Grenze International Journal of Engineering and Technology, June Issue

Machine Learning based Customer Churn Prediction with Random Forest Algorithm

Kalyani Nivrutti Mahajan[1], Dr. Ram Joshi[2] and Prof. K.V. Deshpande[3]
[1-3]JSPM's Rajarshi Shahu College of Engineering, Tathawade,Pune,Maharashtra,India – 411033
Email: kalyanimahajan.1398@gmail.com, rbjoshi_it@jspmrscoe.edu.in, kvdeshpande_comp@jspmrscoe.edu.in

***Abstract*—The basic random forests are combined with a sampling method, cost sensitive learning, and improved balanced random forests (IBRF), which utilized to predict client churn. We may repeatedly learn the best features of IBRF by changing the distribution of the classes and toughening the punishment for misclassifying the minority class. They are demonstrated to considerably increase prediction accuracy when compared to other methods such as artificial neural networks, decision trees, and class-weighted core support vector machines when applied to a database of credit debt clients from an anonymous commercial bank in China (CWC-SVM). To evaluate these algorithms' qualities, they are compared and graded. The sampling plan and data processing are explained in great detail.**

***Index Terms*— Support vector machine, Random Forest classifier, prediction, Churn.**

I. INTRODUCTION

This essay will examine customer churn, one of the most significant problems now affecting businesses and a popular topic in CRM. Customer churn, or the likelihood that customers will stop doing business with a company within a predetermined period of time, has become a significant issue for many businesses, including those in the publishing, investment, insurance, electric utility, health care, credit card, banking, Internet, telephone, and cable services sectors. Customer turnover, of course, has a direct impact on customer retention rates and, thus, a customer's lifetime worth to a business.
It is easy to discover that devoted consumers contribute the majority of a company's earnings and that acquiring new customers is more expensive than keeping existing ones by looking at the present of a customer's lifetime profit to a business. So, it's essential for businesses to keep up their relationships with current clientele.

II. ALGORITHM RANDOM FOREST

To address classification and regression problems, the Random Forest Algorithm, a very well- liked supervised machine learning method, is used. A forest is made up of numerous different species of trees, and the forest will be more vigorous the more trees there are. Similar to this, the number of trees in a Random Forest Algorithm increases the algorithm's accuracy and ability to solve problems. A classifier known as Random Forest employs numerous decision trees on various subsets of the input data to boost the predicted accuracy of the dataset. It is based on the concept of ensemble learning, which is the process of combining different classifiers to solve a difficult problem and improve the performance of the model.

Grenze ID: 01.GIJET.9.2.553

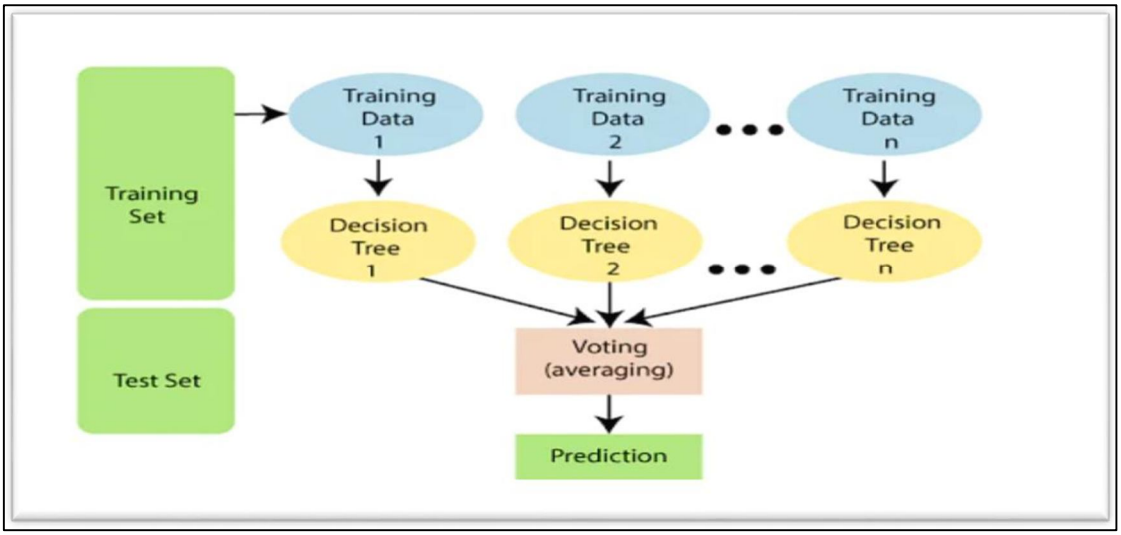

Fig 1 Random Forest

Step 1: Choose a random sample from a specific training set or data collection.
Step 2: This algorithm will create a decision tree for each training set of data.
Step 3: Voting will be done using the decision tree's average
step 4: Choose the prediction outcome that garnered the most support as the final prediction outcome.
This combination of different models is referred to as an ensemble. Ensemble uses two methods:

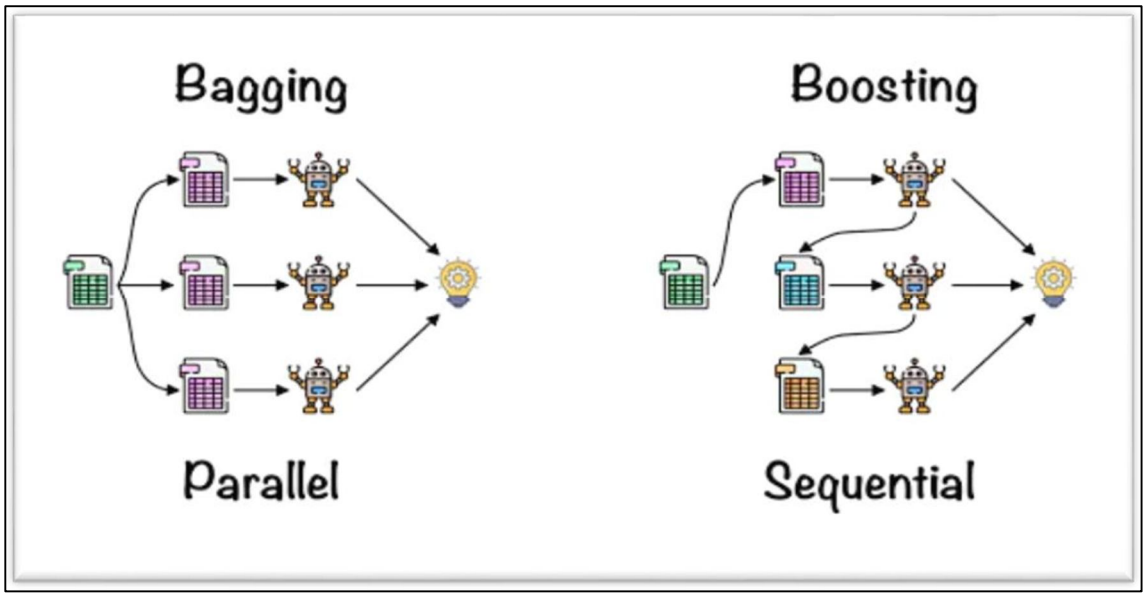

Fig 2 Bagging And Boosting

Bagging**:** A distinct training subset is created by the process of bagging using a sample training dataset. A majority vote is required to determine the result.
Boosting**:** is the process of transforming weak learners into strong ones through the development of subsequent models with the aim of reaching the highest level of accuracy. Examples include ADA BOOST and XG BOOST.
Bagging: Given the aforementioned theory, Bagging algorithm is used by Random forest is concluded. So let's look more closely at this concept. Bagging in random forests is also known as Bootstrap Aggregation. The process starts with any initial random data. Data is then arranged into samples known as Bootstrap Samples. This process is known as bootstrapping. Each model is additionally trained independently, resulting in unique results known as Aggregation. The results are combined in the final stage, and a majority vote is used to determine the outcome. An Ensemble Classifier is used at the process' Bagging stage.

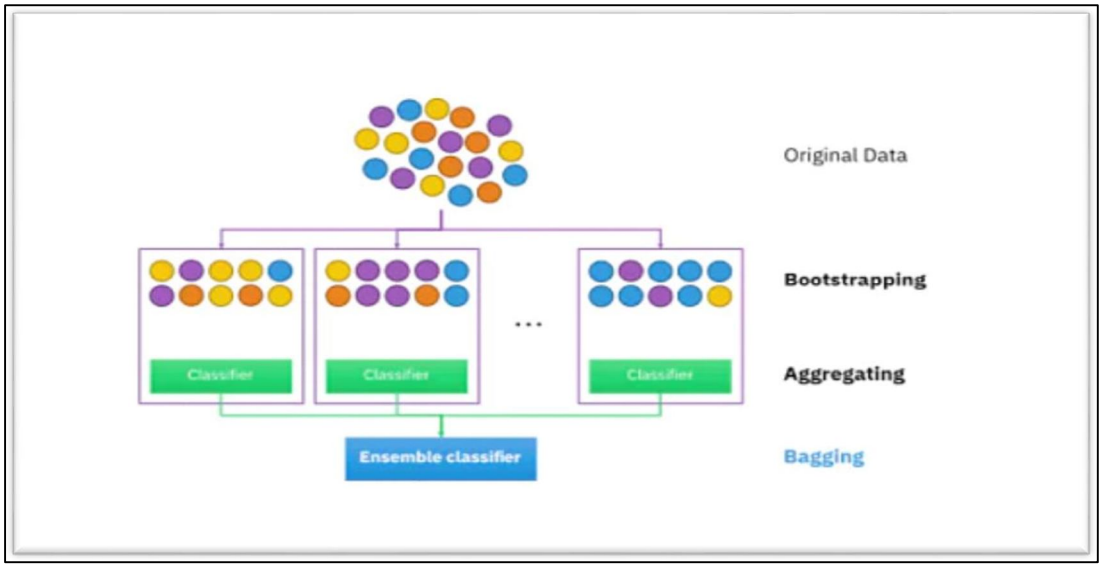

Fig 3 Bagging

III. MATHEMATICAL MODEL

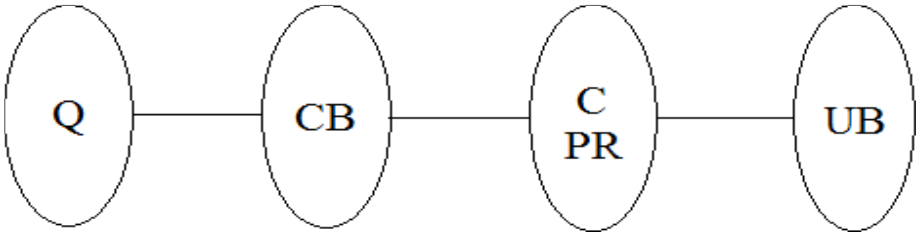

Fig 4 Mathematical Model

- Where Q denotes a user-entered query, such as a URL, and CB denotes preprocessing.
- C stands for "feature selection."
- PR = Ensemble algorithm is used.
- UB stands for "predict outcome."
- Theorem of Sets
- Assume S is a system that receives vocal data
- S = In, P, Op, S = S = S = S = S = S = S = S = S = S = S.
- Recognize Input as well as
- if in = q
- Where,
- Q = Input from the user (vocal dataset)
- Process P should be identified as
- CB, C, PR = P
- Where,
- CB = System checks whether or not the entered list is PD infected.
- C = Copy the URL (universal resource locator).
- PR = Ensemble algorithm is used.
- Designate Output Op as
- UBOP = UBOPOPOPOPOPOPOPOPOPOPOPOPOP
- Where,
- UB =predict the result

=Conditions of Failure and Success=

- Failures:
- A large database might increase the amount of time it takes to get information.
- Failure of the hardware.
- Failure of the software.
- Success:
- Datasets can be used to find the information you need.
- The user receives a quick response
- Space Complexity: The space complexity is determined by how the identified patterns are presented and visualised. The more data is stored, the more room is required.
- Complexity of Time:
- Check the number of patterns in the datasets=n

If (n>1), retrieving information can take a long time. As a result, this algorithm has an O(nn) time complexity.

Block Diagram

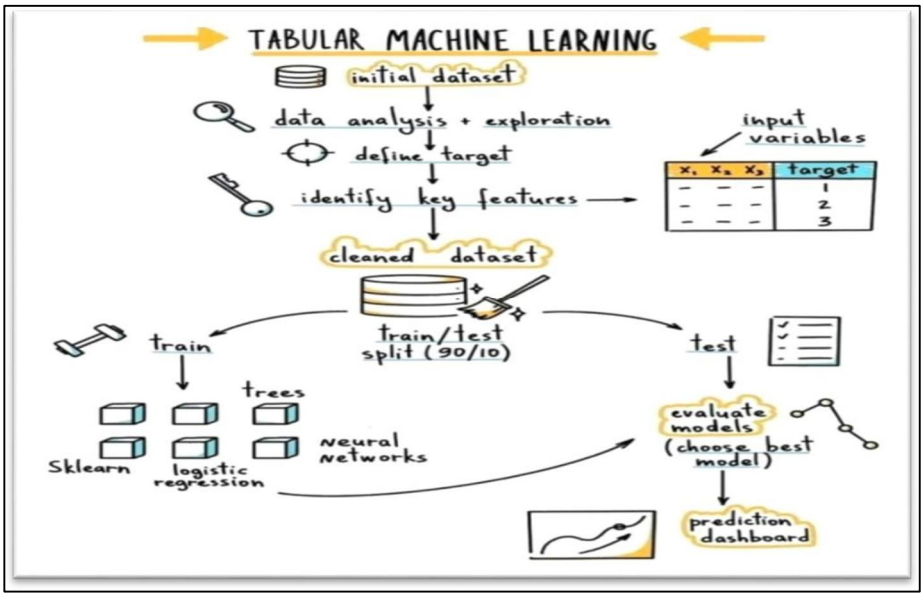

Fig 5 Block Diagram

IV. CHALLENGES

1 It's challenging to create a dataset from a consumer
2 It can be difficult to keep track of phone reservations & cancellations.
3 It can only be used with one operating system.

V. RESULTS

```
Accuracy_Train=((3432+498)/(4930)*100)
print(Accuracy_Train)
```

```
79.71602434077079
```

```
from sklearn.metrics import classification_report
print(classification_report(train['Churn'], train['Predicted']))
```

```
              precision    recall  f1-score   support

           0       0.80      0.95      0.87      3611
           1       0.75      0.36      0.49      1319

    accuracy                           0.80      4930
   macro avg       0.78      0.66      0.68      4930
weighted avg       0.79      0.80      0.77      4930
```

```
Accuracy_test=((1471+191)/(2113)*100)
Accuracy_test
```

```
78.65593942262187
```

```
from sklearn.metrics import classification_report
print(classification_report(y_test, test['Predicted']))
```

```
              precision    recall  f1-score   support

           0       0.80      0.94      0.87      1563
           1       0.67      0.34      0.45       550

    accuracy                           0.79      2113
   macro avg       0.74      0.64      0.66      2113
weighted avg       0.77      0.79      0.76      2113
```

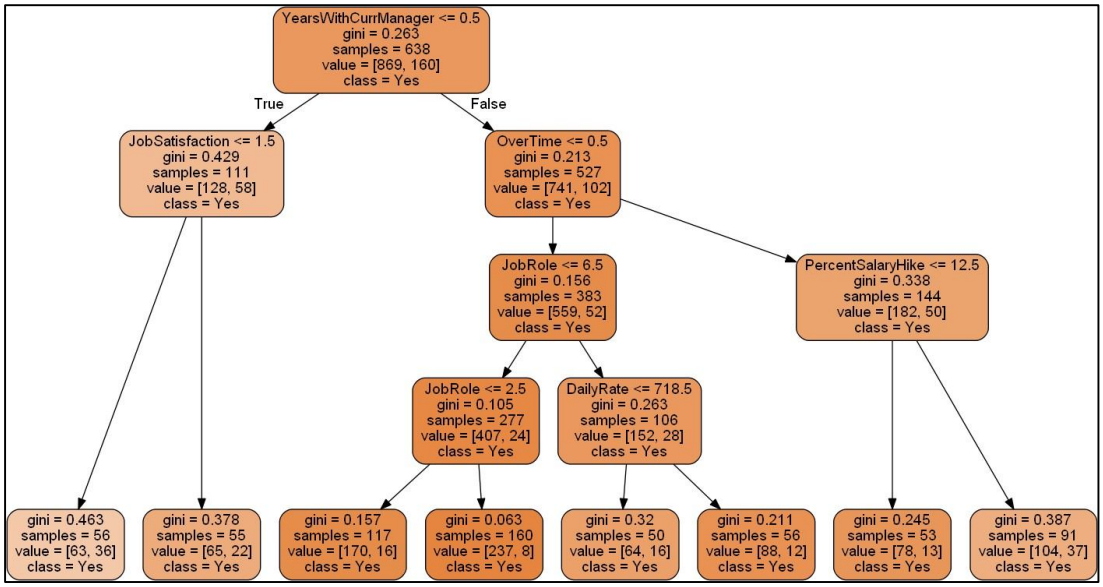

Fig 6 Design Tree

VI. Conclusion

This article provides a novel method for predicting banking industry churn termed IBRF. By combining a sampling technique with cost-sensitive learning, IBRF has the advantage of altering the class distribution and punishing minority class misclassification more severely. The best weak classifiers are created by artificially equating the class priors, from which the best features are then iteratively trained. According on experimental results on bank databases, our method performs better than earlier random forest algorithms like balanced random forests and weighted random forests. The top-decile lift of IBRF is higher than that of ANN, DT, and CWC- SVM. IBRF has a lot of promise when compared to traditional approaches due of its scalability and quicker training and running rates. Effectiveness and generalizability should be improved by recent studies. Internal variables are used by IBRF to examine sample distribution. Despite the fact that it has been shown that the conclusions are not sensitive to the values of these variables, lowering their values in subsequent tests may improve the method's capacity to predict outcomes more precisely.

Acknowledgement

This research was supported by JSPM's Rajarshi Shahu College Of Engineering, Tathawade, Pune .The author is thankful to the guide Dr. Ram Joshi , Prof. K .V . Deshpande for providing guidelines for this research.

References

[1] S Neslin, S Gupta, W Kamakura, J Lu, C Mason, "Defection Detection: Improving Predictive Accuracy of Customer Churn Models", Working Paper, Teradata Center at Duke University, 2004.3

[2] Wells, Melanie, “Brand ads should target existing customers” Advertising Age, pp26-47, 1993.
[3] CP Wei, IT Chiu, “Turning telecommunications call details to churn prediction: a data mining approach”, Expert Systems with Applications, Vol. 23, Issue 2, pp103-112, 2002.
[4] Mozer, M. C., Wolniewicz, R., Grimes, D. B., etc, “Churn reduction in the wireless industry”, Advances in Neural Information Processing Systems, pp935-941, 2000(12).
[5] Eiben, A.E.; Koudijs, A.E.; Slisser, F. Source, “Genetic modeling of customer retention”,LectureNotes in Computer Science, Vol. 1391, pp178, 1998
[6] J. R. Quinlan, “Decision trees as probabilistic classifiers”, Proc. 4th Int. Workshop Machine Learning, pp31–37, Irvine, CA, 1987.
[7] Wai-Ho Au, Keith C. C. Chan, and Xin Yao, “A Novel Evolutionary Data Mining Algorithm with Applications to Churn Prediction”, IEEE Transactions on Evolutionary Computation, Vol. 7, Issue 6, pp532-545, 2003
[8] LUO Ning, MU Zhichun, “Bayesian Network Classifier and Its Application in CRM”, ComputerApplication, Vol. 24, No. 3, pp79-81, 2004
[9] Y Zhao, B Li, X Li, W Liu, S Ren, “Customer Churn Prediction Using Improved One-Class Support Vector Machine”, Lecture Notes in Computer Science, Vol. 3584, pp300-307, 2005
[10] Lariviere, B., Van den Poel, D. (2004). Investigating the role of product features in preventing customer churn by using survival analysis and choice modeling: The case of financial services. Expert Systems with Applications, 27, 277–285
[11] Breiman, L. (2001). Random Forests. Machine Learning, 45, 5–32.
[12] Lariviere, B., Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. Expert Systems with Applications, 29, 472–484
[13] Liaw, A., Wiener, M. (2002). Classification and regression by random forest. R News,
[14] 2(3), 18–22. [14] Chen, C., Liaw, A., Breiman L. (2004). Using
[15] random forests to learn imbalanced data, Technical Report 666. Statistics Department of University ofCalifornia at Berkeley.
[16] Dekimpe, M. G., & Degraeve, Z. (1997). Theattrition of volunteers. European Journal of Operational Research, 98(1), 37–51.
[17] Rust, R. T., & Metters, R. (1996). Mathematical models of service. European Journal of Operational Research, 91(3), 427–439.
[18] Mozer, M. C., Wolniewicz, R., Grimes, D.B., Johnson, E., Kaushansky, H. (2000). Predictingsubscriber dissatisfaction and improving retention in the wireless telecommunication industry. IEEE Transactions on Neural Networks, Special issue on Data Mining and Knowledge Representation, 690–696
[19] Neslin, S., Gupta, S., Kamakura, W., Lu, J., Mason, C. (2004). Defection detection: improving predictive accuracy of customer churn models. Working Paper, Teradata Center at Duke University.
[20] S. Babu, D. N. Ananthanarayanan, and V. Ramesh, ‘‘A survey on factors impacting churn in telecommunication using datamininig techniques,’’ Int. J. Eng. Res. Technol., vol. 3, no. 3, pp. 1745–1748, Mar. 2014.
[21] C. Geppert, ‘‘Customer churn management: Retaining high-margin customers with customer relationship management techniques,’’ KPMG & Associates Yarhands Dissou Arthur/Kwaku Ahenkrah/David Asamoah, 2002. Dataset: https://www.kaggle.com/c/customer-churn-prediction-2020/data