iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VIII **Month of publication:** August 2021

DOI: https://doi.org/10.22214/ijraset.2021.37467

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

Bank Customer Churn Prediction Using Machine Learning

S. Likhit Kumar[1], G. Sandilya Srinivas[2], J. Pavan Ganesh[3], M. Shailaja[4]
[1, 2, 3]*Department of Electronics and Computer Engineering, Sreenidhi Institute of Science and Technology, Hyderabad – 501301, Telangana, India*
[4]*Associate Professor, Department of Electronics and Computer Engineering, Sreenidhi Institute of Science and Technology, Hyderabad – 501301, Telangana, India*

Abstract: ***Banking is one of the highly competitive sectors where customer relations is of utmost importance for any bank. Each customer is considered as a customer for life by the banks. The term "Customer Churn" refers to the state in which the customer or the subscriber stops involving in business transactions with a company or a service provider. To deal with this, the paper presents the work done towards predicting the customer churn rate, using machine learning models which will indicate whether a customer will leave the bank or not based on many factors, this in turn will help the bank in knowing which category of customers generally tend to leave the bank. Further the banks can bring in exciting offers so that it can retain its customers. In this predictive process popular models such as logistic regression, decision trees, random forest and other boosting techniques have been used to achieve a decent level of accuracy, for the banks to rely upon.***
Keywords: Banking, Gradient Boosting, Customer Churn, Machine Learning, Random Forest Classifier.

I. INTRODUCTION

Customer churn, also known as customer attrition is the term coined for the probability of whether an existing customer continues his/her transactions with the organization or not. The probability factor of this parameter depends on numerous other factors in various industries like banking, telecom and few other sectors. In intense market scenarios that are prevailing today in every sector, it becomes important for the organizations to keep a track of customer churn and the various reasons causing the customer to stop their transactions with the company. Almost every organization is well-versed with the concept that retention of the existing customers will save a lot of money as trying to acquire new customers will cost five to six times the cost to retain an existing customer. Therefore, each organization out in the market started to understand and analyse the various factors that might be the cause for a customer or client to leave the organization's business.

Once importance of customers and customer churn was identified by various leading companies, they started to gather data on customer behaviour like how often does a customer purchase a product, etc, with this, the collection, storage and managing of the customer data on a time basis became really crucial for the companies. Thus, the decision-making process shifted from being an event-driven approach to a data-driven approach. New technologies like Big Data, Cloud storage and machine learning supported the companies' data gathering, processing and analysis phase. Therefore, the whole process shifted to statistical analysis as opposed to predictive analysis.

This paper deals with the customer churn in a banking sector and highlights the various factors that affect the attrition of customers from the bank. It also sheds some light on how a bank can predict the rate of customer churn by making use of well-known and popularly used machine learning models.

II. EXISTING SYSTEM

Many approaches were explored and used in predicting the churn in various industries like telecom, banking, etc, most of the approaches have utilized the power of data mining techniques and machine learning algorithms. The significant amount of previous work focused on utilizing only a single approach of data mining to pull data from appropriate sources, and the others concentrated on numerous machine learning algorithms to interpret churn rate. Back in the day with SVM-POLY using AdaBoost as the best possible model, supervised machine learning approaches have been employed in customer churn prediction problems (Vafeiadis, Diamantaras, Chatzisavvas & Sarigiannidis, 2015). Decision tree, Random-forest, Multilayer perceptron, and SVM are the most usual approaches applied for forecasting customer churn rate. Neural networks, support vector machines and logistic regression models are the approaches that were widely used to forecast customer churn.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

III. PROPOSED SYSTEM

This paper proposes the use of machine learning and data mining techniques to differentiate the customer who are in the risk of being churned and customers who are happy and satisfied with the products and services of the bank. The algorithms and various technologies employed to predict the customer churn for a bank are discussed in the following sections in- detail. This paper discusses about the use of several data pre-processing steps and machine learning algorithms and techniques like logistic regression, AdaBoost, etc, and compares each of their prediction power by taking into consideration the accuracy of each machine learning models and selecting the best model with highest accuracy, to forecast the customer churn. A front-end application which is a simple webpage is provided to the end-user to enter the details of a customer to check whether he/she is at a risk of being churned or is happy and satisfied with the products and services of the bank.

IV. METHODOLOGY

The following are the software requirements for building the model and web application, to get the desired level of results:

1) Python Programming
2) Jupyter Notebook
3) Visual Studio Code
4) Streamlit

The following are Python libraries utilized to obtain the results:

a) Pandas
b) NumPy
c) Matplotlib
d) Seaborn

The following are the machine learning models and modules obtained from the sklearn library in Python. The accuracy of these model's prediction is compared and the model with the best accuracy is chosen for future predictions.

A. Logistic Regression

Logistic regression is a basic classification technique. It belongs to the linear classifiers group and is similar to polynomial and linear regression. Logistic regression is a simple and rapid method for predicting results, and it is ideal for you to use. Although it is primarily a binary classification approach, it may also be used to address multiclass issues such as this concept.

B. Decision Tree Classifier

A decision tree is a popular machine learning method. Internal decision- making logic is shared, which is not available in black box algorithms like Neural Network. When compared to the neural network approach, it takes less time to train. The number of records and attributes in the data supplied can impact the temporal complexity of decision trees. The decision tree is a non-parametric or distribution-free classifier that does not rely on probability distribution assumptions. Decision trees can handle large amounts of data and yet make accurate predictions.

C. Random Forest Classifier

Random-forest is a learning method that is supervised. This classifier may be used for both classification and regression. Random forest is also very adaptable and simple to utilise. Trees make up a forest. It is considered that the greater the number of trees available, stronger a forest becomes. Random forests create decision trees based on randomly selected data samples, obtain forecasts from each tree, and use the voting concept to select the best potential answer. Random forest also provides a fairly decent measure of feature value.

D. K-Neighbors Classifier

The supervised machine learning method K-nearest neighbours (KNN) is a kind of KNN algorithm. KNN is relatively simple to implement in its most basic form, yet it performs a wide range of classification tasks. It is a slow learning algorithm since it does not have an exclusive training phase. Rather, it trains on all of the data while categorising a new data point or instance. KNN is a non-parametric learning method since it makes no assumptions about the raw data.

©IJRASET: All Rights are Reserved

E. AdaBoost Classifier

Boosting algorithms have been more popular and well-known among data science and machine learning enthusiasts in recent years. These enthusiasts want to use boosting algorithms to win tournaments since they provide great accuracy. The data science projects provide as a platform for learning, researching, and developing practical solutions to a variety of corporate and government problems. Boosting algorithms combine many low- accuracy (or weak) models to produce a high-accuracy (or strong) model.

F. Gradient Boosting Classifier

Gradient boosting classifiers are a collection of machine learning algorithms that combine a number of weak learning models into a powerful prediction model. When applying gradient boosting, decision trees are frequently employed. Gradient boosting models are gaining popularity as a result of their efficiency in categorising compound datasets, and have recently been utilised by enthusiasts to win a number of Kaggle data science challenges. The idea behind "gradient boosting" is to take a bad hypothesis or a bad learning algorithm and make a series of changes to it that will improve the hypothesis's strength.

G. XGBoost Classifier

XGBoost is an open-source software package that allows you to use C++, Java, Python, R, Julia, Perl, and Scala to create a regularising gradient boosting framework. It's compatible with Linux, Windows, and Mac OS X. Its purpose, according to the project, is to produce a "Scalable, Portable, and Distributed Gradient Boosting (GBM, GBRT, GBDT) Component." It may run on a single system as well as distributed processing systems like as Apache Hadoop, Apache Spark, and Apache flink.

V. STEPS FOR CUSTOMER CHURN PREDICTION

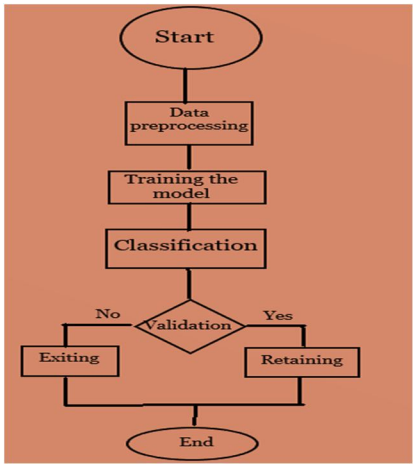


Fig. 1 Steps For Customer Churn Prediction

1) The dataset considered here is obtained from Kaggle and name of the dataset is “Bank customer churn data”. The dataset consists of 1000 records and 14 features.
2) Data pre-processing is the most basic and important step in any machine learning project. Data processing involve certain steps like Data cleaning, Data imputation, Dealing with outliers, Data transformation and Data visualization.
3) The next step is training the model and classification is done using different algorithms
4) Validation provides with the information whether a customer is churned or not.
5) It involves two cases based on the input values entered, the prediction is made in such a way that customers are retained or exited
6) If its **Yes** the customer are retained and he/she is with the bank.
7) In case of **No** customer exit’s which is nothing but the customer is churned.
8) The work presented in this paper also intends to make a front end (website) where the values entered give the output which help the company know whether the customer/client is staying or leaving the organization/bank.
9) This paper intends to predict the best accuracy rate by comparing different algorithms and the output is displayed on the website.

©IJRASET: All Rights are Reserved

VI. PIMPLEMENTATION

Before the model is constructed there are numerous pre-processing steps that need to be done in order for the model to be properly constructed. The previously mentioned machine learning models are employed in the final model construction and utilization.

The model construction can be divided into three major parts:

1) Splitting the Predictors(X) and the Target variable(y).

While splitting the features into Predictors and Target variables, this paper uses the iloc method to split the features column wise. In this case, 10 features are available as X. Those features are:

a) Credit Score
b) Geography
c) Gender
d) Age
e) Tenure
f) Balance
g) Number of products
h) Whether the customer has credit card or not
i) Whether the customer is an active member or not
j) Estimated Salary of the Customer

2) Splitting the X and y further. That is the test-train split. Now, having 4 variables. X_train, X_test ; y_train, y_test. The train data is 80% of the complete data whereas the test data is 20% of the complete data sampled randomly. This paper uses the train_test_split python library to split the data and defined the random state as 42.

3) Fitting the Required machine learning model using the algorithms.

The algorithm used are imported from the Scikit Learn library which is an open- source library.

Models are trained using the train data that was divided earlier using the train test split.

To predict the values, the test data from the train test split can be used.

For the purpose of evaluation, this paper uses the Confusion matrix and accuracy parameter derived from the confusion matrix parameters.

$$\text{Accuracy Score} = \frac{\text{Number of Correct PreAdictions}}{\text{Total Number of Predictions Made}}$$

The evaluation Metrics used were Train/Test split validation as well as K fold Cross validation techniques.

Since these two are different validation techniques, there will be a slight difference in the accuracy scores of the same model.

With the models used, the accuracies obtained are:

Table I
Various Ml Models And Their Accuracy In Predicting Customer Churn Rate

Models Used	Accuracy
Logistic Regression	77.30%
Decision Tree Classifier	78.93%
Random Forest Classifier	87.22%
K-NN Classifier	83.52%
AdaBoost Classifier	83.55%
Gradient Boosting Classifier	84.99%
Xtreme Gradient Boosting Classifier	86.72%

©IJRASET: All Rights are Reserved

Out of these all the models used, Random Forest classifier gives the highest accuracy of 87.2%.

But the validation technique used here was Train/Test split validation. Another validation technique is checked.

LOOCV (Leave one out Cross Validation).

Using LOOCV, the following accuracy scores were obtained:

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
gbc	Gradient Boosting Classifier	0.8634	0.8593	0.4664	0.7823	0.5840	0.5086	0.5332	0.2260
rf	Random Forest Classifier	0.8587	0.8474	0.4477	0.7711	0.5654	0.4883	0.5146	0.1750
lightgbm	Light Gradient Boosting Machine	0.8578	0.8516	0.4907	0.7313	0.5868	0.5050	0.5199	0.1960
ada	Ada Boost Classifier	0.8573	0.8433	0.4823	0.7352	0.5821	0.5005	0.5169	0.0840
xgboost	Extreme Gradient Boosting	0.8560	0.8390	0.4976	0.7178	0.5874	0.5036	0.5163	0.5200
et	Extra Trees Classifier	0.8403	0.8281	0.4054	0.6929	0.5107	0.4232	0.4452	0.1770
lda	Linear Discriminant Analysis	0.8308	0.8229	0.3098	0.7031	0.4289	0.3472	0.3879	0.0610
ridge	Ridge Classifier	0.8253	0.0000	0.2232	0.7561	0.3438	0.2765	0.3458	0.0410
dt	Decision Tree Classifier	0.7910	0.6898	0.5177	0.4952	0.5053	0.3731	0.3738	0.0140
lr	Logistic Regression	0.7878	0.6667	0.0588	0.3699	0.1006	0.0520	0.0819	0.6880
nb	Naive Bayes	0.7840	0.7543	0.1123	0.4087	0.1757	0.0964	0.1230	0.0080
knn	K Neighbors Classifier	0.7573	0.5218	0.0790	0.2357	0.1177	0.0164	0.0198	0.0300
svm	SVM - Linear Kernel	0.6914	0.0000	0.2202	0.1173	0.1334	0.0235	0.0301	0.0300
qda	Quadratic Discriminant Analysis	0.2062	0.5000	1.0000	0.2062	0.3419	0.0000	0.0000	0.0790

Fig. 2 Accuracy Scores for Various Models

Also, as part of the proposed system, this paper put-forth a real-time webapplication, as shown below, in which the bank can feed in the customer details which in turn will help the bank to understand which customer is at a risk of being churned.

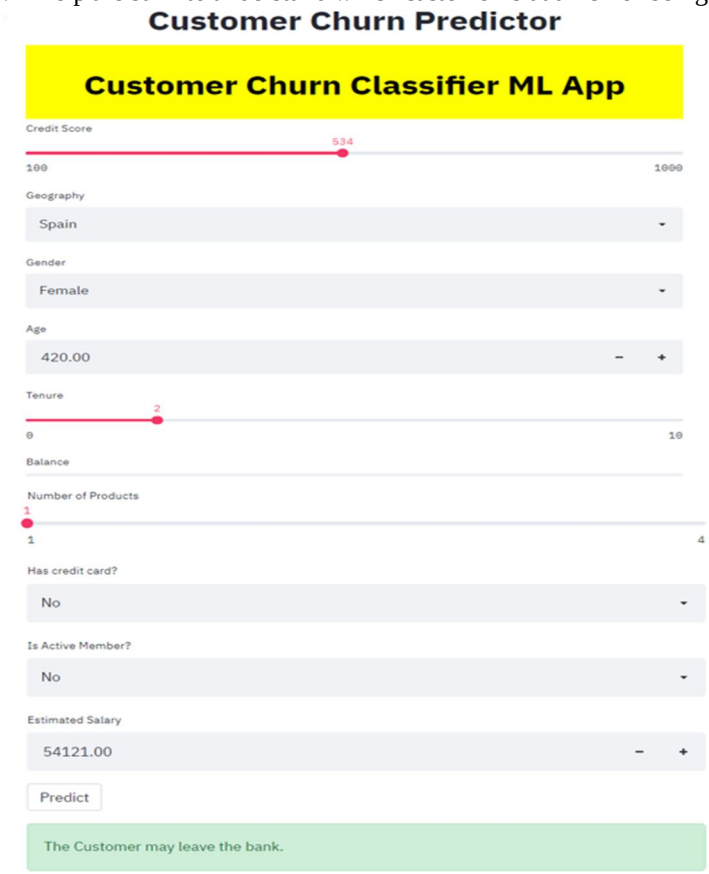


Fig. 3 Customer Churn Predictor Web App

©IJRASET: All Rights are Reserved

VII. CONCLUSION And FUTURE ENHANCEMENTS

The banking sector around the world has been undergoing significant structural transformations in its business model to achieve profitability goals, after an extended period of time passed by with low interest rates. In an attempt to resist low financial margins, there has been a substantial rise in the commissions to improve the revenue. Due to the thought to shift towards digital transactions there has been a change of operations and shutting down of various branches across the country. Certain decisions like these have had an enormous effect on customer contemplation and significant increment in churn rates.

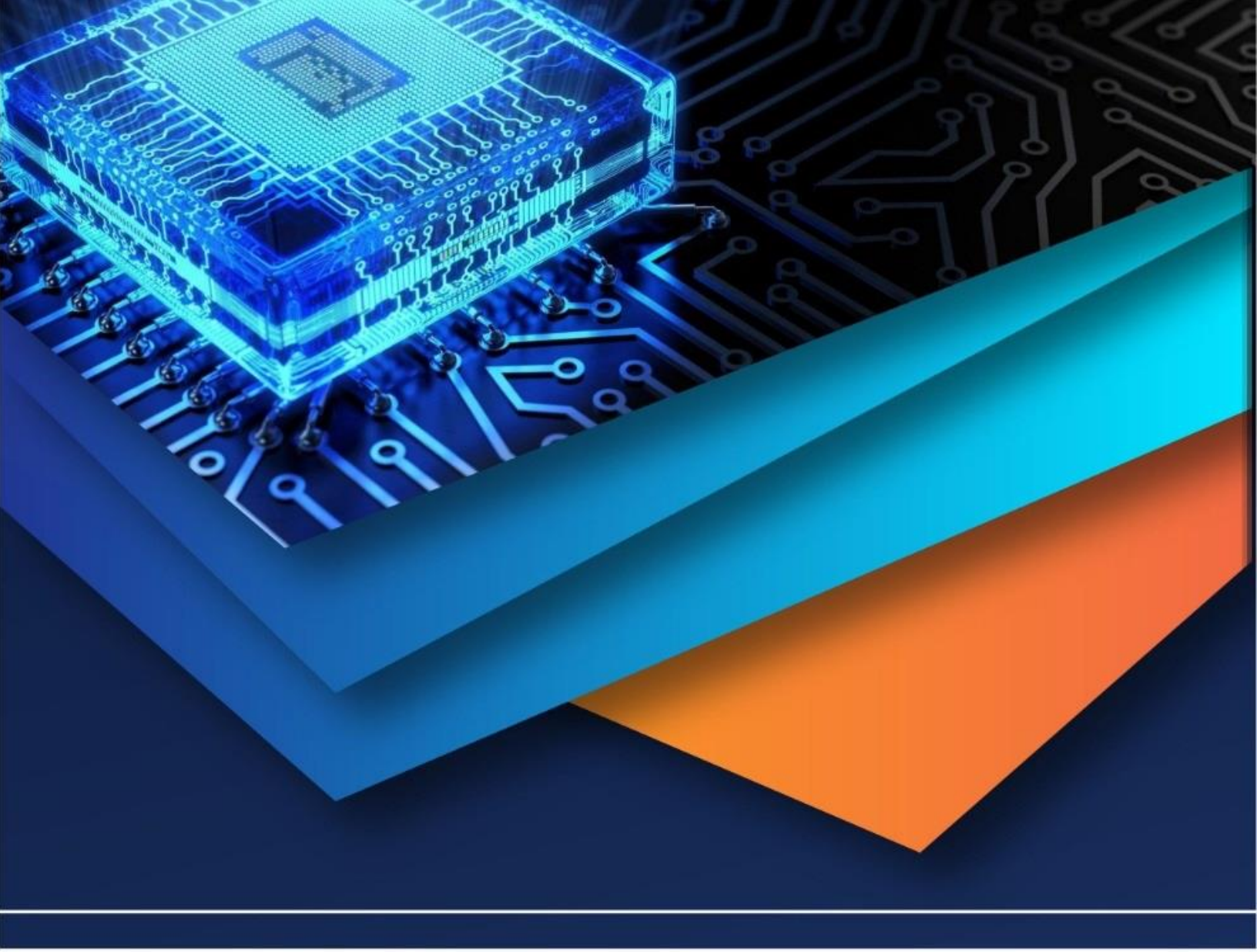
At the end, this method proposed in the publication attained its main goal to predict the customer churn rate effectively and reliably. It has provided a way to keep a timely check on the customers and their interaction with the bank. As and when the bank identifies certain customers who are having the risk of being churned, the bank can put-forth certain schemes and marketing strategies to keep the customer involved and engaged with the bank. Thus, a bank can be successful in retaining its customers and those customers in turn will eventually bring in new customers to the bank.

The current work presented in this paper takes into consideration the datasets from a certain period of time and not the entire data since the establishment of the bank. Therefore, the obtained results might be biased with respect to the data available to us during the time of prediction. In the end, the proposed system of prediction can be enhanced by getting access to the different datasets from various years since the inception of the bank. Thus, improving the accuracy and reliability of the prediction resulting from our machine learningmodel.

REFERENCES

[1] AMIN, Adnan, et al. Customer churn prediction in telecommunication industry using data certainty. Journal of Business Research, 2019, 94:290-301.
[2] Qureshii SA, Rehman AS, Qamar AM, Kamal A, Rehman A. Telecommunication subscribers churn prediction model using machine learning. In: Eighth international conference on digital information management. 2013. P.131—6.
[3] ULLAH, Irfan, et al. A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. IEEE Access, 2019, 7: 60134-60149.
[4] Umayaparvathi V, Iyakutti K. A survey on customer churn prediction in telecom industry: datasets, methods and metric. Int Res J Eng Technol. 2016; 3(4):1065—70.
[5] Abbasimehr H, Setak M, Tarokh M (2011) A neuro-fuzzy classifier for customer churn prediction. International Journal of Computer Applications 19(8):35–41.
[6] Asthana P (2018) A comparison of machine learning techniques for customer churn prediction. International Journal of Pure and Applied Mathematics 119(10):1149–1169.
[7] Burez J, Van den Poel D (2009) Handling class imbalance in customer churn prediction. Expert Systems with Applications 36(3):4626–4636.
[8] Coussement K, De Bock KW (2013) Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. Journal of Business Research 66(9):1629–1636
[9] Hadden J, Tiwari A, Roy R, Ruta D(2006) Churn prediction: Does technology matter. International Journal of Intelligent Technology 1(2):104–110.
[10] Hadden J, Tiwari A, Roy R, Ruta D (2007) Computer assisted customer churn management: State-of-the-art and future trends. Computers & Operations Research 34(10):2902– 2917.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)