International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 5, pp 114-119, May-2023

https://doi.org/10.47001/IRJIET/2023.705013

An Integrated Bank Customer and Credit Card Holder Churn / No Churn Analysis System Using Machine Learning

[1]Vandit Talwadia, [2]Shubh Kumar Jain, [3]Lavesh Chanchawat, [4]Suma Keerthana Pepeti

[1,2,3,4]Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, India

***Abstract* - Customer churn is a well-known issue in most sectors, there are various reasons why customers leave banks such as customers aren't getting the results they want, customer service needs improvement, customers believe your competitors can do a better job, customers no longer see the value in bank's product, and customers believe your product is too expensive or too cheap. Hence it's critical to develop a perfect predictive model designed in support of customer churn that could be used to formulate a customer retention strategy. This topic is much more important in markets where competition is high and acquiring new customers is more difficult than retaining existing customers. There is a need to build a Bank customer churn/no churn analysis system with best performance and accuracy that will be of great use in banking enabling them to prevent revenue loss, reduce marketing and sales costs and improve the quality of customer service. Coupled with Business Intelligence (BI) tools like Tableau, business executives can make sense of big data. The research focuses on using the machine learning model random forest to help the machine learning model evaluation and interpretability for customer churn analysis in having bank account and are associated with the credit card services, even though multiple models are employed for this analysis.**

Keywords: Churn, Prediction, Machine Learning, Random Forest Classifier, Tableau, Accuracy.

I. INTRODUCTION

Customer churn is one of the biggest and most difficult problems for businesses like credit card companies. While not the most exciting thing to look at, customer churn metrics can help businesses improve customer retention. There is involuntary turnover, such as when a customer is unable to pay their credit card bill and no longer stays with the credit card company. Reasons for customer churn can vary and will require domain knowledge to pinpoint, but some of the most common reasons are lack of product usage, poor service, and better pricing in other places. Whatever the reason may be specific to different industries, one thing that applies to every industry is that the costs of acquiring new customers are higher than retaining existing ones. The best way to avoid customer churn is to know your customers, and the best way to know your customers is to use historical data and new customers. Identifying leaving customers will help management categorize customers who are likely to leave early and target customers with promotions, as well as provide insight into factors to consider when customers are loyal. While different models are used to analyze customer churn, the project focuses on using the machine learning model's random forest to support customer model evaluation and interpretability row. Machine learning to analyze customer churn.

II. LITERATURE REVIEW

1. Personalized Customer Churn Analysis with Long Short Term Memory (1). Ahmet Tuğrul Bayrak Predicting churn customers in the fast food industry using LSTM algorithm. The ascending competitive environment leads fast-food customers to select among the available products corresponding to personal preferences. Therefore, most of the customers in the fast-food sector may not be considered potential loyal customers. Accuracy of the trained model when Random Forest Regression is found to be 67.76 % only and SVM model gave the accuracy of 70.85% which are very less to rely on.

2. Machine Learning Based Customer Churn Prediction In Banking (2). Manas Rahman, V Kumar, The study only used a small amount of data, and also highly imbalanced. But real commercial bank data would be much larger. By oversampling, both of these headaches up to a certain degree can be resolved. The model examined KNN, SVM, Decision Tree, RF classifiers under different conditions for this study. A better result is achieved when using the RF classifier together with oversampling (95.74%). This research paper is based on the project which has used small dataset which doesn't align to real bank dataset. Huge amount of data should be used to build the model and cover all possible scenarios.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 5, pp 114-119, May-2023

https://doi.org/10.47001/IRJIET/2023.705013

3. Robust Model for Churn Prediction Using Supervised Machine Learning. (3) Anurag Bhatnagar, In this paper, a telecom company dataset is used to predict customer churn. KNN and Logistic Regression of the supervised machine learning algorithm were applied after cleaning the dataset. The concluded accuracy in the research paper is less as compared to accuracy in other research papers. Techniques can be used to improve this accuracy score. Using KNN supervised learning; accuracy score was 88.5% and 86.5% when Logistic regression was used.

4. Profit Optimizing Churn Prediction for Long-Term Loyal Customers in Online Games (4). Eunjo Lee, This paper proposes a churn prediction process considering the expected profit of the online game by referring to the existing research methods and applies them to the live game that has been in service for over nine years to verify its effectiveness. Low accuracy of the supervised models used is GBM and XG Boost. Paper stares that using GBM, accuracy score was 37.89 and when XG Boost was used it was 38.66.

5. Prediction of customer attrition of commercial banks based on SVM model (5). Benlan Hea, Yong Shic, Qian Wand, Xi Zhaoc By changing the sample distribution, Random sampling method has a higher degree of recognition. Therefore, we use random sampling method to improve SVM method, and select F measure to evaluate the predictive power Due to the imbalanced characteristics of the actual churn dataset, SVM model cannot predict the churners effectively and only general evaluation criteria cannot measure the predictive power of the model.

6. A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector (6). Irfan Ullah, Basit Raza, in this paper, the researchers have been looking at the key factors of churn to retain customers and solve the problems of CRM and decision maker of a company. In this study, a customer churn model is provided for data analytics and validated through standard evaluation metrics. The obtained results show that our proposed churn model performed better. Random Forest and J48 produced better F-measure result that is 88%. One of the drawback of this algorithm is visualization becomes difficult.

III. SYSTEM ARCHITECTURE

First of all, we gathered the dataset and performed the required cleaning techniques followed by customer classification which will help in determining the model that can be used.

The end user (Administrator of the bank database) will register himself and will login into the system with his credentials, he must enter the OTP received on his registered email id, after login he will be required to enter the details of the customer to evaluate the churn status of the customers who are accountholders in the bank and customers using credit card service of the bank.

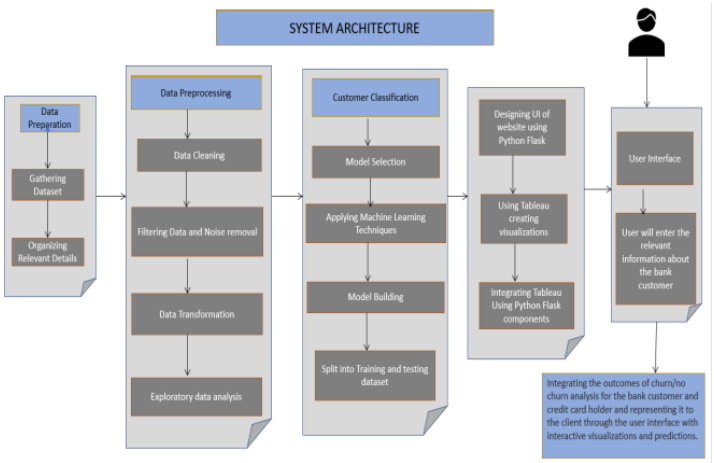

Figure 1

Admin can download the Churn Report generated for the customer and can navigate to Tableau storyboard for interactive visualizations based on different correlation between features. Figure1 explains the complete architecture.

IV. METHODOLOGY

1. Data Preparation:

Datasets from various sources containing attributes like Customer salary, credit score, age and more factors for Bank Customer churning and Credit Card holder churning will be the important requirements of our project.

2. Data Cleaning / transformation:

Here we will be fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within our dataset. Exploratory Data Analysis (EDA), here we will be discovering patterns, spotting anomalies, working on test hypothesis.

3. Feature Selection:

When creating a predictive model, the process of feature selection involves lowering the number of input variables. In some circumstances, reducing the number of input variables might enhance the efficiency of the model while also lowering the computing cost of modelling.

Figure 2 depicts the selected features for our model

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 5, pp 114-119, May-2023

https://doi.org/10.47001/IRJIET/2023.705013

```
Data columns (total 15 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Dependent_count           10127 non-null  int64
 1   Total_Relationship_Count  10127 non-null  int64
 2   Months_Inactive_12_mon    10127 non-null  int64
 3   Contacts_Count_12_mon     10127 non-null  int64
 4   Credit_Limit              10127 non-null  float64
 5   Total_Revolving_Bal       10127 non-null  int64
 6   Avg_Open_To_Buy           10127 non-null  float64
 7   Total_Amt_Chng_Q4_Q1      10127 non-null  float64
 8   Total_Trans_Amt           10127 non-null  int64
 9   Total_Trans_Ct            10127 non-null  int64
 10  Total_Ct_Chng_Q4_Q1       10127 non-null  float64
 11  Avg_Utilization_Ratio     10127 non-null  float64
 12  Age                       10127 non-null  int32
 13  Bank_Relationship_Period  10127 non-null  int32
 14  Attrition_Flag            10127 non-null  int64
```

Figure 2

4. Model Selection:

Based on the different accuracies from different machine learning classification models, the one with greatest accuracy will be selected to be the ML model for the project. From the survey we observed that using the Random Forest Machine learning model along with other techniques we can improve the accuracy score and performance for better prediction outcomes and analysis.

5. Designing UI of our webpage using Flask:

Flask is an open source app framework in Python language. It helps us create web apps for data science and machine learning in a short time. It is compatible with major Python libraries such as scikit-learn, Keras, PyTorch, SymPy (latex), NumPy, pandas, Matplotlib etc.

6. Using Tableau – creating visualization on available data:

Using Tableau, we are going to create an interactive dashboard which will help to infer patterns from the data and analyse it.

7. Integrating tableau with Flask (Integration Feasibility):

Steps to integrate tableau dashboard with Flask:

- Copy the Tableau Public Dashboard embedded code.
- Use Flask Components.

After importing Flask and Flask components, input the copied embed code into the template file.

Deploy the Flask app to a Server.

8. Final Integration:

Finally integrating the outcomes of churn/no churn analysis for the bank customer and credit card holder and representing it to the client through the user interface with interactive visualizations and predictions.

V. APPLIED ALGORITHMS

1. Logistic Regression:

Given the values of one or more predictor variables, the logistic regression model forecasts the likelihood of the binary outcome variable (often labelled as 0 or 1). A logistic function, which converts any real-valued input to a number between 0 and 1, is used to represent the relationship between the predictor variables and the likelihood of the outcome variable.

The formula for logistic regression can be expressed as:

$p(y=1|x) = 1 / (1 + e^{-(\beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta kxk)})$

2. Decision Tree Classifier:

A supervised machine learning approach called a decision tree classifier creates a model with a tree-like structure to generate predictions based on a set of rules discovered from training data. Each node on the decision tree represents a choice or test of a feature, and each branch on the decision tree represents a potential result of that choice or test. The following is the decision tree classifier's mathematical formula, which is based on the ideas of information gain and entropy:

Let D be a collection of training data that includes samples (x1, y1), (x2, y2),..., (xm, ym), each of which is composed of a feature vector (xi) and a class label (yi) that corresponds to it. Let T be the decision tree classifier that will be built, and let A be a list of potential attributes that might be utilised to divide the data.

Calculate the entropy of the target variable, which is given by:

$H(Y) = - \Sigma p(y) \log 2 p(y)$

Where p(y) is the proportion of samples in D that belong to class y.

For each attribute $a \in A$, calculate the information gain (IG) that can be obtained by splitting the data based on the values of a. The information gain is defined as:

$IG(D, a) = H(Y) - \Sigma |Di|/|D| H(Y|Di)$

Where |Di| is the number of samples in subset Di of D, and H(Y|Di) is the entropy of the target variable Y given subset Di.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 5, pp 114-119, May-2023

https://doi.org/10.47001/IRJIET/2023.705013

As the decision tree's root node, pick the attribute a with the highest information gain.

Repeat steps 1-3 for each partition until all samples are put into a single class. Recursively partition the data based on the values of the chosen attribute.

3. Gradient Boosting:

The basic goal of gradient boosting is to iteratively expand the ensemble of decision trees while minimising the loss function. The objective is to reduce this disparity, which is measured by the loss function as the difference between the target variable's anticipated values and actual values.

Gradient boosting's mathematical formula is as follows:

Set the prediction value f0(x) to the target variable y's mean as its initial value.

$f0(x) = mean(y)$

For each iteration t = 1,2,...,T, do the following:

a. Calculate the negative gradient of the loss function L(y, f(x)) with respect to the current prediction value ft-1(x) for each training sample (xi, yi).

$ri = - \partial L(yi, f(x)) / \partial ft-1(xi)$

b. Create the new prediction function ht(x) by fitting a weak learner to the negative gradients ri (often a decision tree).

c. Determine the ideal step size t by reducing the loss function relative to the step size:

$\alpha t = \text{argmin } \alpha \, L(y, ft-1(x) + \alpha ht(x))$

d. Update the prediction function by adding the new tree multiplied by the step size:

$ft(x) = ft-1(x) + \alpha t \, ht(x)$

The final prediction value is the sum of all the trees in the ensemble:

$F(x) = \Sigma t=1..T \, ft(x)$

4. Random Forest Classifier:

The advantage of the Random Forest Classifier method is that it lessens the over fitting problem that is present in single decision trees. The approach can produce a diverse set of trees that can enhance the generalization performance of the model by using several trees with random feature and data selections.

The formula for the Random Forest Classifier can be written as follows:

For each tree t = 1,2,...,T, do the following:

a. Randomly select a subset of the training data with replacement.

b. Randomly select a subset of the features.

c. Build a decision tree using the selected data and features.

For a given input sample x, make a prediction by aggregating the predictions of all the decision trees in the forest:

$f(x) = \text{argmax}(\Sigma t=1..T \, I(yt = k) / T)$

VI. RESULTS

	CrossVal_Score_Means	CrossValerrors	Algo
0	0.963036	0.005664	RandomForestClassifier
1	0.960779	0.004611	Gradient Boosting
2	0.945965	0.005247	ExtraTreesClassifier
3	0.936089	0.007464	DecisionTreeClassifier
4	0.898843	0.006628	Logistic Regression

Figure 3

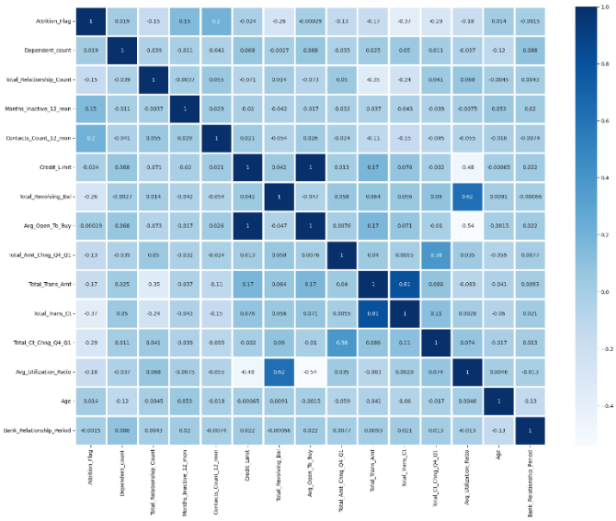

Figure 4

By observing at the specific values in the heatmap in Figure 4, some key observations are:

The target variable "Attrition Flag" is negatively correlated with "Total_Trans_Amt" and "Total_Trans_Ct", which suggests that customers who conduct more transactions are less likely to churn.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 5, pp 114-119, May-2023

https://doi.org/10.47001/IRJIET/2023.705013

"Total_Revolving_Bal" and "Avg_Utilization_Ratio" are negatively correlated with several variables, indicating that customers who carry a high balance on their credit cards and have a high utilization ratio are more likely to churn.

"Total_Ct_Chng_Q4_Q1" and "Total_Amt_Chng_Q4_Q1" are positively correlated with "Total_Trans_Ct" and "Total_Trans_Amt", which suggests that customers who increase their transaction activity are more likely to continue their relationship with the company.

"Dependent_count" and "Months_Inactive_12_mon" have weak correlations with most other variables, suggesting they may not be strongly related to customer churn.

"Credit_Limit" and "Avg_Open_To_Buy" are highly correlated with each other, which is expected since they are both measures of a customer's available credit.

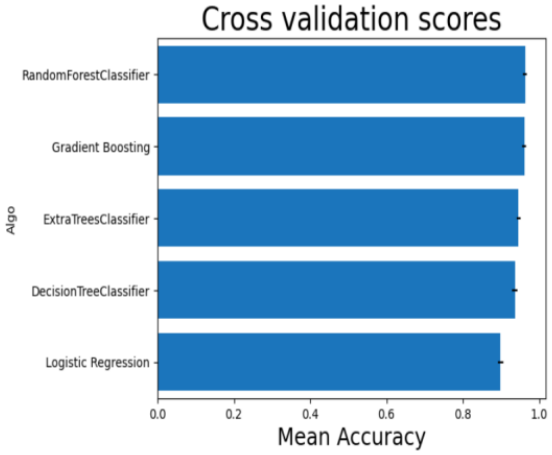

Figure 5

Based on the accuracies obtained from cross-validation on a customer churn dataset (7), the table shows the performance of five different machine learning algorithms.

The first column "CrossVal_Score_Means" shows the mean accuracy scores obtained from cross-validation. The second column "CrossValerrors" shows the standard errors of the mean accuracy scores.

According to the table in Figure 3, the Random Forest Classifier algorithm achieved the highest accuracy score of 0.963036, with a standard error of 0.005664. The Gradient Boosting algorithm achieved the second-highest accuracy score of 0.960779, with a standard error of 0.004611.

The Extra Trees Classifier algorithm achieved an accuracy score of 0.945965, followed by the Decision Tree Classifier algorithm with an accuracy score of 0.936089. The Logistic Regression algorithm achieved the lowest accuracy score of 0.898843.

In summary, the table in figure 3 and the graph in figure 5 provides a comparison of the accuracies of five different machine learning algorithms for predicting customer churn, with the Random Forest Classifier algorithm performing the best and the Logistic Regression algorithm performing the worst.

VII. CONCLUSIONS

In case of Random Forest classifier, we observed some significant differences when compared with other algorithms. As per the research it gave an accuracy of nearly around 96% which is higher than previous algorithms discussed and it can handle large datasets efficiently.

We trained a machine learning model to predict customer churn using several machine learning algorithms, including Random Forest Classifier (RFC), Decision Tree Classifier (DTC), and Logistic Regression (LR).

After experimenting with these models, we observed that the accuracy was not as high as we expected. We suspected that the problem could be with the dataset. On further analysis, it was found that the dataset was heavily biased towards non-attrited customers, which made it difficult for the model to identify patterns that could help predict churn more accurately. To improve the accuracy of the model, we tried to perform hyper parameter tuning by optimizing the model parameters, but it did not yield significant improvements. Ultimately, we realized that the quality of the dataset (7) (8) was limiting the accuracy of the model, and we needed to address the bias in the data to improve the performance of the machine learning model.

ACKNOWLEDGEMENT

We would like to express our profound gratitude to Dr. Vinod V. Kimbahune, Head of Department Computer Engineering, and Dr. Lalit Kumar Wadhwa, Principal Dr. D. Y. Patil Institute of Technology Pimpri Pune for their contributions to the completion of our project We would like to express my special thanks to our mentor Mrs. Rucha Madali for her time and efforts she provided throughout the year. Your useful advice and suggestions were really helpful to us during the project's completion. In this aspect, we are eternally grateful to you. We are deeply grateful to all the staff members of Computer Department, Dr. D. Y. Patil Institute of Technology Pimpri Pune for supporting us in all aspects.

REFERENCES

[1] Ahmet Tuğrul Bayrak; Asmin Alev Aktaş; OkanTunalı; OrkunSusuz; Neşe Abbak, International Conference on

ISSN (online): 2581-3048
Volume 7, Issue 5, pp 114-119, May-2023
https://doi.org/10.47001/IRJIET/2023.705013

Big Data and Smart Computing (BigComp) 2021 Publisher: IEEE.

[2] Manas Rahman; V Kumar, 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA) 2020 Publisher: IEEE.

[3] Anurag Bhatnagar; Sumit Srivastava, 9th International Conference on Advanced Computing (IACC) Publisher: IEEE.

[4] Eunjo Lee; Boram Kim; Sungwook Kang; Byungsoo Kang; Yoonjae Jang; Huy Kang Kim, Transactions on Games (Volume: 12, Issue: 1, March 2020) Publisher: IEEE.

[5] Benlan He Yong, Shi Qian, WanXi Zhao, Procedia Computer Science Volume 31 International Conference on Information Technology and Quantitative Management, ITQM.

[6] Irfan Ullah, Basit Raza, Ahmad Kamran Malik, National Research Foundation of Korea (NRF) Publisher: IEEE.

[7] https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers

[8] https://www.kaggle.com/code/sudhanshu2198/bank-customer-churn-prediction/input

AUTHORS BIOGRAPHY

Vandit Talwadia,
Student, Computer Engineering,
Dr. D. Y. Patil Institute of Technology Pune, Savitribai Phule Pune University, India.

Lavesh Chanchawat.
Student, Computer Engineering,
Dr. D. Y. Patil Institute of Technology Pune, Savitribai Phule Pune University, India.

Shubh Kumar Jain,
Student, Computer Engineering,
Dr. D. Y. Patil Institute of Technology Pune, Savitribai Phule Pune University, India.

Suma Keerthana Pepeti,
Student, Computer Engineering,
Dr. D. Y. Patil Institute of Technology Pune, Savitribai Phule Pune University, India.

Citation of this Article:

Vandit Talwadia, Shubh Kumar Jain, Lavesh Chanchawat, Suma Keerthana Pepeti, "An Integrated Bank Customer and Credit Card Holder Churn / No Churn Analysis System Using Machine Learning" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 7, Issue 5, pp 114-119, May 2023. https://doi.org/10.47001/IRJIET/2023.705013

****

