Journal of Population Therapeutics & Clinical Pharmacology

RESEARCH ARTICLE
DOI: 10.47750/jptcp.2023.30.11.039

Prediction Of Churning Customers for Bank and Telecommunication Sectors

K.Saranya[1*], S.Vijayashaarathi[2], N.Sasirekha[3], K.Dinesh[4], K.Hariharan[5]

[1,2,3]Assistant Professor, Department of ECE Sona College of Technology(An Autonomous Institution) Salem, India

[4,5]UG Scholar, Department of ECE Sona College of Technology(An Autonomous Institution) Salem, India

***Corresponding author:** K.Saranya, Assistant Professor, Department of ECE Sona College of Technology(An Autonomous Institution) Salem, India, Email: saranya.k@sonatech.ac.in

Submitted: 27 March 2023; Accepted: 16 April 2023; Published: 09 May 2023

ABSTRACT

In this paper the prediction of churn customers and non-churn customers were done. Basically a single customer is very important for an industry, or for organization etc... This project will predict or identify the churn customers and the non-churn customers in the given dataset of the organization. Some Machine Learning algorithms are used in this project to predict the churn customers range and accuracy of the algorithm in finding the churn rate. We have created a web based app, which requires some details about the customers of the organization to predict whether Churn or Non churn customer.

Keywords: *KNN, AUROC, Support Vector Machine, Logistic Regression, Decision Tree, Random Forest*

INTRODUCTION

Customer churn is the biggest issue faced by the companies or industries. Preventing the customers from churning, It is now crucial for business growth and development to try to keep clients. Because each and every customers are important to the industries or the organization. The main reason for the customers to churn the company is due to not satisfied on the industries service or not satisfied on the product of the company.

For Bank sector: Attrition of customers is also referred to as churn customer. It means customer decide to leave the organization and decide to stop making a relationship with the organization or stop using the products of an company during the specific period of time is know as churning rate. This customer churn will directly affect the revenue of the company.

The need for prediction of the churning customer project is to identify the customer those who are "Not Satisfy" (churn customer) from the service provided by the organization. By observing the dataset and analyzing methods were performed to identify whether the customer is agitate or continue with the service provide by the company. So, that analysis of data is used for observing the data to find the valuable information using machine learning and data mining techniques. The banking industry faced more difficulties to hold customers, Based on various reasons the customers will move towards the other bank, for example, other banks may provide better financial

service at low charge, location of the bank, high quality services, low rate of interest, etc...

Thus, the prediction model is used to identify clients who are likely to leave in the future. These models use historical churner data to look for patterns that might be shared with current clients. If any commonalities are discovered, the current customers are then labeled as churners. It is difficult for the organization to bring the new customer as like as old customer for their organization. Also, serving long-term clients is simpler and less expensive than losing a client, which causes the bank to lose profit. The old customers may give referrals for the organization which is a high benefit for the organization. In this study, various machine learning models, including logistic regression (LR), decision trees (DT), and K-nearest neighbours (KNN), are applied to the dataset provided by the company for prediction. The effectiveness of the models is evaluated by examining the accuracy, recall, and other factors, and the results are presented.

EXISTING SYSTEM

In most of the customer churn model they have used algorithm like decision tree to find out whether the customer is churn or not churn. But the decision tree algorithm is not always suitable for the situations or problem which is in complex form. By reducing the data's in the dataset will increase the accuracy of the decision tree. But according to Machine Learning algorithms "the more data, the more Accuracy".

PROPOSED METHOD

In this paper, we employed a number of machine learning methods, including the random forest algorithm (RF), the decision tree (DT), the logistic regression algorithm (LR), and the support vector machine (SVM) [5] and determine which algorithm is best suitable for our project for the given dataset with highest accuracy among the above algorithms. Because in Machine learning the accuracy of the algorithms will differ for different datasets provided by the organization.

The very first step in this project is data preprocessing in which the filtering of data and converting the data into similar form is done and then we make further selection. In the next step we have used the algorithm which gave the highest accuracy among the above algorithms to do the classification process. Training data and testing data are separated from the dataset. The20 to 25% of the dataset is utilized to test the model, with the rest 75 to 80% used as training data. After the classification step, we do the analysis on the result obtained from the algorithm.

METHODOLOGY

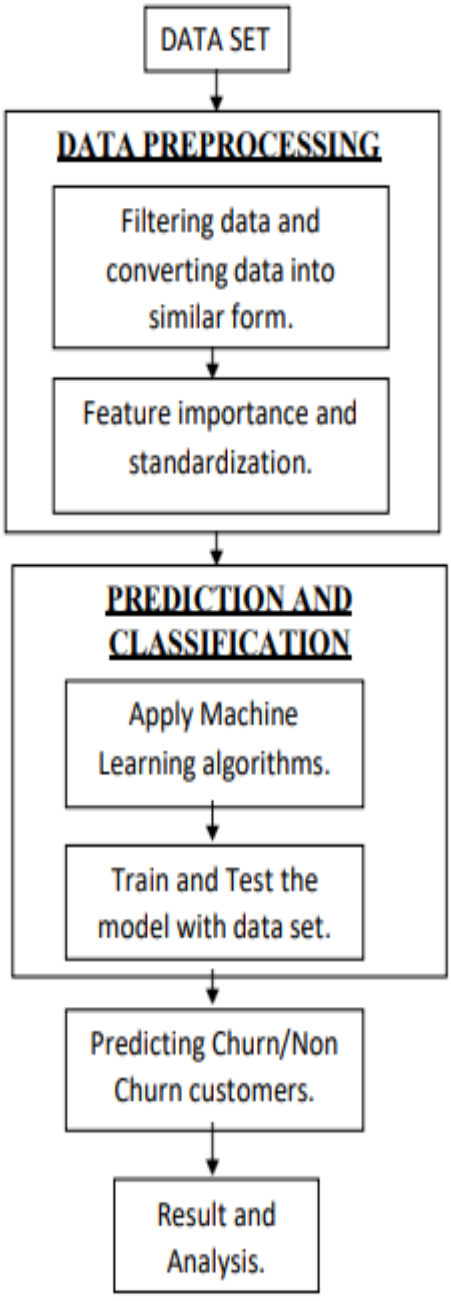
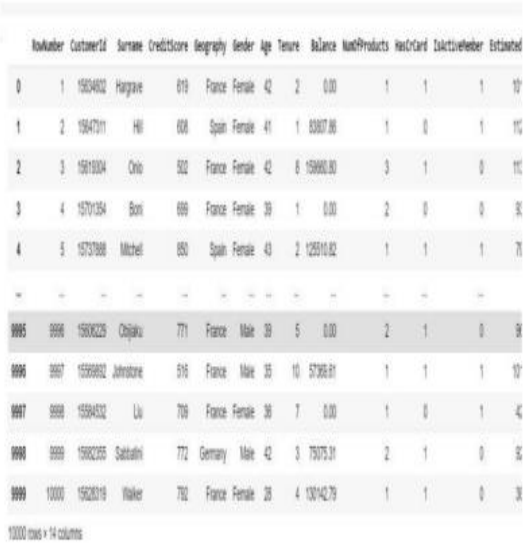


FIG 1: Methodology

J Popul Ther Clin Pharmacol Vol 30(11):e385–e391; 09 May 2023.

A) Dataset

The dataset is the starting point in this project. It contains the information about the customers. For example: name of the customer, age, No. of products purchased or NO. of accounts, credit score of the customer etc...By using the dataset only we can train the machine to give the better output for the problem or situation.



	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	Estimated
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	10
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	11
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	11
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	9
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	7
...	...	...	...	...	...	...	...	...	...	...	...	...	
9995	9996	15606229	Obijiaku	771	France	Male	39	5	0.00	2	1	0	9
9996	9997	15569892	Johnstone	516	France	Male	35	10	57369.61	1	1	1	10
9997	9998	15584532	Liu	709	France	Female	36	7	0.00	1	0	1	4
9998	9999	15682355	Sabbatini	772	Germany	Male	42	3	75075.31	2	1	0	9
9999	10000	15628319	Walker	792	France	Female	28	4	130142.79	1	1	0	3

10000 rows × 14 columns

FIG 2: Sample dataset

B) *Data Preprocessing*

B. I. Dataset Filtering

Basically the dataset is the collection objects, values etc about the customer. Sometimes the Datasets may contain null values and also may contain duplicate values. It will decrease the accuracy of the algorithm [14]. So, the Process of Dataset Filtering is the process of removing the null values and duplicate rows and values in the dataset.

B. II. Splitting The Dataset

In splitting the dataset process the content of the dataset is splitted two separate parts. The first is a TRAINING SET, while the second is a TESTING SET. The training set is utilized to teach the machine how to get the desired results. Additionally, the testing set is utilized to determine whether or not the machine is producing the required or accurate results

```
# Splitting the Dataset into Dependent and Independent Variables
X = final_dataset.iloc[:, [0,1,2,3,4,5,6,7,9,10,11]]
y = final_dataset.iloc[:, 8].values
```

```
[33] X.head()
```

	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Geography_Germany	Geography_Spain	Gender_Male
0	619	42	2	0.00	1	1	1	101348.88	0	0	0
1	608	41	1	83807.86	1	0	1	112542.58	0	1	0
2	502	42	8	159660.80	3	1	0	113931.57	0	0	0
3	699	39	1	0.00	2	0	0	93826.63	0	0	0
4	850	43	2	125510.82	1	1	1	79084.10	0	1	0

FIG 3: Splitting dataset

C) *Prediction And Classification*

C. I. Logistic Regression

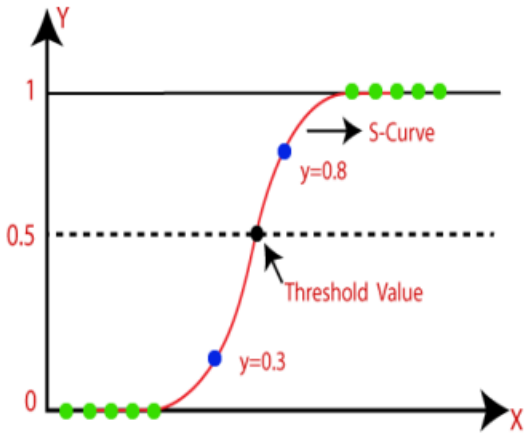


FIGURE 4: Logistic Regression.

One of the most often used Machine Learning algorithms is logistic regression. The Supervised Learning approaches include this algorithm. Using logistic regression, a dependent variable's output is predicted. The Logistic Regression's output resembles discrete values [12]. It may be Yes or No, true or false, 0 or 1, etc., but instead of giving an exact number between 0 and 1, it gives probabilistic values that are between 0 and 1.

C. II. Decision Tree

Decision Tree is one of the famous supervised algorithm.

It is the classifier algorithm with tree structure. Each leaf node in a tree structure reflects the result, whereas branches indicate the decision-making processes [2]. The two nodes in a decision tree are the Decision Node and Leaf Node.

J Popul Ther Clin Pharmacol Vol 30(11):e385–e391; 09 May 2023.

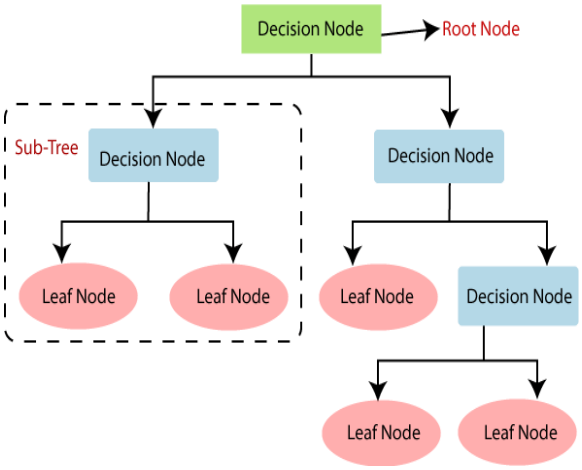


FIG 5: Decision tree

C. III. Random Forest Algorithm

Another supervised learning method is the Random Forest algorithm. It can be applied to machine learning tasks involving both classification and regression [13]. Random Forest is a classifier that takes the average of several decision trees on a given dataset to increase the dataset's predictive accuracy [11]

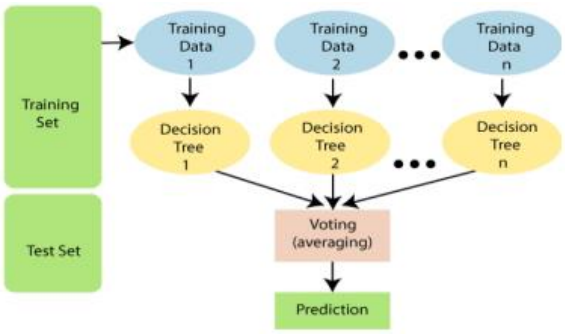


FIG 6: Random Forest algorithm structure

C. IV. K-Nearest Neighbor (KNN)

One of the most basic algorithms for machine learning and a supervised learning method is K-Nearest Neighbor. Assuming that the new and old data are comparable, the K-NN method places the new sample in the category that matches the old categories the closest. Once all the data has been recorded, a new data point is categorized using the K-NN technique based on similarity [3].

It can be applied to problems involving classification as well as regression. It is also known as a lazy learner algorithm since it saves the training dataset rather than learning from it immediately. Instead, it performs an action while classifying data by using the dataset.

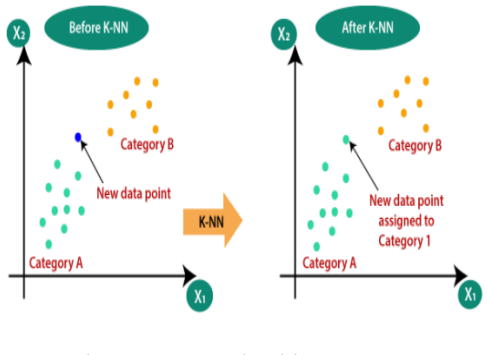


FIG 7: Random Forest algorithm structure

Standardization

When continuous independent variables are measured at various scales, the idea of standardization is brought into play [9]. Machine learning algorithms that weight inputs often use optimization techniques like gradient descent. One such technique is to rescale values using a distribution value between 0 and 1.

Data Visualization

Data summaries, test data analysis, and model output analysis are only a few of the numerous analytical activities for which data visualization is crucial [10]. Seeing the best in others is one of the simplest ways to establish a connection.

FIG 8: Data Visualization

Correlation Between All Features Heat Map

A heat map is a visual representation of multivariate data in a matrix of columns and

J Popul Ther Clin Pharmacol Vol 30(11):e385–e391; 09 May 2023.

rows. The association between various numerical variables can be described using heat maps, which can help to highlight patterns and abnormalities.

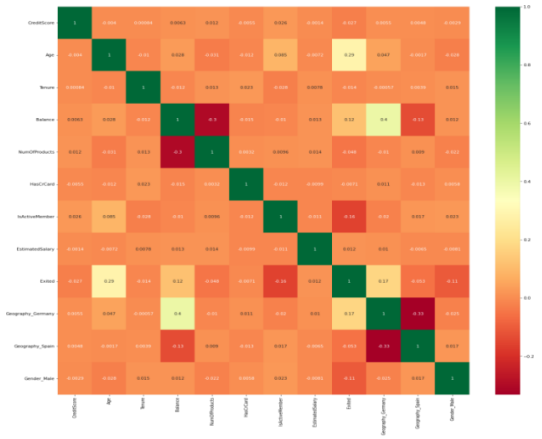


FIG 9: HEAT MAP

RESULT AND DISCUSSION

Logistic Regression Output

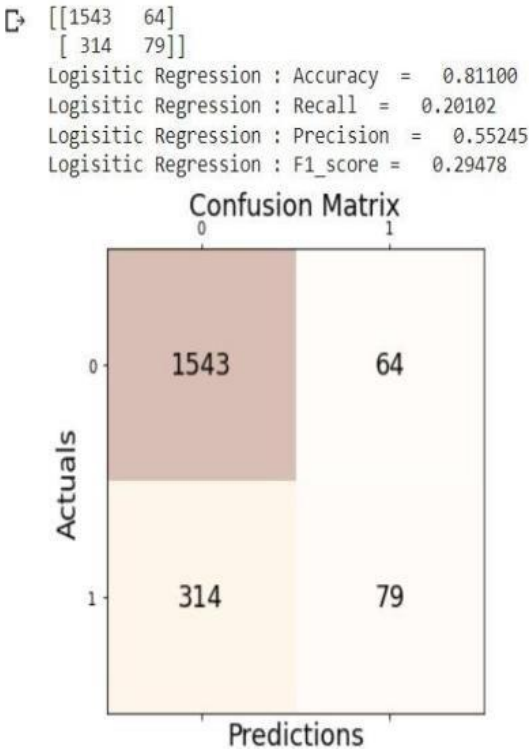


FIG 10: Confusion Matrix for Logistic Regression

Accuracy : 81%

Random Forest Algorithm Output

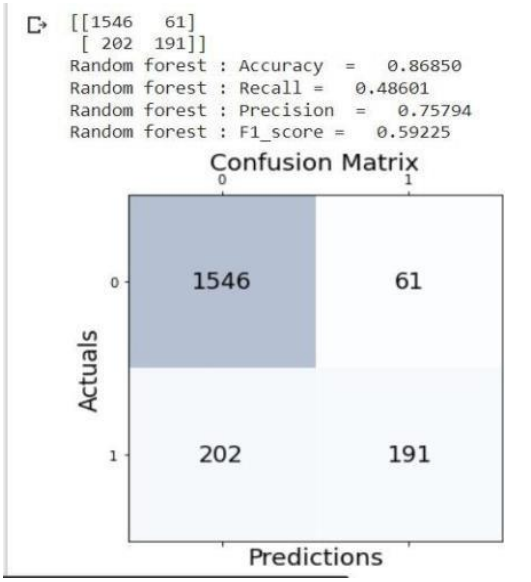


FIG 11: Confusion Matrix for Random Forest

Accuracy : 86%

Decision Tree Output

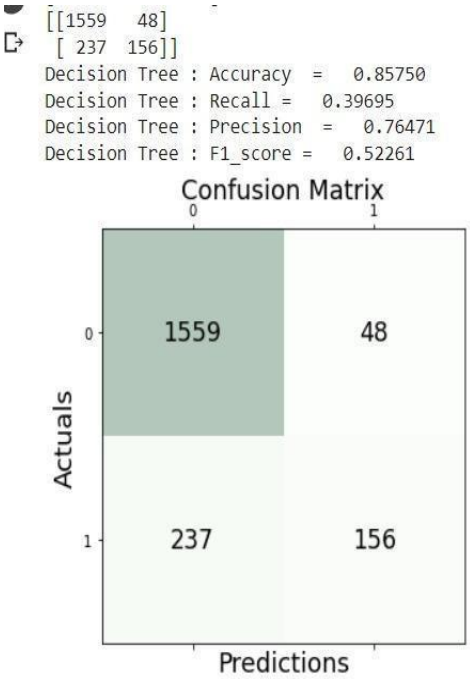


FIG 12: Confusion Matrix for Decision Tree.

Accuracy : 85%

J Popul Ther Clin Pharmacol Vol 30(11):e385–e391; 09 May 2023.

Overall Accuracy

TABLE 1: Overall Accuracy for algorithms

Algorithm	Accuracy	Auroc Score
Random Forest	86.8%	85.5%
Logistic Regression	81%	77%
Decision Tree	85.7%	83%
K-Nearest Neighbors	83%	73.3%

Output

Confusion Matrix Of Knn

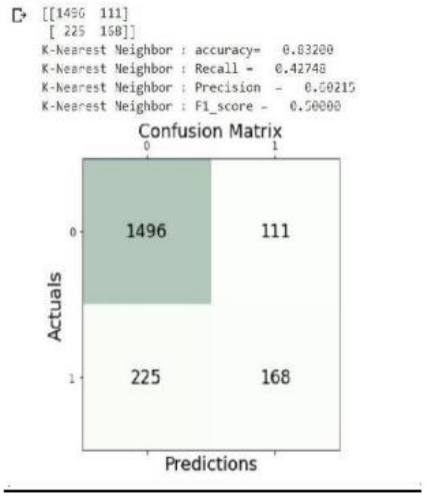


FIG 12: Confusion Matrix for K-NN

Accuracy And Auroc Score

FIG 13: AUROC score for Random forest,Decision tree, Logistic Regression

Auroc Curve

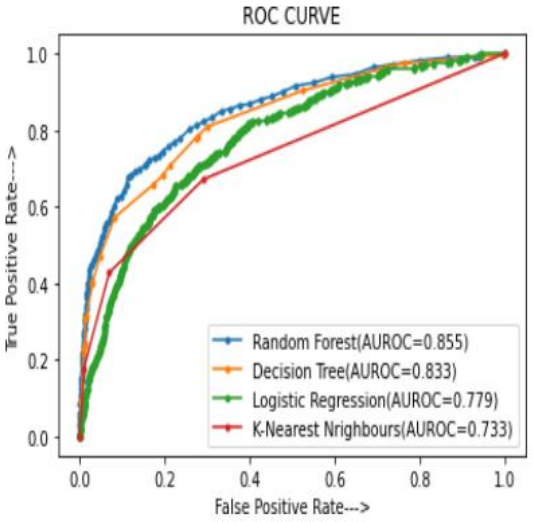


FIG 14: AUROC curve for Random Forest, Decision Tree Logistic Regression and K-NNalgorithm

Advantages

- Best tool to identify the customer satisfaction.
- Easy to use.
- High Accuracy.
- No need of high end computers.
- Best for the big industries which want to know the number of churn customers and non customers to enhance their business [15].

CONCLUSION

The importance of this type of project in the bank, Telecommunication Company, etc... is to help companies to identify the customers whether they satisfied with the service provide by the organization or not. It is well recognised that one of the most significant sources of revenue for businesses is the ability to predict attrition [6]. Therefore, the main goal of this paper was to develop a system that can forecast client attrition. To test and train the model, the sample dataset is divided into 80% for training and 20% for testing We have used four different algorithms in this paper. they are Logistic Regression, Decision Tree, K–Nearest Neighbor, and Random Forest algorithms. In this Random Forest algorithm predict the range of churners in the dataset and gives the higher accuracy of 86%[1].

J Popul Ther Clin Pharmacol Vol 30(11):e385–e391; 09 May 2023.

REFERENCES

1. Abhishek and Ratnesh ,"Predicting Customer Churn Prediction in Telecom Sector Using Various Machine Learning Techniques", In the Proceedings of 2017 International Conference on Advanced Computation and Telecommunication, Bhopal, India, 2017.
2. Kiran and Surbhi , "Customer Churn Analysis in Telecom", Industry International Conference for Reliablity,Noida , India , 2015.
3. L. Ning, L. Hua, L. Jie, Z. Guangquan, "A customer churn prediction model in telecom industry using boosting", IEEE Trans. Ind. Inform. 10 (2014) 1659– 1665.
4. T.Vafeiadis K.I Diamantaeas, G.Sarigiannidis K.Chatzisavvas "Customer churn prediction in telecommunications", Simulation Modelling: Practice and Theory 55 (2015) 1-9.
5. M.A.H. Farquad, Vadlamani Ravi, S. Bapi Raju "Churn prediction using comprehensible support vector machine: An analytical CRM application", Applied Soft Computing 19 (2014) 31–40.
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. Brânduşoiu, I., Toderean, G., Beleiu, H.: Methods for churn prediction in the pre-paid mobile telecommunications industry. In: 2016International conference on communications (COMM), pp. 97– 100. IEEE (2016)
8. Dahiya, K., Bhatia, S.: Customer churn analysis in telecom industry. In: 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), pp. 1–6 (2015)
9. Chin-Ping Wei and I-Tang Chiu.:The churn prediction technique for customer retention analysis.
10. Gürsoy UŞ (2010) Customer churn analysis in telecommunication sector. İstanbul Üniversitesi İşletme Fakültesi Dergisi 39(1):35–49
11. Zainab Jamal, Randolph E. Bucklin proposeda paper on "Improving the diagnosis and prediction of customer churn". This tells about how to improve the prediction of the system.
12. John Hadden, Ashutosh Tiwari, Rajkumar Roy, and Dymitr Ruta .: Churn Prediction: Does Technology Matter?.
13. Chris Rygielski, Jyun-Cheng Wang, David and C.Yen, discuss Neural Networks as data mining technique for customer relationship management (2002) neural networks provide a more powerful and predictive model than other techniques.
14. Qureshi, S.A., Rehman, A.S., Qamar, A.M., Kamal, A., Rehman, A.: "Telecommunication subscribers' churn prediction model using machine learning.
15. Yogesh Beeharry,Ristin Tsokizep Fokone "Hybrid approach using machine learning algorithms for customers' churn prediction in the telecommunications industry.