IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 25, Issue 4, Ser. I (July. – August. 2023), PP 27-33
www.iosrjournals.org

Analysis And Prediction Of Customer Churn Using Machine Learning - A Case Study In The Banking Sector.

Mr.Sindikubwabo Emmanuel[1], Dr.Ndengo Marcel[2]
[1](*Data science, African centre of excellence in Data science, University of Rwanda, College of Business and Ecomomics, Rwanda)*

Abstract:
Customer turnover is a global issue that has an impact on the banking business. This study aims to raise knowledge of whether a customer is likely to switch banks depending on requested services. The several ensemble machine learning algorithms namely, Ada-Boost (AD), Random Forest (RF), Light Gradient Boosted Machine (LGB), and CatBoost (CT) combined to make the proposed technique for our research called Super learner. The study compared the super learner that comes from ensembles to the one produced under the combination and configuration of weaker learners machine learning models precisely, Decision Tree (DT), K Nearest Neighbours (KNN), Support Vector Machine (SV M), and Logistic Regression(LR). We used a churn for bank customer data-set from Kaggle. Super Learning algorithms helped us to categorize customers who are likely to change from one bank to another bank and those who are not. Both of the super learners were able to outperform all of the employed machine learning models. The machine learning evaluation measures assisted us in deciding that the super learner produced by ensemble machine learning models, with an accuracy of 87.7%, was the optimal model to utilize in our research using the data set from kaggle, which recorded past customer's bank information.

Background: *Customer turnover in traditional banks is a global issue as customers shift to high-tech banks for better services [1]. The World Retail Banking Report 2019 shows that 66.8% of existing clients plan to or have already used non-traditional banks. Traditional banks are threatened by non-traditional financial sector competitors, making it harder for them to retain their customer base. Minimizing customer churn is essential as acquiring new customers is costly [2]. To prevent turnover, businesses now study and monitor client behavior to identify potential churners early and take proactive action to retain clients and boost profitability. However, identifying churn is challenging due to the vast customer base of large banks, leading to a demand for automated solutions using machine learning techniques to understand complex patterns in data. The study proposes a "super learner" model that combines various machine learning approaches for improved churn prediction accuracy.*

Materials and Methods: *This study uses an effective advanced machine learning approach to predict client attrition in a bank as early as possible . In the banking industry, many machine learning algorithms have been used to predict the possibility of client turnover. The majority of these techniques (basic models and a few ensemble learners like random forest and extra gradient boost) were trained on both actual data from local banks and publicly available data from machine learning repository.*

Kaggle. According to the findings of several research investigations, the choice of characteristics and proper approach might impact the prediction outcomes of the chosen model. This work intends to use feature selection (SelectfromModel) along with the combined machine learning-based model called super learner, which have received less attention in previous studies. The sections that follow offer an overview of each of the methodologies listed above.

Results: *The study utilized four weak learners: Decision Tree, Logistic Regression, K-Nearest Neighbors, and Support Vector Machine, to predict outcomes from a Kaggle dataset. KNN, SVM, and DT outperformed LR, with accuracy rates of 86.52%, 86.38%, and 83.99% respectively. The research also employed four ensemble machine learning models, including Ada-Boost, Cat-Boost, LightGradientBoost, and Random Forest. RF, AD, and CB outperformed LGB, with accuracy rates of 87.35%, 87.10%, and 87.10%. The study then constructed two super learner models using logistic regression as a meta model. Both super learners outperformed individual models, with the ensemble-based super learner achieving an accuracy of 86.7%, making it the optimal model.*

Conclusion: *The study aimed to detect customer turnover in banking by using various prediction methods. K-Nearest Neighbor, Decision Tree, Support Vector Machine, and Logistic Regression were employed to build the first super learners. Built-in ensemble models like Random Forest, Light Gradient Boosting, Cat-Boost, and Adaboost were also used. The proposed method, a super learner ensemble model, outperformed other machine learning models with an accuracy of 87.7%.*

Key Word: *Customer churn, Machine Learning*

Date of Submission: 20-07-2023 Date of Acceptance: 30-07-2023

I. Introduction

Customer turnover has increased in traditional banks, as customers migrate to new high-tech banks to better access their desired services, making it a global issue [1]. Worldwide bank executives are well aware of how important it is to boost client happiness. Customers' demands for financial services are increasing as they make a greater utilization of new technologies in other areas of social life. In the next couple of years, 66.8% of existing banking clients plan to or have already utilized a bank account from a non-traditional business, according to the World Retail Banking Report, 2019 [2]. As stated in [3], traditional banks are threatened by these non-traditional financial sector competitors, according to 55% of bank executives. This scenario of distinct competition makes it harder for traditional banks to hold onto their current customer base. Customer churn, the act of specific customers leaving their current firm to use the services of a rival organization, is becoming more prevalent [4].Since acquiring new consumers can be up to five times more expensive than retaining and satisfying current ones, numerous studies have shown that minimizing customer churn could result in financial savings [5, 6]. As a result, managing client connections is becoming an increasingly important investment for organizations in order to prevent turnover. Companies now study and monitor their client behavior to spot clients who are more likely to churn early on as a result of the need to protect their revenues. Businesses can take proactive action in this way to keep clients and boost profitability. However, some businesses fail to keep their consumers. A study looked into why clients in South Africa left or switched banks [7]. This research looked at 1.7 million social media posts about customer churn from five major South African banks. People expressed their feelings on various services via various web platforms. This enabled financial organizations to identify areas where they may improve their services. Most banks across the world take client satisfaction into account and make every effort to im- prove it. It is due to the manner that technology is evolving that customer perceptions and needs are changing. But, there are still difficulties in identifying churn specifically in the banking in- dustry. First, the customer base of large banks often numbers in the tens of millions. Human intervention-based approaches to churn reduction do not scale up well. Second, they can't evolve quickly enough to meet shifting customer demands. Third, despite the fact that banks divide their clientele among local managers, it is still challenging to identify consumer trends by hand, espe- cially if the bank manages a sizable clientele. These characteristics drive the demand for automated solutions that can identify the non-trivial patterns of consumer behavior that may indicate future churn in advance in these enormous data sets. These qualities drive the employment of machine learning techniques, which offer supervised learning approaches that have demonstrated the ability to understand complex patterns in the data (without human intervention) and generalize well to previously unexplored data. In the academic literature, several strategies have been used to predict customer churn, with em- phasis on decision trees [8, 9], k-nearest neighbors [10, 11], elastic net [12], logistic regression[13], SVMs [14], and random forests [15]. Despite offering accurate predictions, the majority of these studies concentrated on using a single statistical models and, occasionally, used significantly di- minished bases in their experiments. Our study will merge all of the previously described machine learning models into a single model called super learner in order to improve model prediction ac- curacy. Our study will merge all of the previously described machine learning models into a single model called super learner in order to improve model prediction accuracy. The study was carried out by providing the reader with an overview and allowing us to have the first chapter, which is called the introduction. The study was based on the limitations of earlier research studies that employed machine learning models to forecast client attrition.

II. Material And Methods

Statistical methods*:* This study uses an effective advanced machine learning approach to predict client attrition in a bank as early as possible.In the banking industry, many machine learning algorithms have been used to predict the possibility of client turnover. The majority of these techniques (basic models and a few ensemble learners like random forest and extra gradient boost) were trained on both actual data from local banks and publicly available data from several machine learning repositories such as Kaggle and our world in data, among others. According to the findings of several research investigations, the choice of characteristics and proper approach might impact the prediction outcomes of the chosen model. This work intends to use feature selection (SelectfromModel) along with the combined machine learning-based model called super learner, which have received less attention in previous studies. The sections that follow offer an overview of each of the methodologies listed above.

Super Learner

Based on the weak performance of some of the machine learning models, weak learners, that have been used, we come up with the idea of using a super learner. The super learner method has been one of the best combinations of weak learners. It combines weak learners and uses cross-validation to estimate the risk of the data which greatly improved their collective performance as shown from the figure below.

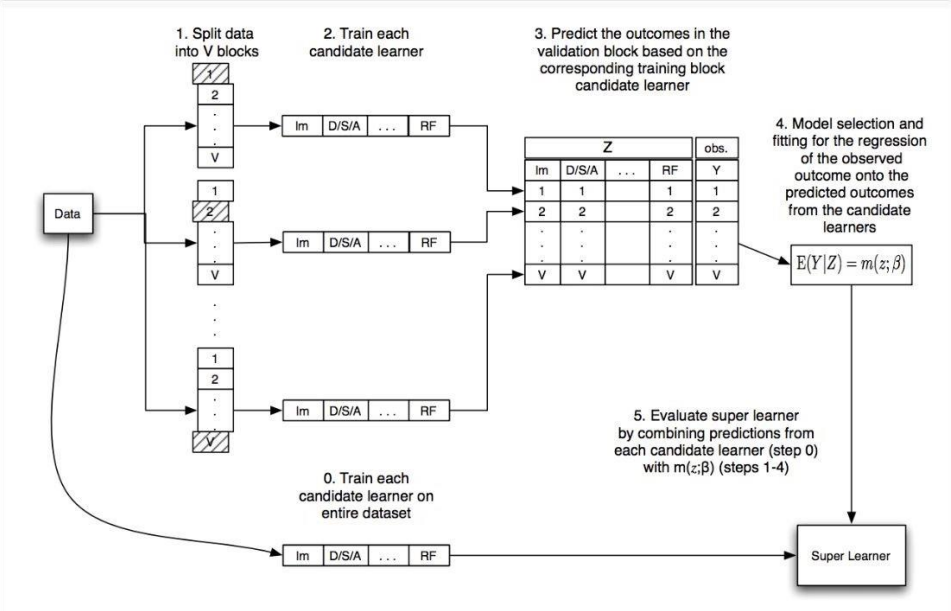

Figure 1: Super Learner

Figure 1 is the architecture that shows in more detail how the super learner is built from the weak learners. Data here refer implicitly to the training data set and when step 0 is mentioned on the entire data set, it implicitly refers to the training set as well. After splitting the training set into K data sub-sets. For the K^{th} data sub-set ($1 \leq k \leq K$) , the algorithm fits each one of the weak learners to the data and outputs corresponding pre-results. All the pre-results are gathered into a new data frame whose columns can be named according to the weak learners. In step 4.of Fig. 1, the regression algorithm is used to learn the best possible linear combination of the features of the new data set. In this work, not only we used the regression technique for the final prediction but we also interchanged their positions.

Study Location: The study was supposed to be done in the bank of Rwanda, bank of Kigali but due to lack of data, Kaggle data were used.

Sample size and features: The churn for bank customer data set from Kaggle [15] has been investigated in this study. The dataset was used to train our ensemble models. Following data wrangling of the data set consisting of 10, 000 samples and 14 features, we implemented the aforementioned ensemble models to generate predictions toward our research aims. Because the target variable column in the dataset is termed exited, the task is a binary classification problem. The target variable column (exited) shows whether the consumer will continue to use the bank. Looking further into the class (exited) column, it should be noticed that there were nearly 4 times more clients who chose to stay in the bank than left, as seen in Figure 2. The target label, 1 indicates that the client's bank account has been closed (customer exited), whereas 0 indicates that the client's bank account is still retained, as shown in Figure 2. Overall, the data set includes the 13 predictor variables that will be used to train our models and make predictions. The figure 2 .below provides more information about the dataset and describes each feature in detail.

No	Feature Name	Feature Description
1	Row number	Instances from 1 to 10000.
2	Customer Id	Individual IDs to identify bank customers.
3	Surname	Last Name of customer.
4	Credit Score	Customer's credit score.
5	Geography	the nation that the consumer is from.
6	Gender	Whether Male or Female.
7	Age	Age of the Client.
8	Tenure	the length of time the consumer has been a bank customer.
9	Balance	Customer's balance in the bank
10	Num of Products	number of bank products being used by the customer (savings account, mobile banking, internet banking etc.)
11	Has Cr Card	Whether the customer has a credit card with the bank is indicated by a binary flag.
12	Is Active Member	Whether the consumer is an active member of the bank is indicated by a binary flag.
13	Estimated Salary	estimated compensation in US dollars for the client.
14	Exited	Binary flag 1 if the client's bank account was closed and 0 if the client was retained.

Figure 2: Explanation on Data set's features

Data Analysis and Discussion: we included the summary statistics of the dataset (Figure 3). This table demonstrates that there were no missing values (total count for each attribute is the same as the length of the dataset). It is also worth noting that numerical features have a broad range of values.This implies that we should standardize or normalize data before fitting our models.

	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000
mean	650.528800	38.921800	5.012800	76485.889288	1.530200	0.70550	0.515100	100090.239881
std	96.653299	10.487806	2.892174	62397.405202	0.581654	0.45584	0.499797	57510.492818
min	350.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.000000	11.580000
25%	584.000000	32.000000	3.000000	0.000000	1.000000	0.00000	0.000000	51002.110000
50%	652.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000
75%	718.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500
max	850.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000

Figure 3: Summary statistics

It is as vital to understanding how the dataset attributes correlate with one another as it is to have an overview of what the data looks like. Depending on the degree of the linear correlations that exist between characteristics, certain features may need to be excluded from the analysis when they contain comparable information. Similarly, it is necessary to assess the extent to which each predictor variable influences the class attribute, in which case a correlation matrix would be required. In this regard, we generate a heatmap using the correlation matrix to better understand the interactions between each feature, allowing us to find the feature with the largest effect on our target variable (Figure 4).

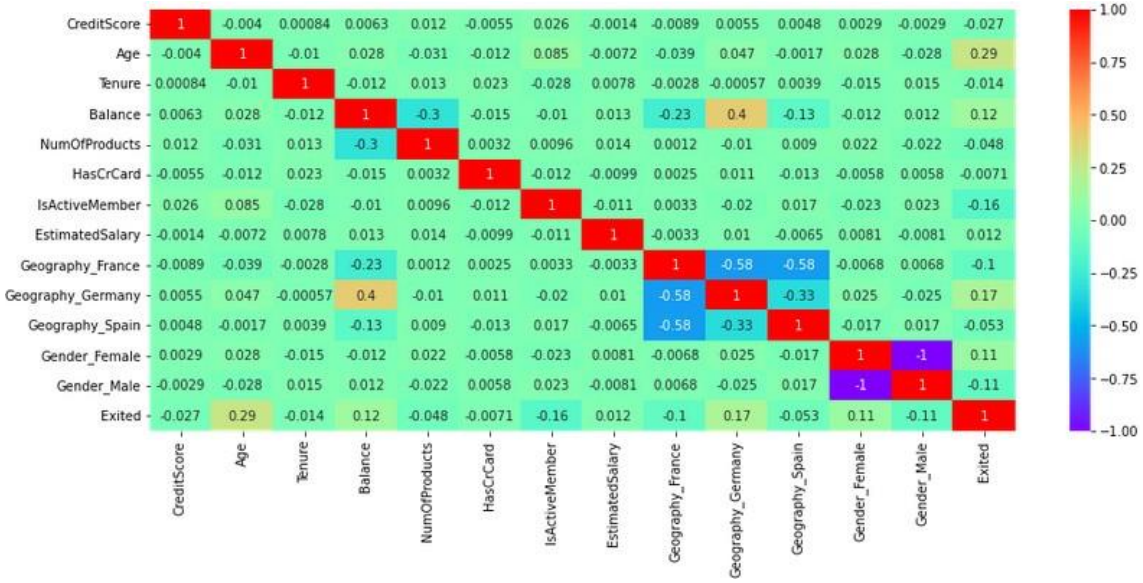

Figure 4: Linear relationship of each attribute with the target variable

In addition to the correlations between each attribute and the target variable, the class distributions are substantially skewed, a machine learning problem that is known as class imbalance. As a resultof this issue, many categorization learning algorithms may have not provided accurate predictions for the minority class. According to the graphical depiction (Figure 5), the dataset contains roughly four times more samples from the positive class (customers who remained) than from thenegative class (exited customers).

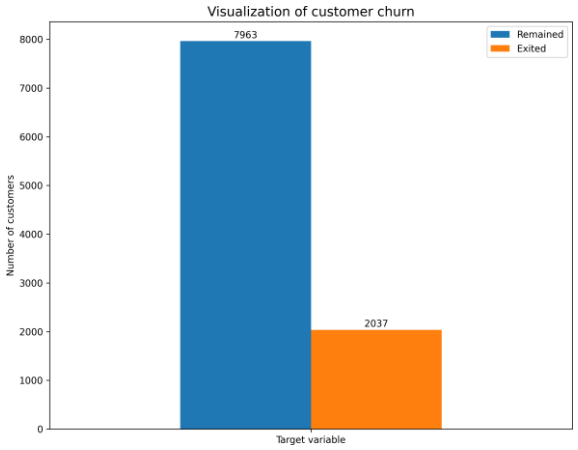

Figure 5: Class distribution

III. Result

The study employed four weak learners: Decision Tree (DT), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). The prediction was made using the dataset from Kaggle [15] using the aforementioned machine learning algorithms. The data set wasdivided into two data sets, creating a training set with 800 in instances, and a testing set with 200 instances. The training set was also put through cross-validation using 10 folds. The results fromthe prediction after we trained our models are shown in the

Table 1. below.

	Classifier	Accuracy score
0	LogisticRegression	81.02
1	DecisionTreeClassifier	83.99
2	SVC	86.38
3	KNeighborsClassifier	86.52

Table 1: Base Learners Prediction Results

K Nearest Neighbors (KNN), Support Vector Machine (SV M), and Decision Tree Classifier (DT) all outperformed Logistic Regression (LR), which was the last model among the others with an accuracy of 81.02%, among the 4 base learners used in this study. KNN had an accuracy of 86.52%, 86.38%, and 83.99%, respectively. Additionally, for our research, we employed 4 ensemble machine learning models including Ada-Boost (AD), Cat-Boost (CB), LightGradienBoost (LGB), and Random Forest (RF). After training each of these ensemble machine learning models, the prediction was completed, and evaluation metrics revealed that Random Forest (RF), Ada-Boost (AD), and Cat-boost (CB) outperformed LightGradientBoost (LGB), which was the last model among those aforementioned models, with an accuracy of 87.35%, 87.10%,87.10%, 86.95%, re- spectively. Moreover, as shown in Table 2. below, all prediction results from base learners and ensemble models were recorded.

	Classifier	Accuracy score
0	LogisticRegression	81.02
1	DecisionTreeClassifier	83.99
2	SVC	86.38
3	KNeighborsClassifier	86.52
4	CatBoostClassifier	87.10
5	LGBMClassifier	86.95
6	AdaBoostClassifier	87.10
7	RandomForestClassifier	87.35

Table 2: Comparison between Ensembles and Base Learners Prediction Results

Since building the super learner machine learning models from scratch has been the main focus of our study. One super learner was formed by base learners, while the other was made by the ensembles that we recently used above. These two different super learner models were developed. The two super learner models stated above were constructed using the logistic regression model as a meta model. Both of the super learners were able to outperform all of the employed machine learning models. The machine learning evaluation measures assisted us in deciding that the super learner produced by ensemble machine learning models, with an accuracy of 86.7%, was the optimal model to utilize in our research. All model prediction results are recorded and shown in the table 3. Below.

	Classifier	Accuracy score
0	LogisticRegression	81.02
1	DecisionTreeClassifier	83.99
2	SVC	86.38
3	KNeighborsClassifier	86.52
5	LGBMClassifier	86.95
8	Super Learner (base learners)	87.05
4	CatBoostClassifier	87.10
6	AdaBoostClassifier	87.10
7	RandomForestClassifier	87.35
9	Super Learner (Ensembles)	87.70

Table 3: All Model Prediction Results

IV. Discussion

First and foremost, in the business world, banking institutions must do their best to provide services to consumers without making any mistakes, unless they are unavoidable, and not as a result of their willingness or engagement. In this regard, early detection of customer turnover is the key to a sustainable solution. In this work, various prediction methods have been used in predicting customer churn on data set composed of one hundred instances with thirteen columns excluding the target variable. We used K-Nearest Neighbor (KNN), Decision Tree (DT), Support vector machine (SV M), and Logistic regression (LR) to build the first super learners and built- in ensemble models like random forest (RF), light gradient boosting (LGB), cat-boost (CB) and Adaboost (AD). Each model was evaluated based on accuracy. In a nutshell, the proposed method–super learner, which was an ensemble based model– outperformed the other ML models with the accuracy of 87.7%.

V. Conclusion

Based on our findings, we believe the accuracy was lower than expected. This might be due to the fact that the weak learnt used to generate the super learner were sklearn built-in models withdefault configurations. To optimize the performance of our super learner, we would recommend developing super learners utilizing weak learners created from scratch in future works.Finally, this study used a publicly accessible dataset. As a result, it is impossible to state the consequences of underlying unproven assumptions on the effectiveness of the employed models. I recommend that future researchers will use the primary dataset.

Acknowledgement

I would like to take this opportunity to expressmy deepest gratitude to thank My supervisor Dr.Marcel Ndengo who contribute a lot with his tenacious professional and technical guidance for this dissertation to be done.My greatest thanks also go to my family, especially my mum to whom I owe my love and gratitude. I would also like to express my gratitude to the efforts of all the academic administrators of the College of Business and Economics at the University of Rwanda for providing me with all I needed in my project implementation. To all my friends whose names are not mentioned here that have helped and supported me along the way, I thank you from the bottom of my heart. I wish you all the best in life and hope our friendship will last forever.

References

[1]. Z. Wen, J. Yan, L. Zhou, Y. Liu, K. Zhu, Z. Guo, Y. Li, And F. Zhang, "Customer Churn Warning With Machine Learning," In The Euro-China Conference On Intelligent Data AnalysisAnd Applications. Springer, 2018, Pp. 343–350.

[2]. R. A. De Lima Lemos, T. C. Silva, And B. M. Tabak, "Propension To Customer Churn In A Financial Institution: A Machine Learning Approach," Neural Computing And Applications, Pp. 1–18, 2022.

[3]. Pwc, "Retail Banking 2020 Evolution Or Revolution?" 2020. [Online].Available: Https://Www.Pwc.Com/Gx/En/Banking-Capital-Markets/Banking-2020/Assets/Pwc-Retail-Banking-2020-Evolution-Or-Revolution.Pdf

[4]. Z. Wen, J. Yan, L. Zhou, Y. Liu, K. Zhu, Z. Guo, Y. Li, And F. Zhang, "Customer Churn Warning With Machine Learning," In The Euro-China Conference On Intelligent Data AnalysisAnd Applications. Springer, 2019, Pp. 343–350.

[5]. A. Sharma And P. Kumar Panigrahi, "A Neural Network Based Approach For Predicting Customer Churn In Cellular Network Services," International Journal Of Computer Applications, Vol. 27, Pp. 26–31, 08 2011.

[6]. J. Xiao, Y. Xiao, A. Huang, D. Liu, And S. Wang, "Feature-Selection-Based Dynamic Transfer Ensemble Model For Customer Churn Prediction," Knowledge And Information Systems, Vol. 43, No. 1, Pp. 29–51, 2015.

[7]. J. Lappeman, M. Franco, V. Warner, And L. Sierra-Rubia, "What Social Media Sentiment TellsUs About Why Customers Churn," Journal Of Consumer Marketing, No. Ahead-Of-Print, 2022.

[8]. G. Nie, W. Rowe, L. Zhang, Y. Tian, And Y. Shi, "Credit Card Churn Forecasting By Logistic Regression And Decision Tree," Expert Systems With Applications, Vol. 38, No. 12, Pp. 15 273– 15 285, 2011.

[9]. L. Bin, S. Peiji, And L. Juan, "Customer Churn Prediction Based On The Decision Tree In Personal Handyphone System Service," In 2007 International Conference On Service Systems And Service Management. Ieee, 2007, Pp. 1–5.

[10]. M. Eastwood And B. Gabrys, "A Non-Sequential Representation Of Sequential Data For Churn Prediction," In International Conference On Knowledge-Based And Intelligent Information And Engineering Systems. Springer, 2009, Pp. 209–218.

[11]. Y. Zhang, J. Qi, H. Shu, And J. Cao, "A Hybrid Knn-Lr Classifier And Its Application In CustomerChurn Prediction," In 2007 Ieee International Conference On Systems, Man And Cybernetics. Ieee, 2007, Pp. 3265–3269.

[12]. R. Prashanth, K. Deepak, And A. K. Meher, "High Accuracy Predictive Modelling For Customer Churn Prediction In Telecom Industry," In International Conference On Machine Learning And Data Mining In Pattern Recognition. Springer, 2017, Pp. 391–402.

[13]. T. Mutanen, S. Nousiainen, And J. Ahola, "Customer Churn Prediction–A Case Study In RetailBanking," In Data Mining For Business Applications. Ios Press, 2010, Pp. 77–83.

[14]. M. A. H. Farquad, V. Ravi, And S. B. Raju, "Churn Prediction Using Comprehensible Support Vector Machine: An Analytical Crm Application," Applied Soft Computing, Vol. 19, Pp. 31–40, 2014.

[15]. M. Akturk. Churn For Bank Customers. [Online]. Available: Https://Www.Kaggle.Com/ Datasets/Mathchi/Churn-For-Bank-Customers