logistic regression model and believed that this method was better for predicting the loss of customers, while [13] suggested that the SVM model was better. Though the previous studies have exploited numerous data mining methods to predict customer churn, there is not yet a coherent conclusion on the application effect of these models, the accuracy of these models is not ideal, and the potential of processing huge data sets are still being revamped.

Therefore, considering the characteristics of banking customer attrition, this paper uses the Random Forest model combined with a random sampling method to improve the performance of customer churn prediction. Also, we have proposed to use a weighted variable as a feature to eliminate the effect of the under-sampling method. The current paper explores commercial bank’s savings account customer churn prediction based on various statistical & machine learning models (Generalized linear model (GLM), Decision Tree, XG-Boost, ANN, and Random Forest) and uses random sampling method (Under-Sampling), to improve the predictive power of these models, considering the imbalance characteristics of customer churn rate in the data.

1.1. Objectives

To build Statistical/Machine Learning models to identify the high-value SB Customers who are likely to churn, and based on the predictions by the model, generate leads of target SB customers who are likely to churn which in turn would contribute to the retention of customers so that a customize campaign might be scheduled.

1.2. Scope

The Churn Model has been specially designed for SB accounts having an age group of customers between 21 and 50, also maintained the average balance between ₹50,000 and ₹10 Lakh for two consecutive quarters in their savings bank account. This model has been developed for complete Bank-Level data, including all states of the country.

2. Material and Methods

2.1. Data Acquisition

The data required for the model building have been extracted from the data warehouse of the bank using the Structured Query Language (SQL) queries.

Savings Bank customers with Quarterly Average Balance (QAB) between ₹ 50,000/- and ₹ 10 lakh for both quarters (Apr-Jun, 17 & Jul-Sep, 17) in SB account and age between 21 years and 50 years as on 30.04.2017 were acquired. This dataset would be referred to as 'Case Data Set' hereafter.

2.2. Data Contents

The Case Data Set consists of the following information of SB Customers at Bank level.

i. Demographic information: Age, Gender, Population Group, and Marital Status.

ii. Transaction Data: Branch Debit & Credit, ATM, POS, E-COM, GCC, INB, and MBS transactions.

iii. Other Data: Income, Occupation, Existing loans, Deposits, Customer vintage, Joint Ventures & investments (Life insurance, General Insurance, Mutual funds, etc.), KYC compliance, and usage of INB service.

2.3. Churn definition

Since the beginning of data mining, the discovery of knowledge from the databases has been carried out to solve various problems, and it has helped the business come up with practical solutions. Large commercial banks are behind improving revenue due to the increasing loss of customers. The process where one customer leaves one commercial bank or a particular product/service is called churn.

- Including the actual churn (i.e., Account Status Closed + Inactive Account), four conditions have been proposed by the business unit of the bank to define the Churn of SB Customers.

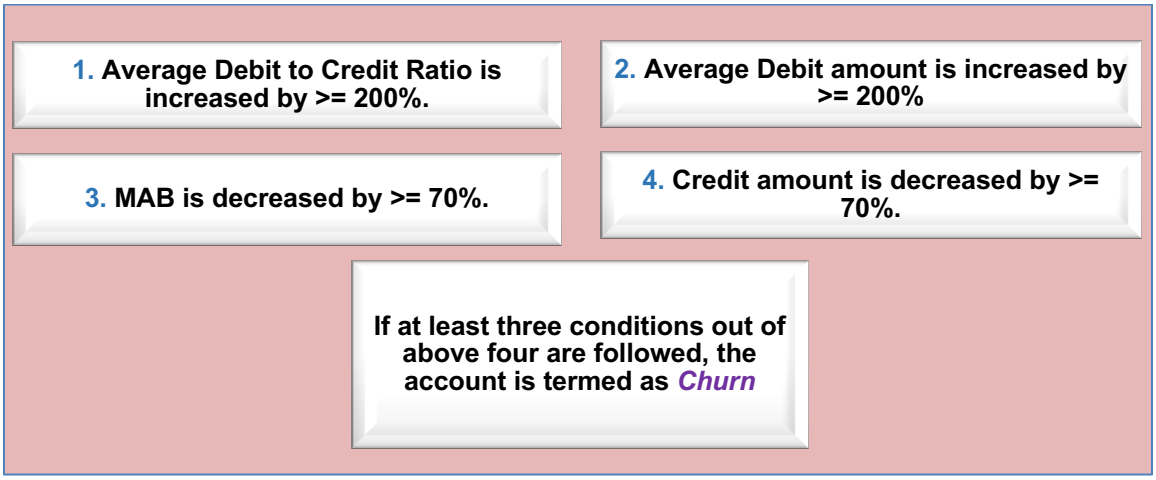

Fig. 1: Conditions to define Churn Definition

- The comparison has been done using the average behavior of customers between two periods (April 2017 to September 2017) and (December 2017 to March 2018).
- A buffer period of two months (October 2017 to November 2017) has been taken so that the behavior of such accounts can be understood more precisely. Meanwhile, the business units will have enough time to act upon the leads (Target customers who have a higher likelihood to churn) and retain the customers before churning actually.

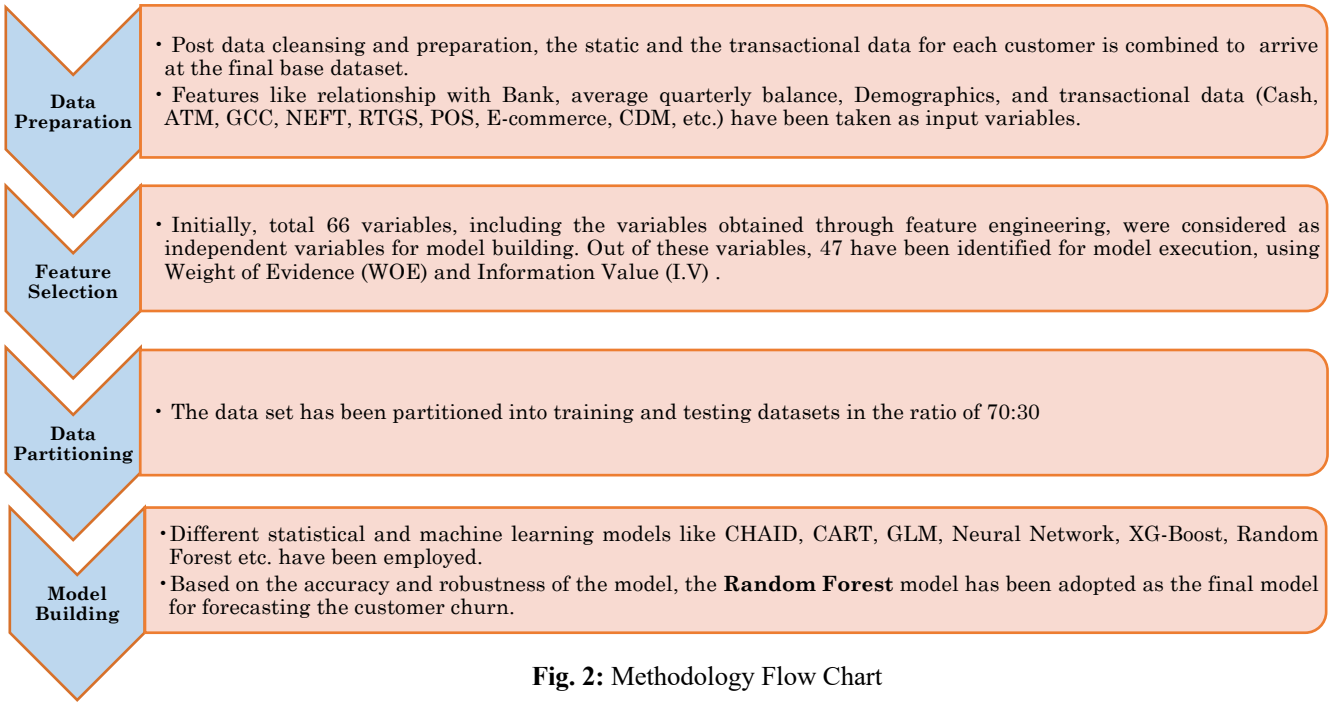

Fig. 2: Methodology Flow Chart

The further details of the whole modeling process are furnished ahead.

2.4. Data Visualization

Various business industries are consulting with analytics experts to extract insights from the enormous data available to businesses regarding their supply chain to help improve decision making. Descriptive Statistical Analysis helps us to understand the data and is an essential part of Machine Learning modeling; this is due to Machine Learning being all about making predictions. On the other hand, statistical analysis is all about drawing conclusions from data, which is a necessary initial step for developing predictive models. In any classification problem, descriptive statistics are used to describe the basic features of the data in a study.

Descriptive statistics are usually utilized for the following purposes:

1) Provide elementary information about variables in a dataset.
2) Identify a potential association between variables.
Graphical and pictorial methods are the most widely used tools for data visualization. Several graphical and pictorial methods are available that enhance the understanding of individual variables and the relationships between variables. In other words, pictorial and Graphical methods provide a visual representation of the data.
A few of descriptive graphs to assess the variable's association with churn rate has been furnished below:

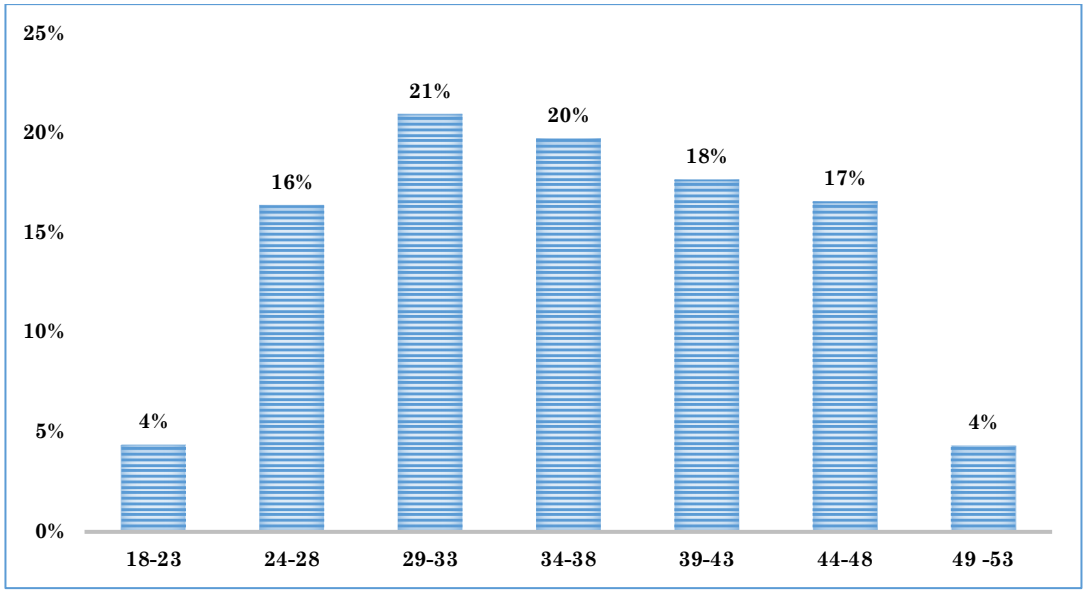

Fig. 3: Age-Wise Distribution of SB customer churn.

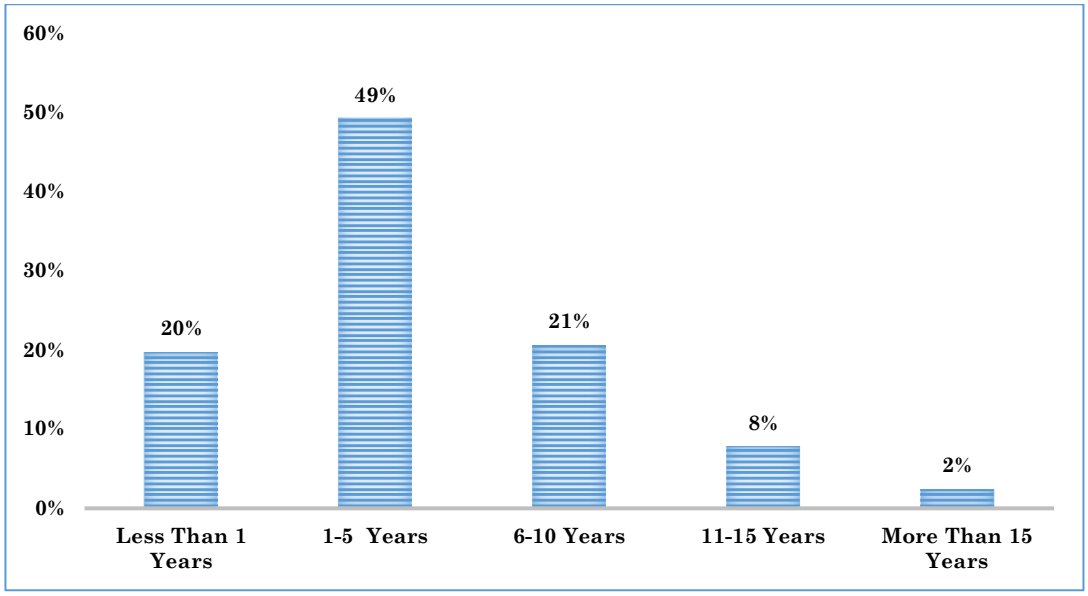

Fig. 4: Customer Vintage Wise Distribution of SB customer churn.

Figure 3 depicts that the highest proportion of churn (59%) lies within the age group 29-49 years. Further, *figure 4* shows that the accounts which are less than five years old, contribute to 69% of the total savings bank churn. Similarly, the descriptive statistics have been found for other features as well, to have an understanding of the churn rate among features.

2.5. Data Cleaning

We discover missing or corrupt data and think of various data cleaning operations to perform, such as marking or removing bad data and imputing missing data. For most of the features, a separate category for missing values has been created so that the model does not lose vital information.

2.6. Feature Selection

After the data visualization and data cleaning, Weight of Evidence (WOE) and Information Value (I.V) have been calculated

for all the desirable features. After that, the strong and moderate predictors, based on I.V, are escalated for further model building.

A brief description of WOE and I.V has been furnished below:

Prior to initiating a binary classification model, an obvious step is to perform variable screening and exploratory data analysis; this is the step where we get to know the data and prune variables that are either ill-conditioned or simply contain no information that will help us forecast the outcome. This is a preparatory step designed to assure that the approaches deployed during the final model-building phases are set up for success.

The I.V and WOE provide an excellent framework for variable screening and exploratory analysis for binary classification. WOE and I.V have been used extensively in the credit risk world for several decades, and the underlying theory dates back to the 1950s. However, it is still not widely used outside the credit risk world, especially in churn prediction.

Salient Features of WOE and I.V

WOE and IV enable one to:

I. Consider each variable's independent contribution to the outcome.

II. Detect linear and non-linear relationships.

III. Rank variables in terms of "univariate" predictive strength.

IV. Visualize the correlations between the predictive variables and the binary outcome.

V. Seamlessly compare the strength of continuous and categorical variables without creating dummy variables.

VI. Seamlessly handle missing values without imputation.

VII. Assess the predictive power of missing values.

I.V and WOE are closely related to concepts from information theory where one of the goals is to understand the uncertainty involved in predicting the outcome of random events given varying degrees of knowledge of other variables [14], [15], and [16]. Therefore, this is a perfect framework for variable screening and exploratory analysis for predictive model building.
[15] gave an expression for the entropy of a probability distribution to introduce Information theory. Entropy is often explained as the level of "disorder," but in the context of information theory, it is better to perceive it as one's level of uncertainty or randomness. In the literature, high entropy refers to high uncertainty and low entropy to low uncertainty.
Based on the similarity of dependent variable distribution, i.e. the number of events and non-events, the WOE helps to transform a continuous independent variable into groups or bins [17].
The WOE can be evaluated by the expression

$$\boldsymbol{WOE = ln\left(\frac{\%\ of\ Non-Events}{\%\ of\ Events}\right)}$$

In a business problem like churn modeling, churn events are often called Bads and Goods for non-churn events. Therefore, the WOE can also be expressed in another way as

$$\boldsymbol{WOE = ln\left(\frac{Distribution\ of\ Goods}{Distribution\ of\ Bads}\right)}$$

Considering the continuous independent variables, we have created bins (categories/groups), and then combined categories with similar WOE values and replaced categories with WOE values. Further, we have used WOE values rather than old input values in our models. Similarly, we have combined categories with similar WOE and then created a new class of a categorical

independent variable with continuous WOE values. In other words, we have used the WOE values rather than the raw categories in our models. Thus, the transformed variable will be a continuous variable with WOE values, and it is the same as any other continuous independent variable. We have clubbed the categories with similar WOE because the categories with similar WOE have almost the same event rate. In other words, the churn behavior of such categories is the same.
Information value is a widely useful technique to select strong predictors in a predictive model. It helps to rank variables based on their importance for predicting the random outcome.
The IV is calculated using the formula

$$\boldsymbol{I.V} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \{(\%\ \boldsymbol{of\ Non-Events}) - (\%\ \boldsymbol{of\ events})\} * \boldsymbol{WOE}$$

Where n is the number of categories or bins for an independent variable, and WOE is the value found for the i^{th} class of the independent variable.

According to [18], by convention, the values of the I.V statistic in credit scoring can be interpreted as follows:
If the I.V statistic is:

Less than 0.02, then the predictor is not useful for prediction (separating the events from the non-events).

a. 0.02 to 0.1, then the predictor has a weak relationship to the Non-events/Events odds ratio.

b. 0.1 to 0.3, then the predictor has a medium-strength relationship to the Non-events/Events odds ratio.

c. 0.3 to 0.5, then the predictor has a strong relationship with the Non-events/Events odds ratio.

d. > 0.5, then the predictor has an apprehensive relationship, which may result in over-fitting.

Using the above criteria, we have opted for 47 features to be used for model building.

2.7. Feature Engineering

We have assessed that some attributes have familiar distributions such as Gaussian or exponential. Thus, we have applied to scale, or log, or other transforms to obtain a better prediction power of the model. Several new features have been computed using the various functions of the raw features available in the case data set.

2.8. *Under-Sampling*

Post cleansing and feature engineering the summary of the churn data used for Training, Testing and Validation is as below;

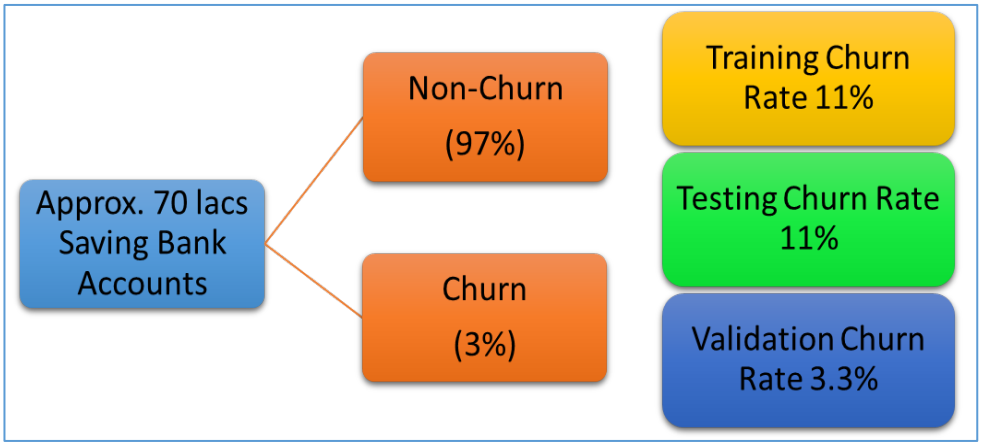

Fig. 5: Churn Rate among data partitions using under-sampling.

A weighted variable has been introduced in the data set to remove the bias caused by under-sampling.

2.9. *Model Selection*

Artificial Intelligence (AI) and machine learning (ML) are making the customer experience more personalized and contextual than ever before. Banks and credit unions are using advanced technology to build websites, emails, digital advertising, social media, and other content more efficiently and effectively; this is increasing marketing return of investment (ROI) as well as customer satisfaction and customer retention.

Considering the necessity to keep up with the competition, data science has become a tool to enhance the business in the banking industry. Indian commercial banks must understand that big data technologies can help them focus their resources efficiently, make sharp-witted decisions, and improve the performance of their products and services. In the current world of information technology and hyped overrated techniques, machine learning has a great future in terms of industrial research and business benefit. The potential of computer algorithms to learn on their own and improve over time generates new opportunities for industries across the board, especially for the banking industry.

The Churn Model has been built in two steps:

In the First Step, all predictors/features which are believed to be influencing customer attrition have been passed to the Logistic Regression, C 5.0, CHAID, ANN, XG-BOOST, and Decision Tree techniques.

In order to improve the efficiency of the model, we further have passed the important predictors, identified in the previous step, through the Random Forest algorithm to build a Machine Learning Model for prediction of customer attrition. Here, the Random Forest model has been trained to predict the binary dependent variable (i.e., either 'Churn' or 'Not Churn').

2.9.1. Random Forests

In many real-world applications of statistical modeling, the cost of a wrong decision may be higher than the advantage of having a high classification rate. Such kind of classification problems represents very resilient tasks because they require highly reliable systems or programming algorithm. To fulfill this need, many classification models have been proposed, among which, classifier ensembles exhibit a successful example. This approach focuses on combining classifiers making uncorrelated errors. In this context, the Random Forests (RF) represents a machine learning algorithm of distinctive interest. A random forests model is made of a suitable ensemble of several decision trees and has proved to be very useful in diverse fields by enhancing the results of decision trees. Random forests algorithm follows a procedure of averaging multiple deep decision trees, trained on various parts of the same training set, intending to control the over-fitting problem of the individual decision tree. In other words, a random forests algorithm is an ensemble learning technique for regression and classification that works by contriving a lot of decision trees at the training stage and provide an output class that is the mode of the classes output by individual trees in case of the classification problem. According to **[**19**],** the framework of the random forest has been immensely successful as a general-purpose classification and regression method.

The term "Random Forests" refers to a general methodology for building an ensemble of L tree-based classifiers $\{T(x; \Theta_k)\}$, where x is an input feature vector and Θ_k is a random vector that governs the growth of the kth tree. The random vector Θ_k is generated independently of the preceding $\Theta_1, \Theta_2 \ldots, \Theta_{k-1}$, but with the same distribution. In bagging, for instance, the random vector is $\Theta_k = \{\theta_1, \theta_2 \ldots, \theta_N\}$, where N is the training set size and θ_i is an integer value randomly drawn (with uniform distribution) from the set $\{1, 2 \ldots, N\}$. In the random subspace method, instead, Θ_k consists of a number of integers randomly drawn from a uniform distribution in the interval $[1; k]$; $k < M$, where M is the number of available features.

The Random Forest approach does not refer to a single algorithm, but rather to a family of methods. The original algorithm proposed by [19], is usually referred to in the literature as Forest-RI and is used as a reference method in most of the papers dealing with RF. For a more comprehensive review and mathematical operations, we refer the readers to [19] and [20]. Given a training set of size N, each consisting of M features, the forest-RI algorithm consists of the following steps:

1) **Random Record Selection:** n samples, at random with replacement, are drawn from the total training data set. The resulting data set will be the training data set of an individual decision tree. Usually, each tree is trained on roughly 2/3rd of the total training data.

2) **Random Variable Selection:** k features are selected at random out of all the predictor variables M and the best split on these features is employed to split the node. By default, k is the square root of the total number of features available in the dataset for classification.

3) Among the values of each of the *k* features drawn, the best binary split is chosen based on the Gini impurity or information gain [21]. Features with the best index values are selected.

4) The tree is grown to its maximum size according to the stopping criterion chosen.

Usually, the node splitting is stopped when one of the following conditions occurs:

i. The number of samples in the node to be split is below a given threshold.

ii. All the samples in the node belong to the same class.

5) Let the tree unpruned. In other words, it is recommended not to prune while growing trees for the random forests.

After the construction of random forests, a sample is labeled according to the Majority Vote rule, i.e. it is labeled with the most popular class among those provided by the ensemble trees. It is worthwhile to mention that in **[**22**],** it is proved that RF does not over-fit since more trees are added, rather its generalization error tends to a limiting value.

Random Forest does not require a **split sampling method (e.g., *n-fold* validation)** to assess the accuracy of the model. It performs internal validation as 2/3rd of available training data is used to grow each tree, and the remaining one-third portion of training data is always used to calculate out-of-bag error to assess model performance.

3. Results and Discussion

The final case data set of all features, including the weighted variable and corresponding churn flag, is divided into training and testing data sets with a ratio of 70:30. The training dataset is used to train different statistical/machine learning algorithms like *Logistic Regression*, *Decision Tree*, *Random Forests*, *ANN*, and *XG-BOOST* to predict the churn.

After trying the above algorithms with different model parameters, the Random Forests model turns out to be the most efficient model for churn prediction, based on the results from training, testing, and validation data sets. Using the under-sampling churn rate is raised to 11% for training and testing purposes. However, the model is validated on original data (Without balancing the churn), which has a 3.3% churn rate. The results for each of the algorithms are reflected in *table 1*. The table exhibits that the Random forests model shows the best churn prediction based on Accuracy, Specificity, and Sensitivity of the model. The other models do not provide significant prediction accuracy for identifying the churn. On the other hand, the XG-BOOST model performs well for testing data, however, it fails to make an accurate prediction of SB churn while validating the model.

Table 1: Model Comparison for Testing and Validation Data Sets.

Model Used	**Testing**			**Validation**		
	Accuracy	**Specificity**	**Sensitivity**	**Accuracy**	**Specificity**	**Sensitivity**
RANDOM FOREST	**78%**	**83%**	**74%**	**79%**	**83%**	**73%**
XG-BOOST	72%	70%	74%	47%	35%	60%
ANN	64%	53%	75%	52%	40%	65%
GLM	58%	35%	80%	54%	30%	78%

Figure 6 shows the Receiver Operating Characteristic (ROC) curve for the churn prediction through the Random Forests model. ROC curve is a performance evaluation tool for a predictive model at various defined thresholds settings. In a ROC curve, the true positive rate (Sensitivity) is plotted in function of the false positive rate (1-Specificity) for different cut-off points. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. A model with perfect discrimination between events and non-events has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore the closer the ROC curve to the top left corner, the higher the overall accuracy of the test **[**23**]**. The figure shows that the random forests model possesses a good discrimination power to predict the churn and non-churn.

Table 2 provides the value of Area under Curve (AUC) and Gini index for the finally opted model, Random Forest. The AUC is the area enclosed by the ROC curve. A perfect classifier has AUC = 1, and a completely random classifier has AUC = 0.5. Usually, a good model scores somewhere in between. The range of possible AUC values is [0, 1]. As per the model result

from *table 2*, the value of AUC is 0.85; it means there is an 85% chance that the model will be able to distinguish between positive class and negative class.

The Gini coefficient can be calculated using the formula:

$$Gini = 2 \times AUC - 1$$

The purpose of using the Gini index is to normalize the AUC so that a random classifier scores 0, and a perfect classifier scores 1. The range of possible Gini coefficient scores is [-1, 1]. The result suggests that the Gini coefficient for the Random Forests model is 0.69, which reflects a good discrimination power of the model.

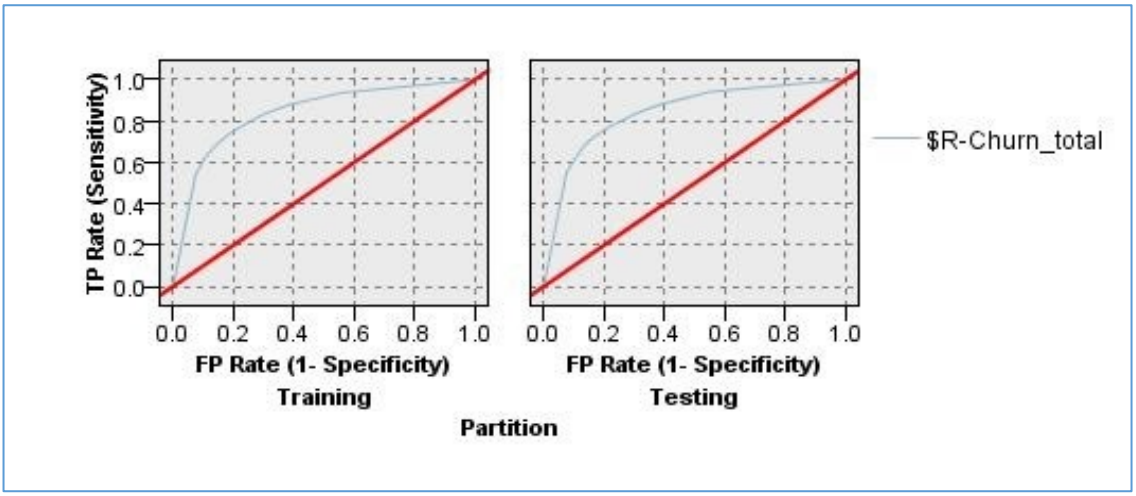

Fig. 6: Receiver Operating Characteristics (ROC) Curve for Random Forest

Table 2: AUC and GINI Index for the Random Forest Model.

Partition	**Training**		**Testing**	
Model	AUC	Gini	AUC	Gini
Random Forest	0.845	0.690	0.844	0.688

Based on the above-discussed criteria, it has been observed that the results obtained from **Random Forests,** are most robust for the scenario.

3.1. *Out of Time Validation*

This model has been validated using fresh (Out of Time) data from 1st July 2017 to 30th Jun 2018. This validation data contain 3.3% of customers as churn.

The model validation (Out of time) results are furnished in *table 3*; the result exhibits that the balanced accuracy of the Random forests has been improved by 1 % for the out of time validation. This could be due to the less variability and noise in the 'out of time validation' data, as compared to the testing data sets. These results may give a motivation to consider the further recalibration of the proposed model to enhance the accuracy and other predictive parameters.

Table 3: Model Performance for out of time validation data.

Accuracy	**Specificity**	**Sensitivity**
80%	81%	79%

3.2. *Most Significant Features*

Figure 7 reflects the most significant variables to predict SB customer churn. It has been found that customer's age, customer vintage, the average balance of the account, and other mentioned variables are the strongest predictors for the

random forest model to predict the churn of SB customers.

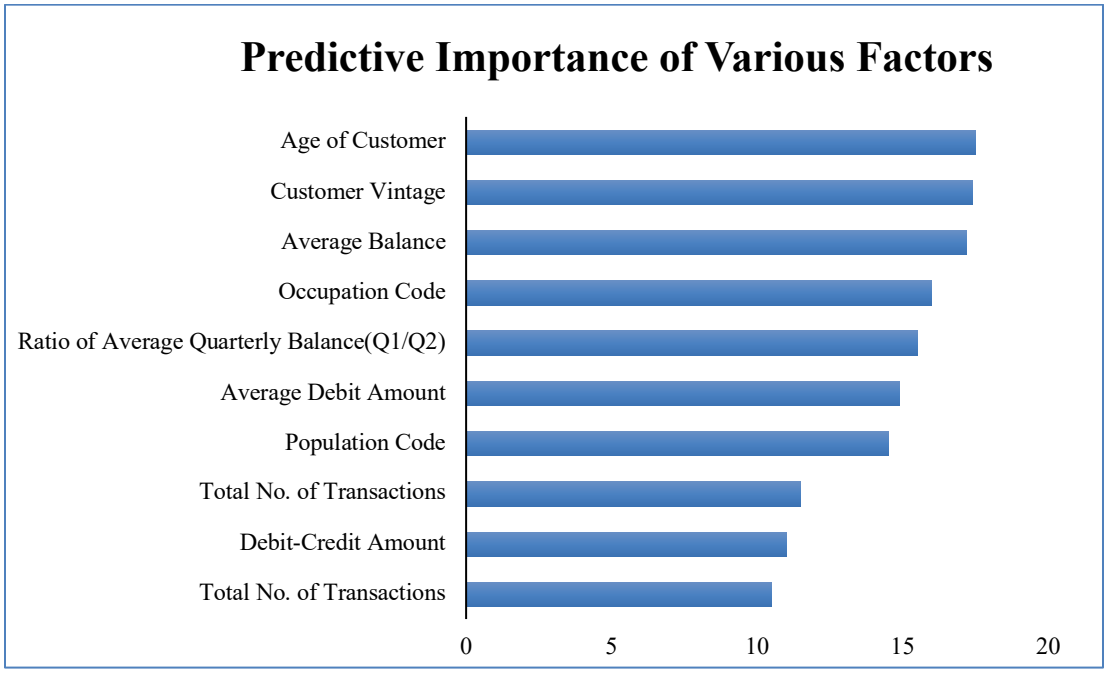

Fig. 7: Most Significant Features for the Random Forest Model.

4. Conclusions

As we have discussed earlier that the retention of customer costs significantly lower than acquiring a new customer, a churn prediction model has been developed to avoid SB customer churn. Several statistical and machine learning models have been employed on the bank level SB customer data. Thereafter, the accuracies for various models have been calculated, and the same are compared using the ROC curve, Gini index & AUC, to opt the best model for the prediction. Further, having a glance at the model comparisons, it can be concluded that the Random Forests model has a better prediction performance among all the suggested models for the given scenario of SB customers.

The commercial banks can deploy the model in order to avoid customer churn, and they may retain the deposits and advances, which are kept by SB customers. The article suggests a customized campaign to be initiated by commercial banks to avoid SB customer churn. Hence, by giving better customer satisfaction and experience, the commercial banks can limit the customer churn and maintain their customer base and CASA ratio.

Acknowledgments

The author expresses his profound gratitude to Mr. Anup Kumar Mahapatra (CGM, ESS, GITC, SBI), Mr. Kunjal Prasad (Former GM, Data & Analytics, GITC, SBI), Mr. Sandipan Sen (GM, Data & Analytics, GITC, SBI), Mr. N.D.S.V. Nageswara Rao (Former DGM, Analytics, GITC, SBI), Mr. Sumanta Kumar Panda (DGM, Analytics, GITC, SBI), Mr. Kamal Kishor Naik (Chief Manager, Analytics, GITC, SBI), Mrs. Priyanka (Manager, Analytics, GITC, SBI) for their valuable suggestions. The author is thankful to Ms. Vrushali Gunjal and Mr. Anup Pillai for their valued support while data extraction and model building. The author also appreciates all the feedbacks received from the editors and the reviewers.

Abbreviations

Area under the curve (AUC)
Artificial Intelligence (AI)
Mobile Banking Services (MBS)
Point of Sale (POS)

Artificial Neural Network (ANN)
Cash Deposit Machine (CDM)
Current Account (CA)
Extreme Gradient Boosting (XG-BOOST)
Green Channel Counter (GCC)
Information Value (I.V)
Machine learning (ML)
Random Forest (RF)
Receiver Operating Characteristics (ROC)
Return of Investment (ROI)
Savings Accounts (SA)
Savings Bank (SB)
Support Vector Machine (SVM)
Weight of Evidence (WOE)

Conflict of interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] He, B., Shi, Y., Wan, Q., Zhao, O. Prediction of Customer Attrition of Commercial Banks based on SVM Model. *Procedia Computer Science*, **31**, 423-430, 2014. https://doi.org/10.1016/j.procs.2014.05.286.
[2] Farris, P.W., Bendle, NT, Pfeifer, PE, Reibstein, DJ. *Marketing Metrics: The Definitive Guide to Measuring Marketing Performance*. Pearson FT Press, New Jersey, USA: 2nd, (p. 432), 2010.
[3] Bhattacharya, C. B. When customers are members: Customer retention in paid membership contexts. *Journal of the Academy of Marketing Science*, **26(1)**, 31- 44, 1998.
[4] Gallo, A. The Value of Keeping the Right Customers. *Harvard Business Review*, U.S./Canada (Oct 2014).
[5] Chandar, M., Laha, A., and Krishna, P. *Modeling churn behavior of bank customers using predictive data mining techniques*, in Proc. National Conference on Soft Computing Techniques for Engineering Applications (SCT-2006), **(3)**, 24-26, 2006.
[6] Popovic, D. and Basic, B.D. Churn Prediction Model in Retail Banking Using Fuzzy C-Means Algorithm. *Informatica*, **(33)**, 243-247, 2009.
[7] Prasad, D. and Madhavi, S. Prediction of Churn Behavior of Bank Customer Customers Using Data Mining Tools. *Business Intelligence Journal*, **5(1)**, 96-101, 2012.
[8] Bahnsen, A.C., Aouada, D, and Ottersten, B. A novel cost-sensitive framework for customer churn predictive modeling. *Decision Analytics,* 2:5, 2015.DOI 10.1186/s40165-015-0014-6.
[9] Burez, J. and Poel, D.V. Handling Class Imbalance in Customer Churn Prediction. *Expert System with Applications*, **(36)**, 4626-4636, 2009.
[10] Xie, Y.Y., Li, X., Ngai, E.W.T., and Ying Weiyun. Customer Churn Prediction Using Improved Balanced Random Forests. *Expert Systems with Applications*, **(36)**, 5445-5449, 2009.
[11] Chiang, D., Wang, Y., Lee, S., and Lin, C. Goal-oriented sequential pattern for Network Banking Churn Analysis. *Expert Systems with Applications*, **25(3)**, 293-302, 2003.
[12] Mutanen, T., Ahola, J. and Nousiainen, S. Customer Churn Prediction-A Case Study in Retail Banking. *ECML/PKDD2006 Workshop*, 13-18, 2006.
[13] Zhao, J. and Dang, X.H. Bank Customer Churn Prediction Based on Support Vector Machine: Taking a Commercial Bank's VIP Customer Churn as the Example. *IEEE*, 1-4, 1998.
[14] Kullback, S. *Information Theory and Statistics*. John Wiley & Sons, USA, 1959.
[15] Shannon, C.E. A Mathematical Theory of Communication, *Bell System Technical Journal*, Vol. 27 (3), pp. 379–423, 1948.
[16] Shannon, C.E. and Weaver, W. *The Mathematical Theory of Communication*. Urbana, University of Illinois Press, 1949
[17] Bhalla, D. Weight of Evidence and Information Value Explained. *Listen Data*, 2015.
[18] Berkel, A.V and Siddiqi, N. *Building Loss Given Default Scorecard Using Weight of Evidence Bins in SAS®Enterprise Miner™ Loss Given Default*. Conference: SGF 2012. https://www.researchgate.net/publication/276956927_Building_Loss_Given_Default_Scorecard_Using_Weight_of_Evidence_Bins_in_SASR_Enterprise_Miner_Loss_Given_Default/citations.
[19] L. Breiman. Random Forests. *Machine Learning*, 45(1), 5–32, 2001.
[20] Criminisi, A., Shotton, J, and Konukoglu, E. Decision forests: A unified framework for classification, regression, density estimation, manifold learning, and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, **7(2-3)**, 81–227, 2011.
[21] C. Gini. Measurement of inequality of incomes. *The Economic Journal*, **31(121)**, 124–126, 1921.
[22] L. Breiman. Bagging Predictors. *Machine Learning,* **24(2)**, 123–140, 1996.

[23] Zweig, M.H and Campbell, G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, **39(4)**, 561-577, 1993.

Prashant Verma is working as Statistician at Analytics Department, Global IT Center, SBI, Navi Mumbai, India, he has completed his Ph.D. from the Department of Statistics, Banaras Hindu University, Varanasi, India. His research interests are in the areas of Business Analytics, Public Health, Population Mathematics, Demography and Reproductive Health. He has been awarded for Junior Research Fellowship-2013, Basic Scientific Research Fellowship-2013, and Senior Research Fellowship-2016, sponsored by University Grant Commission. He has 1 national and 11 international publications with reputed publishers like Springer Open, and BioMed Central. He has published a chapter in the book "Essentials of Statistics in Agriculture Sciences", worldwide distributed by CRC Press (Taylor & Francis Group), 2019. He has received a certificate of appreciation from International Conference on Family Planning (ICFP-2015, Indonesia), Bill & Melinda Gates Institute for Population and Reproductive Health, Johns Hopkins Bloomberg School of Public Health, for serving as an abstract reviewer. Along with 12 national conferences/workshops presentations, he has presented research articles in 7 International conferences including 7th African Population Conference - 2015, Johannesburg, South Africa and 2nd International Conference on Theory and Application of Statistics-2015 at Dhaka University, Bangladesh. One of his articles entitled as "An indirect assessment of Infant mortality estimates in India: A Modified Approach" has been selected for the presentation at the 5th Human Mortality Database Symposium - 2019 "Recent Trends and Future Uncertainties in Longevity", Berlin, Germany. He has also been invited to organize a session in 8th African Population Conference-2019. He has received two travel awards, one from Indian Council of Medical Research (ICMR), New Delhi (AIIMS) to attend an international conference in Johannesburg, South Africa (2015) and second from Indian Council of Social Science Research (ICSSR), New Delhi to attend the 4th Asian Population Conference at Shanghai, China (2018).