PREDICTING CUSTOMER CHURN PREDICTION IN TELECOM SECTOR USING SVM AND RANDOM FOREST

Kirti Malviya[1] , Syed Rameez Ali[2]
[1]Mtech Scholar
[2]Asst. Professor
[1],[2]Department of Computer Science & Engineering, OCT, Bhopal

Abstract: Predicting customer churn in telecommunication industries becomes a most important topic for research in recent years. Because its helps in detecting which customer are likely to change or cancel their subscription to a service. Analysis of data which is extracted from telecom companies can helps to find the reasons of customer churn and also uses the information to retain the customers. So predicting churn is very important for telecom companies to retain their customers. So data mining techniques and algorithm plays an important role for companies in today's commercial conditions because gaining a new customer's cost is more than retaining the existing ones. In this paper we can focuses on machine learning techniques for predicting customer churn through which we can build the classification models such as SVM and Random Forest and also compare the performance of these models.

Keywords: Churn prediction, data mining, telecom system ,Customer retention, classification system, random forest, svm.

1.1. Introduction

In today's technological conditions, new data are being produced by different sources in many sectors. However, it is not possible to extract the useful information hidden in these data sets, unless they are processed properly. In order to find out these hidden information, various analyses should be performed using data mining, which consists of numerous methods.[6]

The Churn Analysis [4] aims to predict customers who are going to stop using a product or service among the customers. And, the customer churn analysis is a data mining based work that will extract these possibilities. Today's competitive conditions led to numerous companies selling the same product at quite a similar service and product quality.

With the Churn Analysis[7], it is possible to precisely predict the customers who are going to stop using services or products by assigning a probability to each customer. This analysis can be performed according to customer segments and amount of loss (monetary equivalent). Following these analyses, communication with the customers can be improved in order to persuade the customers and increase customer loyalty. Effective marketing campaigns for target customers can be created by calculating the churn rate or customer attrition. In this way, profitability can be increased significantly or the possible damage due to customer loss can be reduced at the same rate .

1.2. Related Work

According to (Peng Li, et al , 2017) [1] , From the beginning of the data mining [9] which is used to discover new knowledge's from the databases can helping various problems and helps the business for their solutions. Telecom companies improve their revenue by retaining their customers Customer churn in telecom sector is to leave a one subscription and join the other subscription In these paper they predicting the customer churn by using various R packages and they created a classification model and they train by giving him a dataset and after training they can classify the records into churn or non churn and then they visualize the result with the help to visualization techniques[10]. According to (Chuanqi Wang, et al, 2017) [2] , Telecom Customer churn prediction is a cost sensitive classification problem. Most of studies regard it as a general classification problem use traditional methods, that the two types of misclassification cost are equal. And, in aspect of cost sensitive classification, there are some researches focused on static cost sensitive situation. In fact, customer value of each customer is different, so misclassification cost of each sample is different. For this problem, we propose the partition cost-sensitive CART model in this paper[12]. According to (Kiran Dahiya, et al, 2015) [5] Customer churn plays an important role in customer relationship management (CRM), and they are using various machine learning algorithm to predict customer churn and they found ensemble learning is an best to predict customer churn, but there exist still a lot of problems like how they choose the method of integration and how to choose the strategy, which makes the final ensemble classifier

1.3. Problem Definition

From the problems obligatory through market saturation and value implications, there has been associate identification of a desire for a computer based mostly churn prediction methodology that's capable of accurately distinctive a loss of client ahead, so proactive retention ways is deployed during a bid to retain the client. The churn prediction should be correct as a result of retention ways is pricey [13]. A limitation of current analysis is that alternative studies have focused virtually solely on churn capture, neglecting the problem of misclassification of non-churn as churn. Retention campaigns usually embrace creating service based mostly offers to customers during a bid to retain them.[14]

1.4 Proposed Work

In the proposed system R [8] programming will be used to build the model for churn prediction. It is widely used among statisticians and data miners for developing statistical software and data analysis. R is freely available and a powerful statistical analysis tool which has not yet been explored for building models for churn prediction[3].

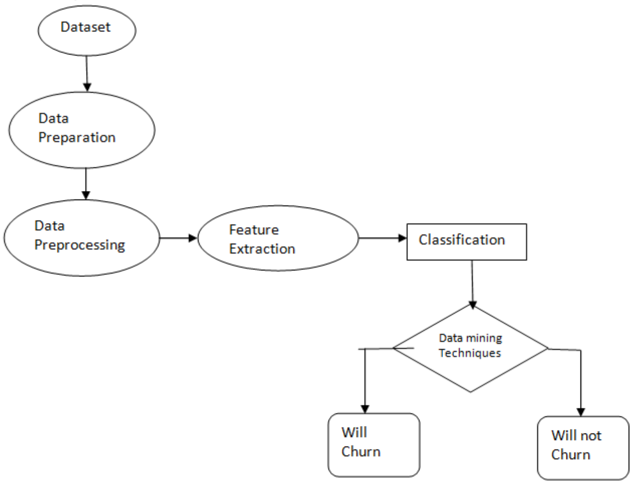


Figure 1. Churn Prediction Framework

In this paper, we proposed different machine learning algorithms to analyze customer churn analysis. Through which we can multiple different models are employed to accurately predict those churn customers in the data set. These models are Support Vector Machine and Random Forest. Our Steps or Algorithm Steps will follow:

1. Dataset:- A telecom dataset is taken for predicting churn which to identify trends in customer churn at a telecom company and the data which we taken is in .csv format. The data given to us contains 7043 observations and 21 variables extracted from a datasets.

2. Data Preparation: Since the dataset acquired cannot be applied directly to the churn prediction models, so we can naming each attributes.

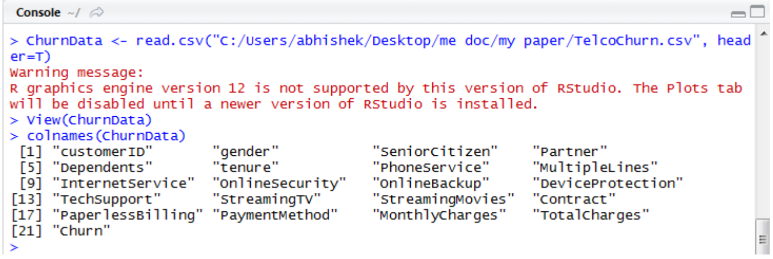
3. Data Preprocessing: Data preprocessing is the most important phase in prediction models as the data consists of ambiguities, errors, redundancy and transformation which needs to be cleaned beforehand.

4. Data Extraction: The attributes are identified for classifying process.

5. Decision: Based on data extraction and classification models we can take a decision whether the employee is churner or not.

1.5 Result Analysis

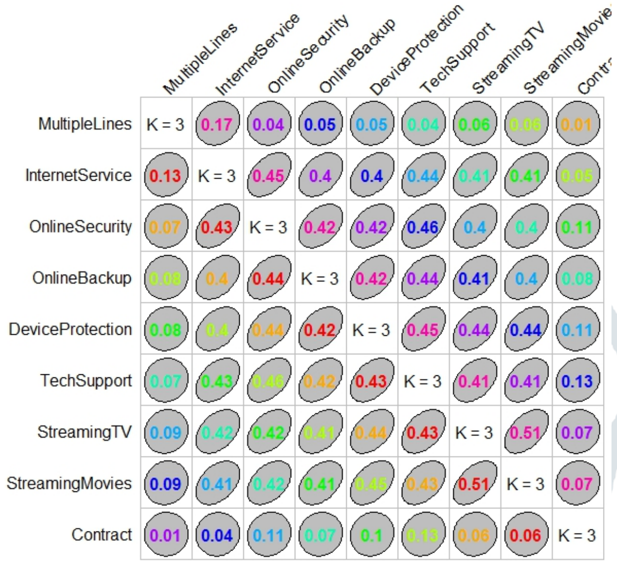
All the experiments were performed using an i5-2410M CPU @ 2.30 GHz processor and 4 GB of RAM running Windows. After that we can install r and Rstudio and then to identify trends in customer churn at a telecom company. The data given to us contains 7043 observations and 21 variables extracted from a data warehouse. These variables are shown in figure 2.



```
> ChurnData <- read.csv("C:/Users/abhishek/Desktop/me doc/my paper/TelcoChurn.csv", head
er=T)
Warning message:
R graphics engine version 12 is not supported by this version of RStudio. The Plots tab
will be disabled until a newer version of RStudio is installed.
> View(ChurnData)
> colnames(ChurnData)
 [1] "customerID"       "gender"           "SeniorCitizen"    "Partner"
 [5] "Dependents"       "tenure"           "PhoneService"     "MultipleLines"
 [9] "InternetService"  "OnlineSecurity"   "OnlineBackup"     "DeviceProtection"
[13] "TechSupport"      "StreamingTV"      "StreamingMovies"  "Contract"
[17] "PaperlessBilling" "PaymentMethod"    "MonthlyCharges"   "TotalCharges"
[21] "Churn"
>
```

Figure-2. Variables or sample values in datasets

Now we started to exploring a data and cleaning a data for machine learning models, we can explore the data by their multiple attributes and find the Correlation between these categorical variables figure 3 shows the relation between these variables.



	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovie	Contr
MultipleLines	K = 3	0.17	0.04	0.05	0.05	0.04	0.06	0.06	0.01
InternetService	0.13	K = 3	0.45	0.4	0.4	0.44	0.41	0.41	0.05
OnlineSecurity	0.07	0.43	K = 3	0.42	0.42	0.46	0.4	0.4	0.11
OnlineBackup	0.08	0.4	0.44	K = 3	0.42	0.44	0.41	0.4	0.08
DeviceProtection	0.08	0.4	0.44	0.42	K = 3	0.45	0.44	0.44	0.11
TechSupport	0.07	0.43	0.46	0.42	0.43	K = 3	0.41	0.41	0.13
StreamingTV	0.09	0.42	0.42	0.41	0.44	0.43	K = 3	0.51	0.07
StreamingMovies	0.09	0.41	0.42	0.41	0.45	0.43	0.51	K = 3	0.07
Contract	0.01	0.04	0.11	0.07	0.1	0.13	0.06	0.06	K = 3

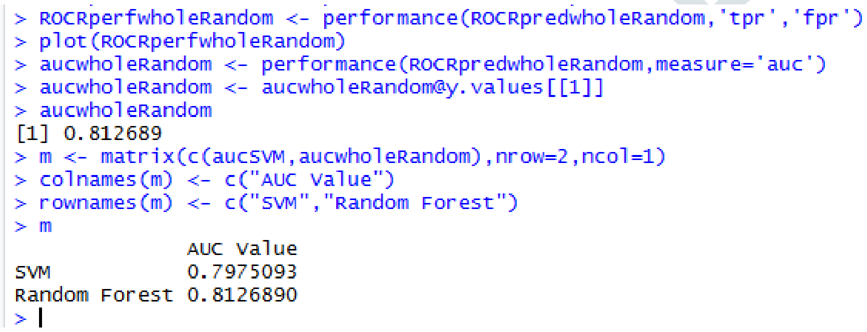
Figure 3. correlation between categorical variables

These graph is to explore relationships between categorical variables and which look like the only highly correlated variables are "streaming movies" and "streaming tv" which is expected.

PREPARATION FOR THE MODEL BUILDING

Now we can build a machine learning models, such as SVM and Random Forest and then we train these classifier and after training we can compute the performance of model and compare their performance. Before starting training we can perform variable selection: We know that we should not include "streaming movies" and "streaming tv" in the same equation. Their correlation from the above section is fairly high.

Now we can split the data into train (75%) and test (25%) dataset and then start learning of model on these train data and we can first compute the performance of both the models on training dataset. Figure 4 shows the computing performance of these two machine learning models.



```
> ROCRperfwholeRandom <- performance(ROCRpredwholeRandom,'tpr','fpr')
> plot(ROCRperfwholeRandom)
> aucwholeRandom <- performance(ROCRpredwholeRandom,measure='auc')
> aucwholeRandom <- aucwholeRandom@y.values[[1]]
> aucwholeRandom
[1] 0.812689
> m <- matrix(c(aucSVM,aucwholeRandom),nrow=2,ncol=1)
> colnames(m) <- c("AUC Value")
> rownames(m) <- c("SVM","Random Forest")
> m
              AUC Value
SVM           0.7975093
Random Forest 0.8126890
> |
```

Figure 4. AUC values of models

According to the AUC values which we have computed, the method that gives us the most accurate model is Random Forest with AUC value of 81.24%. Figure 5 show the AUC curve of these models.

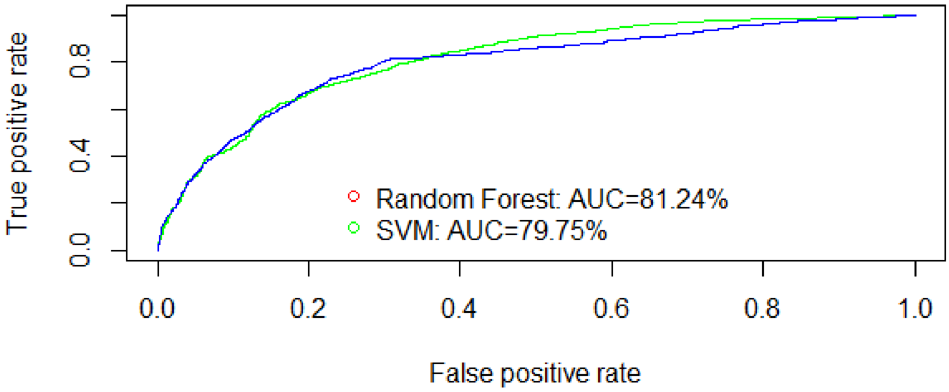
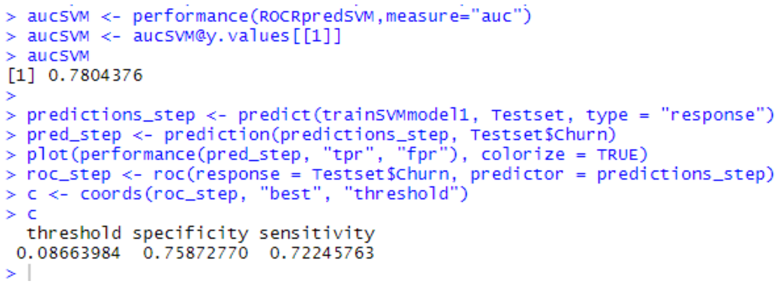


Figure 5. AUC of both the models

According to the AUC curves, the method that gives us the most accurate model is Random Forest with AUC value of 81.24%.

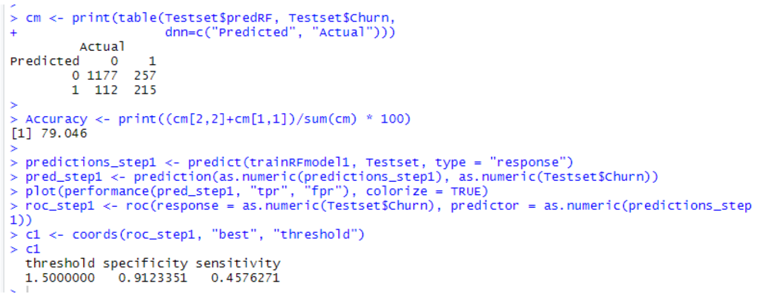
TESTING THE MODELS

After the models gets trained on the training datasets and tested on train dataset we found that random forest perform better, Now we test these models on the test datasets on different performance parameters like specificity and sensitivity. These models predicts the labels as churn or not churn and compared these predicted outcomes to the actual outcomes we gets the performance measures of these models. The testing results of both the models are shown in figure 6 and figure 7.



```
> aucSVM <- performance(ROCRpredSVM,measure="auc")
> aucSVM <- aucSVM@y.values[[1]]
> aucSVM
[1] 0.7804376
>
> predictions_step <- predict(trainSVMmodel1, Testset, type = "response")
> pred_step <- prediction(predictions_step, Testset$Churn)
> plot(performance(pred_step, "tpr", "fpr"), colorize = TRUE)
> roc_step <- roc(response = Testset$Churn, predictor = predictions_step)
> c <- coords(roc_step, "best", "threshold")
> c
  threshold specificity sensitivity
 0.08663984  0.75872770  0.72245763
> |
```

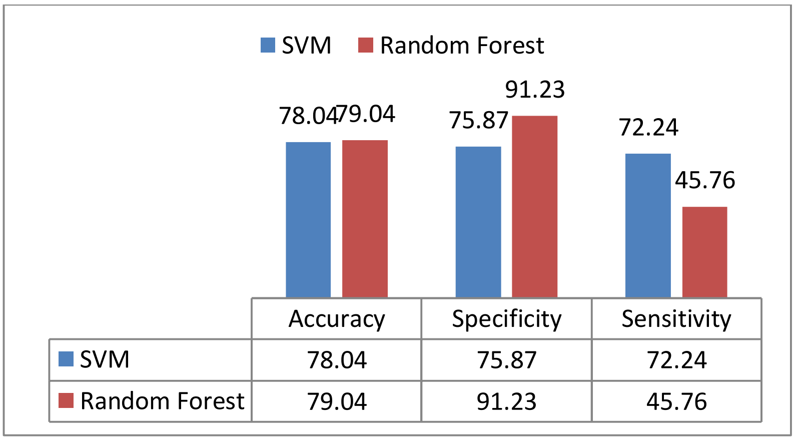
Figure 6. Performance measure of SVM model



```
> cm <- print(table(Testset$predRF, Testset$Churn,
+                   dnn=c("Predicted", "Actual")))
        Actual
Predicted    0    1
        0 1177  257
        1  112  215
>
> Accuracy <- print((cm[2,2]+cm[1,1])/sum(cm) * 100)
[1] 79.046
>
> predictions_step1 <- predict(trainRFmodel1, Testset, type = "response")
> pred_step1 <- prediction(as.numeric(predictions_step1), as.numeric(Testset$Churn))
> plot(performance(pred_step1, "tpr", "fpr"), colorize = TRUE)
> roc_step1 <- roc(response = as.numeric(Testset$Churn), predictor = as.numeric(predictions_step
1))
> c1 <- coords(roc_step1, "best", "threshold")
> c1
  threshold specificity sensitivity
  1.5000000   0.9123351   0.4576271
```

Figure 7. Performance measure of Random Forest model

Based on the AUC curve and the performance measures of both the models we can compare the performance of the models and the conclusion of the performance are shown in figure 8.



	Accuracy	Specificity	Sensitivity
🟦 SVM	78.04	75.87	72.24
🟥 Random Forest	79.04	91.23	45.76

Figure 8. Performance on the models

1.6 Conclusion

In order to retain existing customers, Telecom providers need to know the reasons of churn, which can be realized through the knowledge extracted from Telecom data. In this paper, we train two machine learning models which is SVM and Random Forest and we can say that Random Forest is perform better as compared to SVM because it provides better accuracy and specificity but in terms of sensitivity the SVM model perform better compared to Random Forest.

References:

[01] *Peng Li 1, 2, Siben Li 2, Tingting Bi 2, Yang Liu 2, "* Telecom Customer Churn Prediction Method Based on Cluster Stratified Sampling Logistic Regression" *in IEEE.*

[02] Chuanqi Wang, Ruiqi Li, Peng Wang, Zonghai Chen, “Partition cost-sensitive CART based on customer value for Telecom customer churn prediction” in Proceedings of the 36th Chinese Control Conference 2017 IEEE.

[03] Guo-en Xia, Hui Wang, Yilin Jiang, “Application of Customer Churn Prediction Based on Weighted Selective Ensembles” in IEEE 2016.

[04] Rahul J. Jadhav, Usharani T. Pawar, “Churn Prediction in Telecommunication Using Data Mining Technology”, in *(IJACSA), Vol. 2, No.2, February 2011*

[05] Kiran Dahiya, Surbhi Bhatia, “Customer Churn Analysis in Telecom Industry” in IEEE 2015, 978-1-4673-7231-2/15

[06] N.Kamalraj, A.Malathi’ “ A Survey on Churn Prediction Techniques in Communication Sector” in *IJCA Volume 64– No.5, February 2013*

[07] Kiran Dahiya,KanikaTalwar, “Customer Churn Prediction in Telecommunication Industries using Data Mining Techniques- A Review” in IJARCSSE, Volume 5, Issue 4, 2015.

[08] R Data: http://cran.r-project.org/

[09] Data Mining in the Telecommunications Industryǁ, Gary M. Weiss, Fordham University, USA.

[10] Manjit Kaur et al., 2013.Data Mining as a tool to Predict the Churn Behaviour among Indian bank customers, IJRITCC, Volume: 1 Issue: 9

[11] R. Khare, D. Kaloya, C. K. Choudhary, and G. Gupta, “Employee attrition risk assessment using logistic regression analysis,”.

[12] Praveen et al., Churn Prediction in Telecom Industry Using R, in (IJETR) ISSN: 2321-0869, Volume-3, Issue-5, May 2015

[13] J. Burez and D. Van den Poel, “Handling class imbalance in customer churn prediction,” *Expert Systems with Applications*, vol. 36, no. 3, 2009.

[14] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, “New insights into churn prediction in the telecommunication sector: A profit driven data mining approach,” *European Journal of Operational Research*, vol. 218, no. 1, pp. 211–229, 2012.