I J C T A, 9(34) 2016, pp. 603-621

Enhancing the Performance of the Classifiers for Customer Churn Analysis in Telecommunication Data using EMOTE

[1]S. Babu and [2]Dr. N. R. Ananthanarayanan

ABSTRACT

Customer Churn is the term refers to the customers who are in threat to leave the company. Growing number of such customers are becoming critical for the telecommunication sector and the telecom sector are also in a situation to retain them to avoid the revenue loss. Prediction of such behaviour is very essential for the telecom sector and Classifiers proved to be the effective one for the same. A well balanced dataset is a vital resource for the classifiers to yield the best prediction. All existing classifiers tend to perform poor on imbalanced dataset. An imbalanced dataset is the one, where the classification attribute is not evenly distributed. The cause of poor performance on such dataset is that, the classifiers look for overall accuracy not by taking into account of the relative distribution of each class. Like the other real time applications, the telecommunication churn application also has the class imbalance problem. So it is extremely vital to go for fine balanced dataset for classification.

In this paper, an empirical method **EC_for TELECAM (E**nhanced **C**lassifier **for TELE**communication **C**hurn ***A***nalysis **M**odel**)** using **EMOTE (E**nhanced **M**inority **O**versampling **TE**chnique**)** has been proposed to improve the performance of the classifier for customer churn analysis in telecom dataset. The key idea of the proposed model is that, fine-tuning the misclassified instances into correctly classified instances using their nearest neighbour. To evaluate the performance of the proposed method, Different UCI repository datasets are used with different ratios of imbalance. The experimental result shows that, the proposed method effectively improves the performance of the classifier, through which it extracts the best decision rule for the prediction. In order to perform the Churn analysis, the primary data with 235 samples was collected through structured questionnaire. To extract the decision rule for the churn predicting system, the proposed method was executed on the collected data. In order to prove that, the extracted rules are statistically significant, Discriminant Factor Analysis using SPSS is also carried out. The evaluation results show that, the extracted rules to predict the customer churn are most significant with the related attributes. As an overall, the experimental results show that, the proposed method outperformed and extremely improve the accuracy of the classifier by which it able to achieve the best prediction over the Customer Churn.

Keywords: Customer Churn, Classification, Imbalanced Dataset, Nearest Neighbour, Oversampling.

1. INTRODUCTION

1.1. Telecom and Churn

The telecom services are accepted world–over as an important source of socio-economic growth for a nation. Particularly the Indian telecom has attained a phenomenal growth during the last few years and is expected to take a positive growth in the future also. This rapid growth is possible due to the different positive and proactive decisions of the government and the contribution of both by private sector and the public. Due to the increasing competition, telecom sectors are facing the issue of customer churn. Customer Churn is the term refers to the customers who are in threat to leave the company. Churn is a very critical issue in telecom, because of its association with loss of revenue and the high cost of attracting the new customer.

Dept of Computer Science and Applications, SCSVMV University, *E-mails: [1]babulingaa@gmail.com, [2]nrananthanarayanan@gmail.com*

1.2. Classifier and Imbalanced Dataset

Classification is widely recognized as a significant Data Mining technique for mining the data and predict about the future. By building the pertinent classifier, it is able to predict well about which class the new instance is [9].

In general Classifiers presume that, the dataset instances are uniformly distributed among different classes. The Classifier is able to perform best on the dataset whose distribution among the class is even, but poor on the imbalanced dataset. On the contrary the real world datasets are imbalanced among the distribution of the class attribute. The issue of class imbalance arise when many more instances in one class (Majority Class) and very less in other class (Minority Class) in the training dataset. It is very severe in the context of "Big Data" processing. The first reason is due to fact that the required valuable information is usually represented by a minority class.

The performance of the classifier built based on such imbalanced dataset was extremely good on majority class but very poor on minority class [5]. Conversely, in many of the real cases, the most essential one for the prediction are minority class instances. For example, In Churn dataset 2416 are the total number of instances, in which false class (Not Churned) instances are 2344(97%) and true class (Target Class) instances are 72 (3%). In such scenario, the classifier is able to perform best on classifying the false class (Accuracy is 99.6%), but not well on target class true (Accuracy is 30.3%). Whereas, if the classifier is able to achieve the best prediction on true class (Minority class), it will be most valuable for the telecom industry to identify the customer who will churn and may take suitable action to retain them.

1.3. K-Nearest Neighbour

The classifier is able to build the classification model for the training instances and the same can be used to predict the class labels of unknown samples. In the classification model there are two types of learners are there. They are eager learner and lazy learner. Eager learner develops model from training dataset before receiving the testing dataset. Whereas the lazy learner waits for test set to develop the model. K- Nearest neighbour (KNN) is one of the lazy learners for the purpose of classification. In KNN, prediction of a class label is based on its nearest neighbours. The Figure-1 shown below explains the same. In Figure 1 the data point shown in the middle of the circle along with their 1-, 2-, 3- neighbours are depicted. In Figure 1(a), the data point is assigned with negative class label because nearest neighbour is negative. In Figure 1(b) there is a tie among the nearest neighbour, so the random class label is assigned to data point. In Figure 1(c), out of three nearest neighbour two is positive and one is negative, so in such a case the class label of the data point is assigned with the majority one[7].

Figure 1: Target class with their 1, 2 and 3 Nearest Neighbours

1.4. Weka and IKVM

Weka [20] is a set of Machine learning algorithms for the purpose of data mining task. It also has the visualization tool for the analysis of data together with GUI to perform effortless access on these utilities. Weka performs various functions like data pre-processing, association, regression, classification and

clustering. It was developed on java and the functions presented in weka can be called directly or by means of java code.

IKVM [21] is a tool implemented to run java code on .NET Language. Jeroen Frijters, the Technical Director of Sumatra Software, based in The Netherlands is the main contributor to IKVM.NET. It permits to call all the classes of java using .NET Code. It has following components.

i) A JVM implemented in .NET

ii) A .NET implementation of class libraries of java.

iii) A compiler that converts JAR File(java code) to DLL(.NET IL)

iv) A tool that facilitate interoperability between java and .NET.

By using IKVM, the conversion from Weka jar file to Weka DLL can be done in a single compilation. There after all Classes of Weka are executed through .NET applications.

2. PERFROMANCE MEASURE

Confusion Matrix is the one, through which performance of the machine learning algorithms for binary class problem is measured [10].

	Predicted Positive	**Predicted Negative**
Actual Positive	TP	FN
Actual Negative	FP	TN

Figure 2: Confusion Matrix

In the confusion matrix, TP (True Positive) is the number of positive examples correctly classified as True, TN (True Negative) is the number of negative examples correctly classified as Negative, FN (False Negative) is the number of positive examples incorrectly classified as negative and FP (False Positive) is the number of negative examples incorrectly classified as positive.

Generally accuracy is the basic performance measure of the machine learning algorithms and is defined as

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

The accuracy is sensible in the perspective of balanced datasets. Whereas in the occurrence of imbalanced datasets, it is best to use the other performance measures like Sensitivity (True Positive Rate), Specificity (False Positive Rate), F_Measure, G_Mean, ROC Curve and other similar measures [12]. These measures are defined as

True positive rate (Sensitivity or Recall):

$$\text{TP} = \frac{TP}{TP + FN} \tag{2}$$

True Negative rate (Specificity):

$$SP = \frac{TN}{TN + FP} \tag{3}$$

F-Measure (A harmonic mean of precision and recall):

$$F1 = 2\ X \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

G-Mean (A Geometric mean of precision and recall):

$$G\text{-mean} = \sqrt{Precision\ X\ Recall} \tag{5}$$

The Receiver Operating Characteristic (ROC) curves are the graphical approach for Summarizing and displaying the performance trade off between true positive and false positive error rates of the classifiers [13]. In the ROC Curve the Y-axis represents the Sensitivity (True Positive Rate) and X – axis represents the Specificity (False Positive Rate). The point (0, 1) in the ROC Curve would be the ideal point, (i.e.) it represents that all positive instances are correctly classified as positive and no negative classes are incorrectly classified as positive. In an ROC Curve, the following are the various important points, (0, 0) - States all as a negative class, (1, 1) -States all as a positive class, (0, 1) -Ideal.

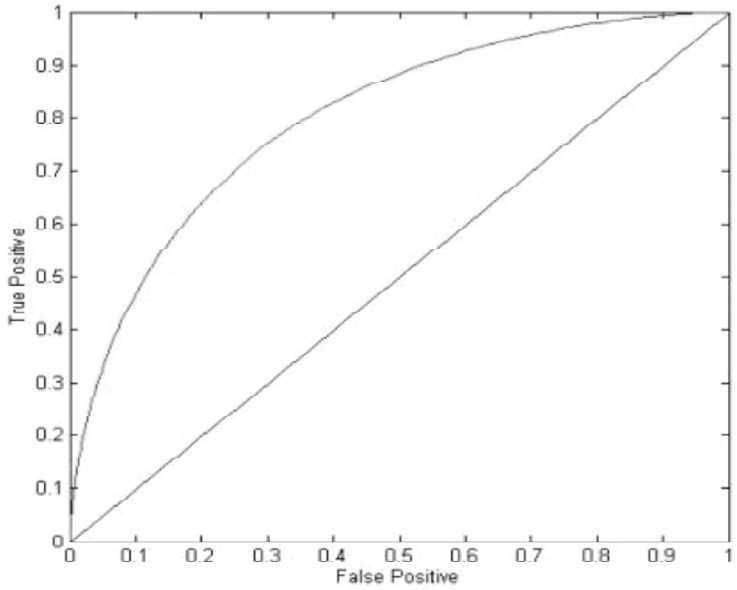

Figure 3: A ROC Curve of a Classifier

The AUC stands for Area Under the Curve is an conventional method to calculate the area under the ROC curve [13]. Accuracy of the classifier is defined by means of the area under the ROC curve. The value of the AUC lies between 0 and 1.0, because it is a segment of the unit square. An area of 1 represents a perfect classifier; an area of .5 represents a less perfect classifier.

3. RELATED WORK

As the churn is a critical issue in telecom, recent researchers shown more interest on churn analysis in telecom and proposed several methods to predict the same. The proposed methods try to predict churn on telecom dataset not by considering whether dataset is well balanced or not. The classifiers may lead to perform poor on the imbalanced dataset.

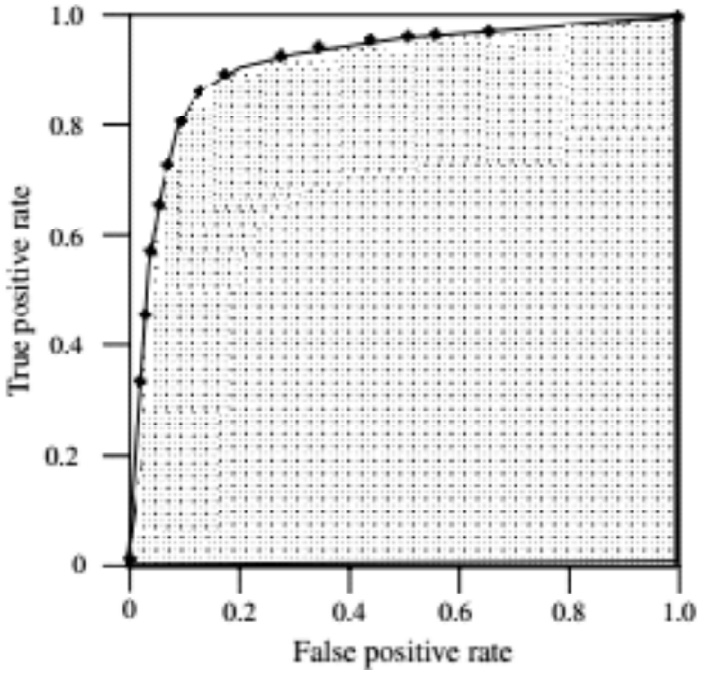

Figure 4: Area under the ROC Curve

To obtain the best prediction system, the imbalance problem needs to be resolved.

Ionut BRANDUSOIU, Gavril TODEREAN are proposed a churn prediction model using support vector machine learning algorithm with four kernel functions namely RBF, LIN, PO and SIG [6]. The churn dataset with 21 attributes of 3333 instances are considered for analysis. Evaluation of result has been done in technical and practical point of view. From the analysis it is proved that, the prediction model that employ with polynomial kernel function performs best. But in practical point of view, the other three models like LIN, POL and RBF performs best in the prediction of churn.

Yi-Fan Wang a, Ding-An Chiang b, Mei-Hua Hsu c, Cheng-Jung Lin b , I-Long Lin d designed a recommender system for wireless network companies to know and to shirk customer churn in telecom industry[8]. Data mining decision tree algorithm used to analyze the dataset taken above the period of three months. The dataset is partitioned into two, the first nine weeks of data is considered as training dataset and rest is considered as testing dataset. Through the proposed work, it has been concluded that, obtained results of the decision tree algorithms are applicable and highly valuable. In addition to this, it is also able to arrive at a new marketing and promotion strategies to control the churn.

Kiran Dahiya, Surbhi Bhatia has introduced a new framework consists of five modules namely Data Acquisition, Data Preparation, Data Pre-processing Data Extraction and Decision to predict the customer churn [14]. Three datasets small (50 instances with 10 attributes), medium (200 instances with 50 attributes) and large (608 instances with 100 attributes) are considered for analysis. Decision tree and Logistic Regression are the mining techniques used to build the customer churn prediction model. From the analysis it is proved that, the performance of the decision tree is far better than the logistic regression. In addition to this, it is also concluded that, fine methods has to be defined and existing methods has to be fine tuned to identity and prevent churn.

P.S. Rajeswari and P. Ravilochanan are initiated a study to observe about the level of customer satisfaction and to analyze the factors that influences the customer churn in prepaid part of Indian telecommunication industry [3]. The primary data about 1102 of Tamil Nadu state of India are collected through questionnaires. The statistical tools like exploratory factor analysis and multiple regressions are considered for data analysis. From the analysis it is concluded that operational factors like, services based

on technology, Network coverage, Internet Speed, complaint resolution system are the forceful elements for the customer churn.

Nitesh V. Chawla has proposed the technique called SMOTE (Synthetic Minority Oversampling Technique) [10]. In the proposed work oversampling was used for minority class and under sampling was used for majority class for balancing the data set. For oversampling, the samples of minority class are populated by creating artificial synthetic examples instead of replicating the real data. Based on required amount of oversampling, five nearest neighbours using k-nn algorithm are taken randomly. From which two are chosen and the new instance was created in the direction of each. To analyze the performance, various datasets and the classifiers like C4.5, Naive Bayes and Ripper are used. The obtained results prove that the proposed method performed well than the other re-sampling techniques.

Maisarah Zorkeflee is presented a new under sampling method to handle imbalance [1]. It is a combination of Fuzzy Distance based Under Sampling (FDUS) and SMOTE. The process divides data set into two classes namely majority (Ai) and minority (Bi) class. Using FDUS method Re-sampling data set is repeated to produce the balanced dataset. During the process, SMOTE is used to balance the dataset, if Ai becomes lesser than Bi. F-measure and G-mean are the measures used to analyze the performance of FDUS+SMOTE using the dataset PIMA, HABERMAN and Bupa. The result of the analysis show that proposed method performed best on balancing the dataset than the other techniques.

Piyasak Jeatrakul is introduced a model to improve the accuracy of the minority class by combining both the SMOTE and the Complementary Neural Network (CMTNN)[4]. In the proposed method for under sampling, CMTNN is applied and for Oversampling, SMOTE is applied. By combining these two methods four techniques are developed to deal with imbalance problem. German, Pima, Spect and Haberman's are the dataset considered for evaluation. To evaluate the performance of the proposed balanced dataset, classifiers ANN, SVM and K-NN are executed on the same. The comparison result of the measures like G-Mean and AUC, suggests that proposed method performs well.

Bao-Gang Hu is presented a work, in which investigation of the cost behaviour related to the classification measures on the imbalanced datasets was done [2]. For study twelve measures are considered, like F-measure, G-mean, precision, recall, Balance Error Rate (BER), Matthews Correlation Coefficient etc.,. A new observation is presented for the above measures by exposing their cost functions in association with the class imbalance ratio. Based on the cost functions, it is able identify, the reason why some measures are suitable to deal with the class imbalance problem. In addition to this, it is also concluded that, G-mean and F-Measure are the suitable measures to prove the performance of the classifiers in the background of imbalance class, because they confirm the suitable cost behaviours in terms of "*a cost misclassification from a small class is greater than from a large class*".

4. MATERIALS AND METHODS

In the imbalanced data set classifiers do not give high level of priority in classifying minority class instances. While constructing the tree also, the minority class instances are treated in the lower branches of the tree and such a treatment may lead to the increase of misclassification rate in the minority class instances. Ultimately the Sensitivity becomes very poor and the overall performance too. So classifiers need to give more importance for the minority class instances also. To do so the misclassified instances are replicated along with their nearest neighbour. With the above brief description, the proposed work is designed in two phases. Which are as follows:

Phase I: Defining the model to Enhance the Performance of the Classifier and Analyzing the same with various datasets.

Phase II: Extracting the Decision Rule for churn prediction from the proposed model using the Primary data collected through questionnaire.

4.1. Materials

4.1.1. Data Description

Different datasets namely German, Yeast, Adult, Pima Indian, Phoneme, Thoracic Surgery, churn from the UCI Machine Leaning repository and Oil Spill data given by Robert Holte [15], are used in the study of Phase I. The datasets mentioned have different ratio of imbalance between the true and false class.

As part of the work is pertained with descriptive research, the primary data were collected and used in the study of Phase II.

4.1.2. Data Collection

Survey method was adopted using the structured questionnaire for the collection of primary data. The attributes of the questionnaire are defined with 5 point scale. In India, Tamil Nadu was the second largest in mobile subscription, so the same was selected as a sampling framework for this study. Totally 235 samples are collected and considered for processing.

4.1.3. Reliability

Reliability is the scale to which, an assessment tool generates stable and consistent results [3]. The reliability of the questionnaire was tested using the SPSS 18 Package. From the result of the test, it is found that, the Cronbach's alpha is 0.879, which in turn indicates that a high level of internal consistency for the scale with the collected samples.

4.1.4. Pilot Study

Before confirming the questionnaire, the field test with 50 respondents who are working in telecom industry is done to fine tune the questionnaire.

4.1.5. Data Pre-processing

Data was Pre-processed by checking the missing values, outliers, skewness and kurtosis. In addition to this, all the attribute values are converted as continuous values by changing Stongly Agree as 5, Agree as 4, Neutral as 3, Disagree as 2 and Strongly Disagree as 1.

4.2. Methods

4.2.1. Phase I

In Phase I, the Enhanced Classification model has been proposed with the main motivation to improve the performance of the classifier. The core idea behind the proposed model (**EC_for_TELECAM**) is that, fine-tuning the misclassified instances into correctly classified instances using their nearest neighbour. The flow starts first by improving the sensitivity (True positive rate) by calling the procedure EMOTE through which, decision rule for true class is extracted. Then By calling EMOTE the specificity (False positive part) of the dataset is also improved, by which decision rule for false class is extracted. Then to improve the overall performance of the classifier, the new classification rule is framed using the extracted decision rules of true and false classes.

ALGORITHM 1: EC_for _TELECAM (Ads)

// **E**nhanced **C**lassifier_ **for** _ **TELE**communication **C**hurn **A**nalysis **M**odel
Input: Actual dataset - **(Ads)**
Output: Dt - > Decision Tree
Dr -> Decision Rule

Step 1: Start.
Step 2: Get the Actual Dataset **Ads.**

// **To improve the Sensitivity of the Dataset**
Step 3: **T_Dr** = Call **EMOTE (Ads, 5, "True")**

// **Decision rule for True class**
Step 5: **TR** = Extract the Decision rule for "**True"** class from **T_Dr**.

// **To improve the Specificity of the Dataset**
Step 6: **F_Dr** = Call **EMOTE (Ads, 5, "False")**
// **Decision rule for False class**
Step 7: **FR** = Extract the Decision rule for **"False"** Class form **F_Dr**.
Step 8: Construct the Predicting rule from **TR & FR**
Step 9: Classify the Actual dataset **(Ads)** with rule from Step 8.
Step 10: Stop.

4.2.2. Phase II

The flow of the procedure **EMOTE** begins by obtaining the imbalanced data set **(Ai)** as input data set. In addition to this, the procedure also takes two additional parameters, namely Number of Nearest Neighbour **(k)** for KNN processing and Class label **(c)** to define about, for which class (True for Sensitivity or False for Specificity) the dataset has to be balanced to improve the performance of the classifier. By setting the class attribute, the classifier is built on the imbalanced data set.

As a primary step, various performance measures like Overall Accuracy, Sensitivity (True positive rate) and Specificity (False positive rate) are computed from the results of the classifier. If the accuracy of the required class label (***C***) is not optimal then, the actual dataset (Ai) is partitioned into two different datasets namely Dataset (CCi) with Correctly Classified Instances and Dataset (Mi) with Misclassified instances.

As the secondary step, the instance from the misclassified instances (Mi) whose class label identical to **c** (True or False, given through parameter) is considered for tuning. To fine tune the selected misclassified instance into Correctly Classified Instance, the nearest neighbours of the misclassified instance from the Correctly Classified Instances are retrieved. Then the selected misclassified instance and its equivalent (based on class label) Nearest Neighbours which as identical class label are populated on the Actual Dataset (Ai). The above step is repeated for each Misclassified instances whose class label belongs to **c.**

As a final step with the enhanced data set the classifier is rebuilt and the above said primary and secondary steps are repeated till the optimal accuracy attained on the class label denoted by **c.**

4. RESULTS AND DISCUSSION

4.1. Phase-I

To evaluate the performance of the proposed work, the C# application has been developed. WEKA is used for data mining process. To utilize the classes of WEKA in the C# code, the source file WEKA.jar is transformed to WEKA.dll using IKVM. The application developed in C# for the proposed method utilizes the classifier C4.5 (J48 in WEKA) to build the classification model. Different datasets are used to evaluate the proposed model **EC_for_TELECAM** The description of all the datasets are shown in Table 1.

ALGORITHM 2: EMOTE(Ai, k, c)

// **E**nhanced **M**inority **O**versampling **TE**chnique

Input: Actual dataset - **(Ai)**, No of Nearest Neighbour - **k,** Class Value - **c**.

Output: Dt - > Decision Tree
Dr -> Decision Rule

Step 1: Start
Step 2: Set the class attribute for dataset **Ai**
Step 3: Set the classifier and build the classifier
Step 4: **do**
Step 5: Classify the dataset **Ai** and Calculate
i) TP – True Positive Rate
ii) FN – False Negative Rate
iii) FP - False Positive Rate
iv) TN – True Negative Rate
Step 6: Compute Accuracy = (TP + TN) / (TP + FN + FP + TN)
Step 7: Compute Sensitivity = TP / (TP + FN)
Step 8: Compute Specificity = TN / (TN + FP)
Step 9: Partition the Actual Dataset **Ai** into Misclassified instances as **Mi** and Correctly Classified instances **CCi**
Step 10: **for i = 0 to count (Mi)-1**
Step 11: if **Mi (i)** = **c** then // **c – Class Value**
Step 12: populate (**Mi (i)** -> **Ai)**
Step 13: find the nearest neighbour **Nn** of the (**Mi (i)**, k) from **CCi** //**k – no of neighbour**
Step 14: **for j = 0 to count (Nn)-1**
Step 15: if **Nn(j)** = **c** then // **c – Class Value**
Step 16: **populate (Nn(j) -> Ai)**
Step 17: end if
Step 18: **Next j**
Step 19: end if
Step 20: **Next i**
Step 21: rebuild the classifier
Step 22: **repeat Steps 4 to 21 until optimum level of accuracy.**
Step 23: return (Dr) // **Returns Decision Rule**
Step 24: Stop.

To test the performance of the classifier, ten cross fold method is adopted to split the dataset into training dataset (90%) and testing dataset (10%). The model has been built on the training dataset and the same was tested on the testing dataset.

Table 1
Description of Datasets used in Experiments

Dataset	*No. of Attributes*	*No. of Instances*	*True Class %*	*False Class %*
German Credit	20	1000	30	70
Churn Telecom	21	2416	3	97
Yeast	8	1485	29	71
Adult	15	27561	28	90
Oil Spil	51	937	4	96
Thoracic Surgery	17	471	15	85
Phoneme	6	5404	29	71

To evaluate the performance of the proposed method different types of test has been carried out. They are,

1. Performance evaluation of the proposed model using C4.5 on various datasets.
2. Performance evaluation of the proposed model using various classifiers on Thoracic Surgery dataset.

3. AUC analysis of the proposed model using various classifiers on Thoracic Surgery dataset.
4. Performance Comparison of proposed model with FDUS+SMOTE and CMTNN+SMOTE
5. Statistical evaluation of the proposed model.

4.1.1. Evaluation of Proposed Model using C4.5

To evaluate the proposed method, the classifier C4.5 was executed on the imbalanced actual dataset using WEKA. From the results of the classifier various performance measures are calculated. To balance the actual dataset, the proposed model was executed on the actual dataset. After balancing, the classifier is executed yet again on the balanced dataset. The calculated values of both executions on various datasets are recorded and presented in Table 2.

The obtained experimental result shown in Table 2 reveals that, the performance of the classifier C4.5 on the imbalanced dataset is not fine on the prediction of True (Sensitivity) class and fine on the prediction of False (Specificity) class. Whereas on the proposed balanced dataset, the performance of the classifier on both the classes are best. Particularly on Thoracic Surgery dataset, the proposed model extremely improves the performance of the classifier on part of Sensitivity. The pictorial representation of comparison of all the measures, on actual and proposed was shown in Figure (5). In figure the Y-axis represents the accuracy of the classifier and the X-axis represents different datasets.

Table 2
Performance Comparison of proposed method on various Datasets using the classifier C4.5

Dataset	Sensitivity		Specificity		Overall Accuracy		G-Mean		F-Measure		AUC	
	Actual	Proposed	Actual	Proposed	Actual	Proposed	Actual	Proposed	Actual	Proposed	Actual	Proposed
German Credit	44.67	***97.67***	85.14	***99.43***	73	***98.9***	61.67	***98.54***	49.81	***98.16***	0.6716	***0.9763***
Churn Telecom	30.3	***96.97***	99.62	***99.91***	97.72	***99.83***	54.94	***98.43***	42.11	***96.97***	0.8461	***0.9988***
Yeast	47.44	***95.81***	86.07	***97.91***	74.88	***97.31***	63.9	***96.86***	52.24	***95.37***	0.8571	***0.9763***
Adult	63.56	***99.23***	93.32	***99.73***	86.15	***99.61***	77.01	***99.48***	68.85	***99.2***	0.8872	***0.9773***
Oil Spil	31.71	***95.12***	97.43	***99.78***	94.56	***99.57***	55.58	***97.42***	33.77	***95.12***	0.5146	***0.9999***
Thoracic Surgery	1.41	***91.55***	99	***100***	84.29	***98.73***	11.81	***95.68***	2.63	***95.59***	0.4878	***0.9931***
Phoneme	79	***99.24***	90.36	***99.53***	87.03	***99.44***	84.49	***99.39***	78.14	***99.06***	0.8992	***0.9858***

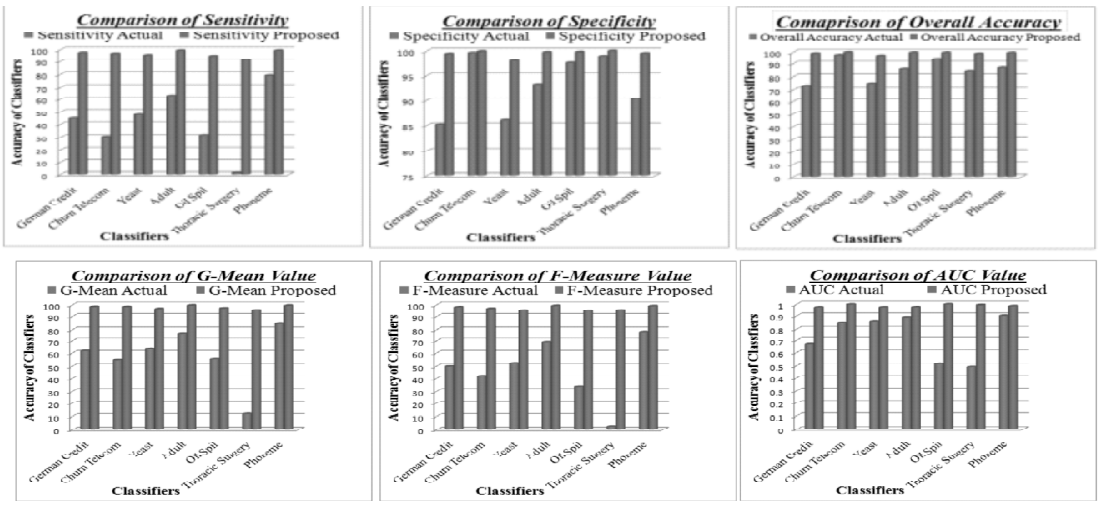

Figure 5: Performance Comparison of proposed method on various Datasets using the classifier C4.5

4.1.2. Evaluation of Proposed Model using Various Classifiers

In contrast to the comparison mentioned above, various classifiers like NB Tree, Random Forest, Simple Cart, KNN and MLP (Multi Layer Perceptron) are executed on the Actual Thoracic Surgery dataset. The results of the experiments are recorded and found that not fair. To improve the performance of the classifier, the proposed model was executed on the actual dataset to balance the same. After balancing, the classifiers stated above are again executed on the balanced dataset. The results of the executions are presented in the Table 3.

Table 3
Performance Comparison of proposed method on various classifiers using the dataset Thoracic Surgery

Dataset	Sensitivity		Specificity		Overall Accuracy		G-Mean		F-Measure		AUC	
	Actual	Proposed	Actual	Proposed	Actual	Proposed	Actual	Proposed	Actual	Proposed	Actual	Proposed
NB Tree	4.23	***90.29***	96.25	***99.75***	82.38	***96.53***	20.17	***94.9***	6.74	***94.66***	0.5773	***0.9682***
Random Forest	9.86	***100***	93.75	***100***	81.1	***100***	30.4	***100***	13.59	***100***	0.6536	***1***
Simple Cart	2.82	***99.03***	92.5	***98.75***	78.98	***98.84***	16.14	***98.89***	3.88	***98.31***	0.4937	***0.988***
KNN	5.63	***97.09***	94.5	***96.5***	81.1	***96.7***	23.07	***96.79***	8.25	***95.24***	0.561	***0.9423***
MLP	16.9	***98.06***	89.5	***99.25***	78.56	***98.84***	38.89	***98.65***	19.2	***98.3***	0.5901	***0.974***

The results of the experiments confirm that, all the classifiers are able to prove their efficiency only on the majority classes but fail to prove in minority classes on actual dataset. The results also reveal that, the proposed model is an efficient one to solve such an issue. In specific the performance of simple cart is very poor when compare to other classifiers on minority class. Still simple cart proves that, the proposed model balances the data set in a best manner to enhance the performance of the classifier.

4.1.3. AUC analysis of the Proposed Model

To further highlight the performance of the proposed model, AUC analysis is also done.

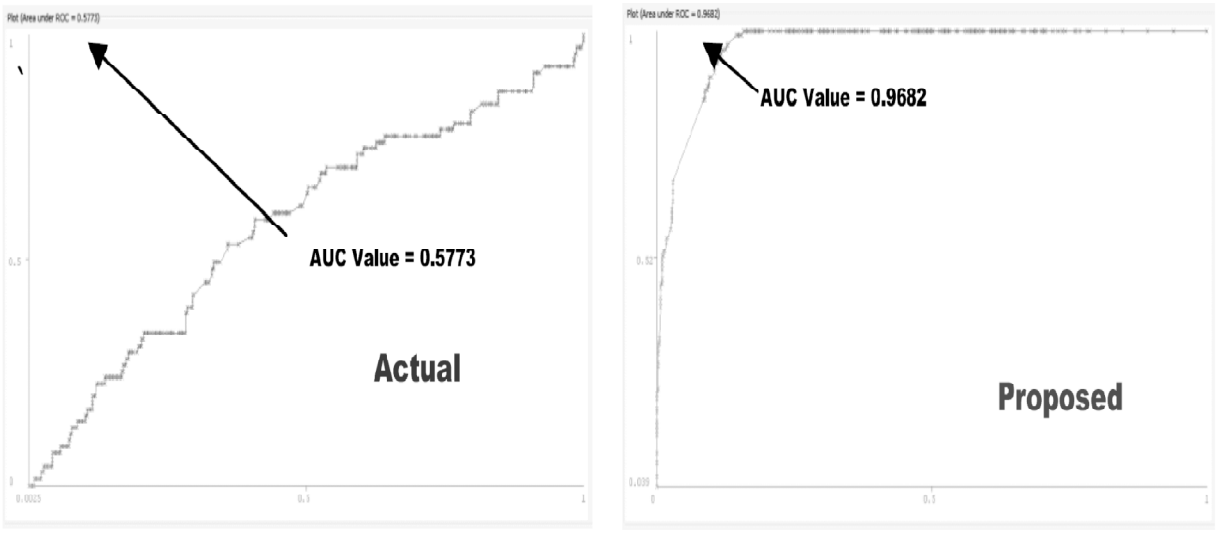

Figure 6: ROC Comparison of Proposed Method with NBTREE as base classifier on Thoracic Surgery Dataset

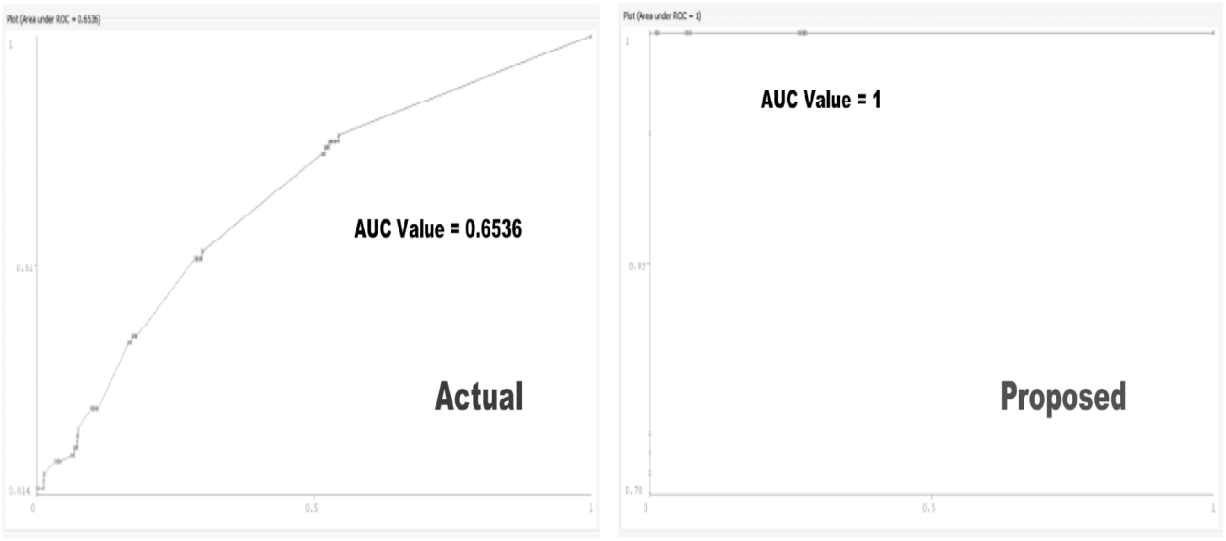

Figure 7: ROC Comparison of Proposed Method with Random Forest as classifier on Thoracic Surgery Dataset

The ROC curves depicted in Figure 6 to 10 are defined using various classifiers like NB Tree, Random Forest, Simple Cart, KNN and MLP (Multi Layer Perceptron) as base classifier. Initially Curves created by executing classifiers on the actual dataset and next on the dataset which is balanced by the proposed model. In ROC curve the steeper the curve (towards the upper left corner) states that, the better the classification. The steep of the ROC curves on the proposed dataset is better than the actual dataset. From the ROC curves, the AUC values also calculated. It also proves that, the proposed model significantly improves the performance of the classifier.

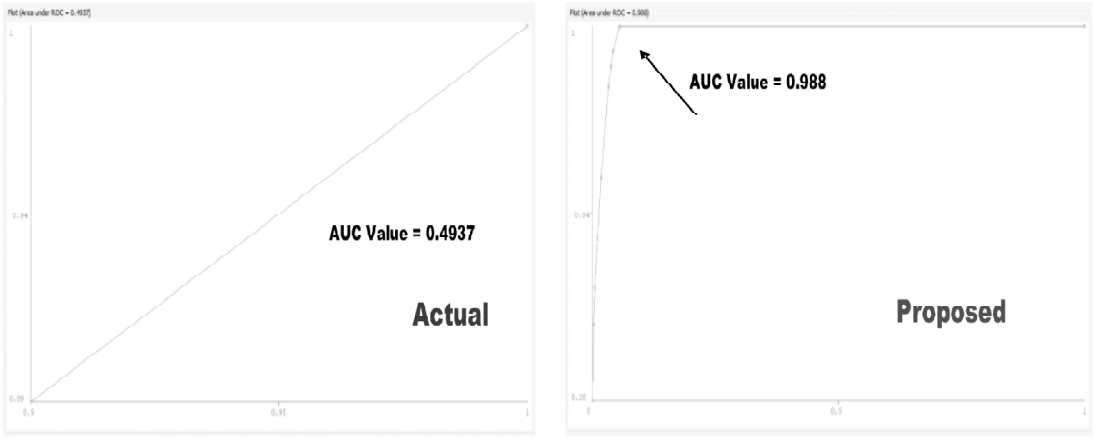

Figure 8: ROC Comparison of Proposed Method with Simple Cart as classifier on Thoracic Surgery Dataset

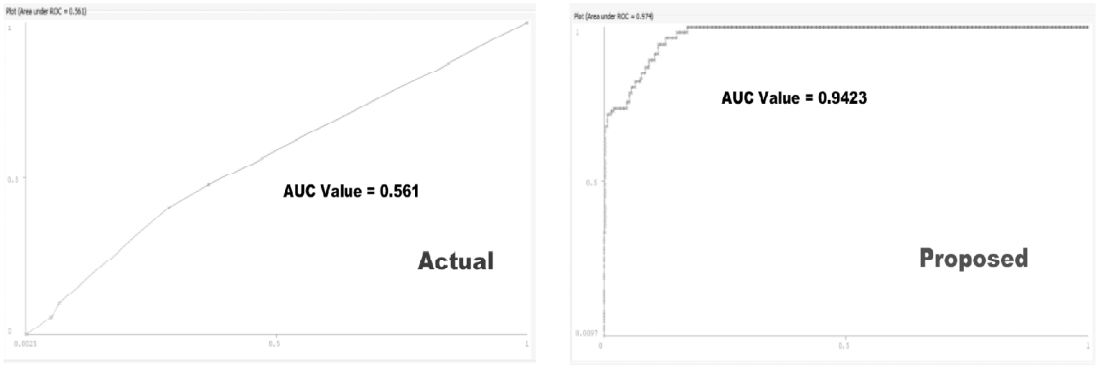

Figure 9: ROC Comparison of Proposed Method with KNN **as base classifier on Thoracic Surgery Dataset**

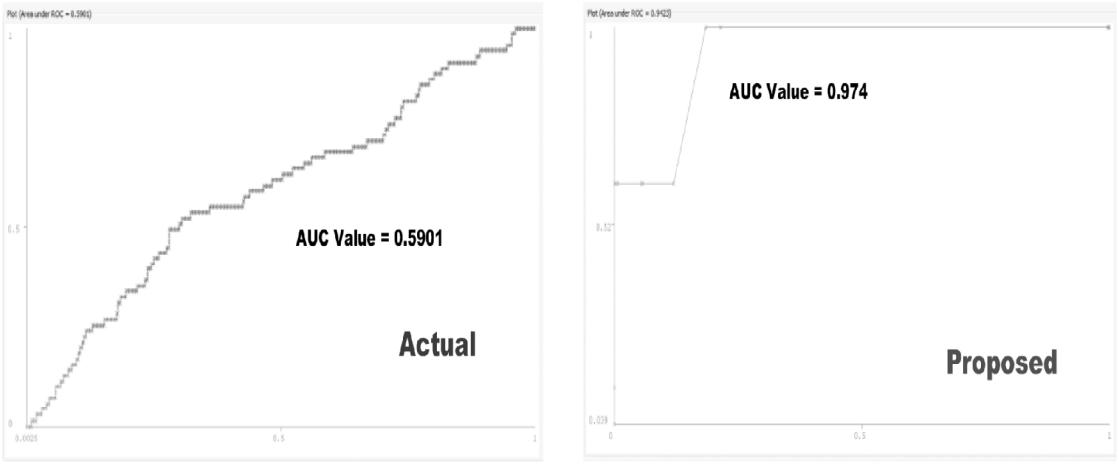

Figure 10: ROC Comparison of Proposed Method with MLP **as base classifier on Thoracic Surgery Dataset**

4.1.4. Comparative Evaluation of the Proposed Model

FDUS+SMOTE [1], which is an integration of Fuzzy distances based under sampling and SMOTE. The efficiency of FDUS+SMOTE is proved by calculating F-measure (relation between precision and recall) and G-Mean (classifier accuracy on both classes) on BUPA, HABERMAN and PIMA Indian datasets and the same is compared with other methods like, SMOTE+ENN [16] and SMOTE+TOMEK [17].

Steep of Curve

To evaluate proposed model with FDUS+SMOTE, the classifier is executed on the same dataset and the measures like G-Mean and F-Measure are calculated and the same is presented in the Table 4 with results of actual and FDUS+SMOTE from [1]. The comparison results proves that, the proposed model performs better than FDUS+SMOTE. In addition to this, through G-Mean it reveals that the classifier accuracy on both the classes is high than FDUS+SMOTE. The F-Measure proves that the proportion of correctly classified instances of both the classes are with greater percentage of F-Measure than FDUS+SMOTE.

Table 4
G Mean and F Measure Comparison of proposed method with other methods

Techniques	Pima Indian Data		Bupa Data		Haberman's Survival Data	
	GM	FM	GM	FM	GM	FM
Original Data	69.14	61.1	81.75	79.83	62.32	50
Proposed	**97.13**	**96.3**	**93.95**	**96.66**	**96.4**	**96**
FDUS+SMOTE	65.44	81.59	80.7	76.69	85.11	85.71

CMTNN+SMOTE [4] is a combined techniques of both Complementary Neural Network (CMTNN) and SMOTE to handle imbalance problem. The performance of CMTNN+SMOTE is proved using the classifier k-NN (k=5) on Pima Indians, Haberman's Survival, German credit and SPECT heart data sets. G-Mean and AUC are the measures calculated and the same is compared with results of other methods like ENN, SMOTE and TOMEK LINKS. To compare the proposed model with CMTNN+SMOTE, the classifier k-NN (k=5) is executed on same dataset and the measures G-mean and AUC are calculated from the output of classifier. The results of proposed model are presented in Table 5 with results of actual and CMTNN+SMOTE from [4]. The results show that the proposed method performed better than CMTNN+SMOTE.

Techniques	Pima Indian Data		German Credit Data		Haberman's Survival Data		SPECT Heart data	
	GM	AUC	GM	AUC	GM	AUC	GM	AUC
Original Data	65.27	0.7665	59.35	0.7483	40.11	0.5741	68	0.8121
Proposed	**94.05**	**0.9253**	**92.48**	**0.9763**	**88.45**	**0.8658**	**93**	**0.8349**
CMTNN+SMOTE	73.95	0.8104	72.35	0.7785	59.3	0.6302	75.3	0.8264

Table 5: G Mean and AUC Comparison of proposed method with other methods

4.1.5. Statistical Evaluation of the Proposed Model

Various evaluations are carried out to prove the efficiency of the proposed model. In all, the proposed model obtained an outstanding improvement over the classifier accuracy. Here the question is, whether this difference is statistically significant or not. To deal with the performance comparison of classifiers, many methods has been described. But the most recommended method is Wilcoxon Signed Rank test for the performance comparison of classifiers [18]. It is a non parametric hypothesis test to compare two related samples and useful to evaluate about the population mean rank of the samples differs or not. The following two tests are done using Wilcoxon Signed Rank test, to check whether the performance difference of the classifier on proposed model is significant or not.

Test: 1

Ho: There is no significant difference in the performance of the classifier on various, actual and proposed datasets.

Input

Reference – Table 2

Output

Table 6
Results obtained from Wilcoxon Signed Rank test for Table -2

	Proposed Sensitivity - Actual Sensitivity	*Proposed Specificity - Actual Specificity*	*Proposed Overall Accuracy - Actual Overall Accuracy*	*Proposed G mean - Actual G mean*	*Proposed F Measure - Actual F Measure*	*Proposed AUC - Actual AUC*
Z	-2.388	-2.388	-2.388	-2.388	-2.388	-2.388
Asymp. Sig. (2-tailed)	0.018	0.018	0.018	0.018	0.018	0.018

Test: 2

Ho: There is no significant difference in the performance of various classifiers on actual and proposed datasets **(Thoracic Surgery).**

Input

Reference – Table 3

Output

Table 7
Results obtained from Wilcoxon Signed Rank test for Table -2

	NB_Tree Proposed - NB_Tree Actual	*Random Forest Proposed - Random Forest Actual*	*Simple Cart Proposed - Simple Cart Actual*	*KNN Proposed - KNN Actual*	*MLP Proposed - MLP Actual*
Z	-2.201	-2.201	-2.201	-2.201	-2.201
Asymp.Sig. (2-tailed)	0.028	0.028	0.028	0.028	0.028

The test statistic results are z=-2.366 and p=0.018 for the first case and z =-2.201 and p=0.028 for the second case. From the results the null hypothesis Ho is rejected as the p value is less than significance value 0.05. Hence it is proved that, in both the cases the differences in the performance of the classifiers are statistically significant at the 0.05 significance level.

4.2. Phase II

To perform churn analysis, the primary dataset collected through questionnaire (235 Instance: 89 True class instances and 146 False class instances) are considered. Approximately, the dataset has the imbalance ration about 1:2. In order to extract the prediction rule and to perform the churn analysis, the proposed method **EC_for_TELECAM** was executed on the dataset. The proposed method improves both sensitivity and specificity by balancing the dataset. The method primarily improves true positive rate (**Sensitivity**) of the dataset and extracts rule for true class. As a secondary step false positive rate (**Specificity**) of data set was improved and the rule for false class is extracted. The prediction rule for the churn predicting system was framed using the rules extracted from primary and secondary step.

The decision rules are framed by constructing the decision tree until the suitable classification is reached. The two mathematical concept involved in constructing the tree are entropy and information gain. Where entropy is overall level of uncertainty and information gain is decrease in entropy.

$$\text{Entropy } E(S) = -\Sigma i \; Pi \; log2Pi \tag{6}$$

$$\text{Gain } (S, X) = \text{Entropy}(S) - \text{Entropy}(S, X) \tag{7}$$

The steps to calculate the Entropy and Information Gain for a continuous attributes are described below. For example to calculate gain for attribute Billing, as a first step Sorting of an attribute values in ascending order is done and as a next step Duplicate values are removed. Then the calculation of gain, using the formula mentioned above is presented in the Table 8.

Table 8
Gain calculation of continuous attribute using C4.5 as base classifier

Attribute Values (Unique)	*5*		*4*		*3*		*2*		*1*	
Interval	<=	>	<=	>	<=	>	<=	>	<=	>
TRUE	188	0	73	115	20	168	8	180	1	187
FALSE	146	0	76	70	12	134	7	139	0	146
Entropy	0.989	0	1	0.957	0.954	0.991	0.997	0.988	0	0.989
Entropy(S,T)	0.989		0.976		0.987		0.988		0.986	
Gain	0.000		**0.013**		0.001		0.000		0.002	

The above process is repeated for all the attributes of the dataset. By comparing all, it is identified that the gain value of billing attribute is higher than the others. So the same is selected as root node for the decision tree. In order to find the next decision node of the tree, the same step is repeated with dataset having instances which has billing attribute as <= 4 (The interval where high gain attained.). This recursive action will lead to the final decision tree. To generate the decision rule and to make the clear view of the decision tree, a path of the each leaf nodes is converted in to IF-THEN rule. The Figure 11 shows the decision rule for the TRUE class, which is the required one in order to expose the prediction system for churned customer

The generated decision tree and rule shows that, Billing, Offers, Accessibility, Mobile Number Portability (MNP) and Tariff Plan are the most significant factors which influence the customer churn in

telecommunication. In addition to this, it also suggests that, these five factors out of fifteen needs to be focused more by the service providers in order to reduce the churn rate.

To evaluate the defined predicting system, the following test was done. They are

1. Cross Validation of Defined Predicting system with actual dataset.
2. Statistical evaluation of Defined Predicting System.

4.2.1. Cross Validation of Defined Predicting System

To check the consistency of the defined predicting system, the cross validation has been carried out with the actual dataset. The results of the predicting system have obtained predictive accuracy of 97.55% which is high about 18.35% than actual and the error rate is 2.45%. In addition, it also has the true positive rate (Sensitivity) as 94.92% and false positive rate (Specificity) as 98.93%. The same was shown in Figure 12, which is the output produced by the developed C# console application of the proposed method.

4.2.2. Statistical evaluation of Defined Predicting System

The statistical test is the effective one to prove significance of the model over the data. In this view, the statistical test is focused to analyze, whether the prediction rules of the predicting system has significance over the Predicting class variable or not. Discriminant Function analysis is useful to deal with such an analysis. It is a statistical test to predict the dependent variable through the independent variable. It is also useful in finding whether a set of variable is effective in the prediction of the dependent variable or not [19].

The discriminant function looks like the following:

$$fk_m = u_0 + u_1 X_1 k_m + u_2 X_2 k_m + ... + u_p X_p k_m \quad (8)$$

Here:

fk_m - The value on the function for case m in the group k

$X_i k_m$- The value on discriminating variable Xi for case m in group k

u_i -Coefficients which produce the desired characteristics of the function.

```
IF BILLING <= 4 AND OFFERS <= 3 AND ACCESSIBILITY <= 4 AND MNP > 3 AND TARIFF PLAN <= 2 THEN
        Classification = TRUE
Else IF BILLING <= 4 AND OFFERS <= 3 AND ACCESSIBILITY <= 4 AND NETWORK COVERAGE <= 4 THEN
        Classification = TRUE
Else IF BILLING <= 4 AND OFFERS > 3 AND SOCIAL MEDIA APPLICATIONS <= 3 THEN
        Classification = TRUE
Else IF BILLING <= 4 AND OFFERS > 3 AND SOCIAL MEDIA APPLICATIONS >3 AND APPLICATIONS <= 4 AND MNP > 4 THEN
        Classification = TRUE
Else IF BILLING <= 4 AND OFFERS > 3 AND SOCIAL MEDIA APPLICATIONS >3 AND APPLICATIONS > 4 THEN
        Classification = TRUE
Else IF BILLING <= 4 AND OFFERS > 3 AND SOCIAL MEDIA APPLICATIONS >3 AND APPLICATIONS > 4 AND TECHNOLOGY <= 4 AND HANDSET ENABLED SERVICES <= 4 THEN
        Classification = TRUE
Else IF BILLING <= 4 AND OFFERS > 3 AND SOCIAL MEDIA APPLICATIONS >3 AND APPLICATIONS > 4 AND TECHNOLOGY >4 AND BRAND > 4 THEN
    IF QUALITY OF SERVICES > 4 THEN
        Classification = TRUE
    END IF
Else IF BILLING > 4 AND APPLICATIONS <= 4 AND VAS SERVICES <= 3 THEN
        Classification = TRUE
Else IF BILLING > 4 AND APPLICATIONS > 3 THEN
        Classification = TRUE
Else IF BILLING > 4 AND APPLICATIONS > 4 AND MNP <= 4 AND TARIFF PLAN <= 3 THEN
        Classification = TRUE
Else IF BILLING > 4 AND APPLICATIONS > 4 AND MNP > 4 AND TECHNOLOGY <= 4 AND VAS SERVICES <= 4 THEN
        Classification = TRUE
Else IF BILLING > 4 AND APPLICATIONS > 4 AND MNP > 4 AND TECHNOLOGY <= 4 AND VAS SERVICES > 4 AND TARIFF PLAN <= 4 THEN
        Classification = TRUE
Else IF BILLING > 4 AND APPLICATIONS > 4 AND MNP > 4 AND TECHNOLOGY >4 AND CUSTOMER CARE SERVICES <= 4 THEN
        Classification = TRUE
Else IF BILLING > 4 AND APPLICATIONS > 4 AND MNP > 4 AND TECHNOLOGY >4 AND CUSTOMER CARE SERVICES <= 4 THEN
        Classification = TRUE
ELSE
        Classification = FALSE

END IF
```

Figure 11: Decision Rules of the Predicting System Obtained from the proposed method

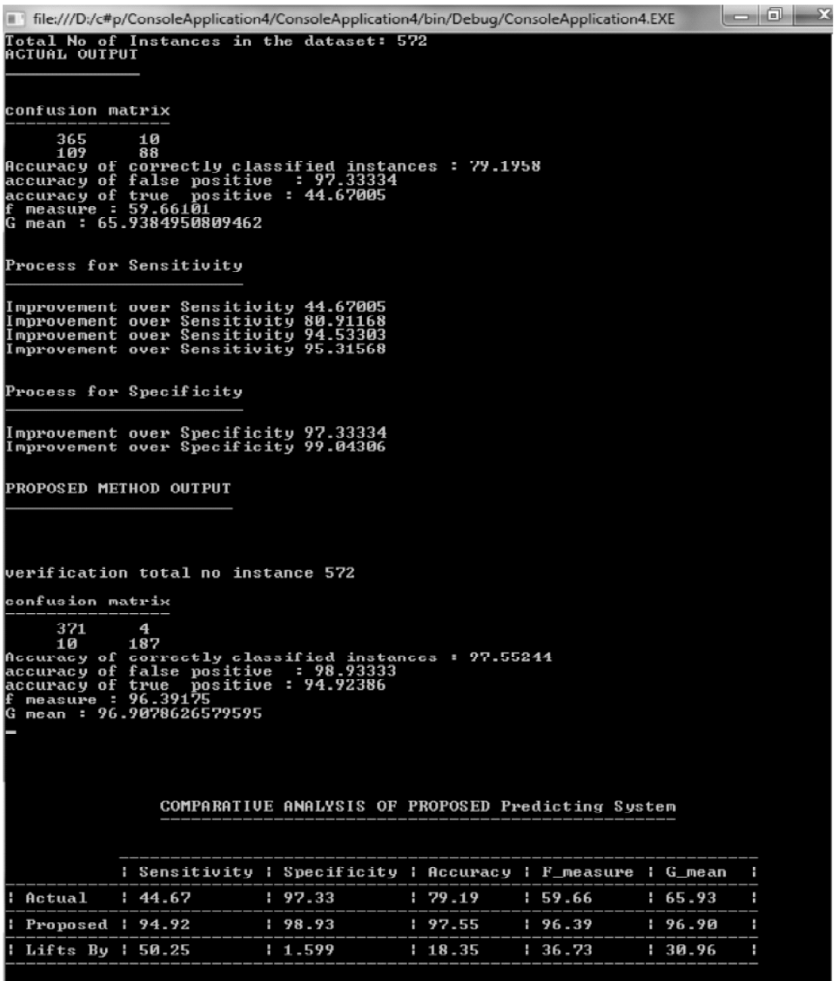

Figure 12 : Output of the C# console application of the proposed method

There are three important variables in the Discriminant Factor Analysis for analyzing the significance of the rule. They are

1. Eigen Value is a value, which is a ratio between explained and unexplained variation in the rule. The rule is good fit to the predictor if the Eigen value is greater than one.
2. Canonical correlation is measure between the dependent variable and rule. A high value represents the high level of association between the dependent variable and rule.
3. Wilks's Lambda is useful to test the significance between dependent variable and rule. A smaller value represents the rule is highly significant with dependent variable.

In order to analyze the significance of the prediction rules, the Null Hypothesis is tested using Discriminant Function analysis.

Ho: The variables involved in the prediction rule have no significance over the dependent variable.

To perform the test, the instances are filtered based on prediction rule and the variables that involved in this rule are tested against the filtered instances. The process is repeated on the randomly selected rules and the results of the test are presented in Table 9.

The values of the Table 9 shows that, the Eigen Values of all the rules except the last two are > 1, the canonical correlation $r_{c>0.35}$ and Wilks Lamda values are all less values (<0.7). The above values indicate that, the rules are good fit to the predictor. In addition to this, p-value<0.005, hence the NULL hypothesis

Table 9
Results obtained from the Discriminant Factor Analysis of Decision Rule (Ref : Figure 11)

	Eigenvalues				Wilks' Lambda			
Function	Eigen value	% of Variance	Cumulative %	Canonical Correlation	Wilks' Lambda	Chi-square	df	Sig.
1	1.810	100.0	100.0	.803	.356	12.914	3	.005
1	1.008	100.0	100.0	.709	.498	27.196	6	.000
1	3.000	100.0	100.0	.866	.250	4.852	1	.028
1	1.585	100.0	100.0	.783	.387	25.645	6	.000
1	.600	100.0	100.0	.612	.625	12.455	3	.006
1	.591	100.0	100.0	.609	.629	18.804	3	.000

Ho is rejected. This proves that, the variables involved in the prediction rule are statistically significant to the dependent variable. As a summary, it is proved that the proposed method is able define best predicting system for customer churn in telecommunication.

CONCLUSION

The issues with imbalanced data set are inherent when used in the process of classification. It impacts the overall performance of the classifier. Many earlier studies focused on various approaches to improve the performance of the classifiers not by considering the imbalance issues. Hence the classifiers are not able to shine on the prediction of minority class instances. Like the other real time applications the telecom churn prediction application also has the imbalance class distribution problem. In such a case, predictions of the customers who need to be identified are very tough.

In this study, an enhanced model **EC_for TELECAM (E**nhanced **C**lassifier for **TELE**communication **C**hurn **A**nalysis **M**odel) is proposed to handle the issue of imbalance in the dataset. To evaluate the proposed method, classifier c4.5 was used on different UCI repository datasets. Various measures like G_mean, F_measure and AUC are calculated and compared with the methods which are widely accepted. The results of the experiments show that, the proposed method well balances the dataset by which it also improves the performance of the classifier. Hence it is concluded that, the role of nearest neighbours of the misclassified instances are more vital in tuning misclassified instances in to correctly classified instances. It is also concluded that, the proposed method is more precious in such a dataset where uniform distribution over the class attributes are not present. In order to define the prediction model for customer churn, the primary data which is collected through questionnaire was used on the proposed method. To test the statistical significance of the rules involved in the prediction model, Discriminant Factor Analysis using SPSS is carried out. The results of the test show that the rules of the prediction model are most significant with the related attributes. As a summary the results of the experiments show that the proposed method EC_for_TELECAM outperformed and defines the best predicting system for customer churn. Through the predicting system it has been concluded that, Billing, Offers, Accessibility, Mobile Number Portability (MNP) and Tariff Plan are the most significant factors which influence the customer churn in telecommunication. Through this work, it is also suggests that, these five factors are more valuable factors which needs to be focused more by the service providers in order to reduce the customer churn rate over the telecom industry.

REFERENCES

[1] Maisarah Zorkeflee, Aniza Mohamed Din, and Ku Ruhana Ku- Mahamud, "Fuzzy And Smote Resampling Technique For Imbalanced Data Sets", Proceedings of the 5th International Conference on Computing and Informatics, ICOCI2015 11-13 August, 2015 Istanbul, Turkey Universiti, Utara, Malaysia.

[2] Bao-Gang Hu, senior member, IEEE "A study on cost behaviours of binary classification measures in class-imbalanced problems", IEEE, 2014.

[3] P.S. Rajeswari and P. Ravilochanan, "An empirical study on customer churn behaviour of Indian prepaid mobile services", Middle-East journal of scienctific research, 2014.

[4] Piyasak Jeatrakul, Kok Wai Wong, and Chun Che Fung, "Classification of Imbalanced Data by Combining the Complementary Neural Network and SMOTE Algorithm", Springer-Verlag, 2010.

[5] Qiang Wang, "A Hybrid Sampling SVM Approach to Imbalanced Data Classification", Hindawi Publishing Corporation, 2014.

[6] Ionut Brandsoiu,"Churn Prediction in the Telecommunications Sector using Support Vector Machine", Annals of the Oradea university, May 2013.

[7] Punam Mulak, Nitin Talhar, "Analysis of Distance Measures using K-Nearest Neighbour algorithm on KDD Dataset", International Journal of Science and Research, 2013.

[8] Yi-Fan Wang, Ding-an chiang, "A recommender system to avoid customer churn : A case study", Expert Systems with Applications, 2009.

[9] S. J. Yen and Y. S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," Expert Systems with Applications, vol. 36, pp. 5718–5727, 2009.

[10] Chawla, N. V, Bowyer, K. W., & Hall, L. O. "SMOTE: Synthetic Minority Over-sampling Technique", 16, 321–357, 2002.

[11] S. Babu, N. R. Ananthanarayanan, V.Ramesh, "A Study on efficiency of Decision Tree and Multi Layer Perceptron to Predict the Customer Churn in Telecommunication Using WEKA", International journal of computer applications (IJCA), volume 140, issue 4, April 16.

[12] Drummond, C., & Holte, R. C., "C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling". In Proceedings of the ICML'03 workshop on learning from imbalanced datasets, 2003

[13] Tom Fawcett, "ROC Graphs: Notes and Practical Considerations for Data Mining Researchers", Intelligent Enterprise Technologies Laboratory, HP Laboratories Palo Alto, HPL-2003-4, January, 2003.

[14] Kiran Dahiya, Surbhi Bhatia, "Customer Churn Analysis in Telecom Industry", IEEE, 2015.

[15] Kubat, M., Holte, R., & Matwin, S. "Machine Learning for the Detection of Oil Spills in Satellite Radar Images. Machine Learning", 30, 195–215, 1998.

[16] Batista, G. E., Prati, R. C., & Monard, M. C. "A study of the behavior of several methods for balancing machine learning training data",. ACM SIGKDD Explorations Newsletter, 6(1), 20-29, 2004.

[17] Batista, G. E., Bazzan, A. L., & Monard, M. C, "Balancing training data for automated annotation of keywords: A case study". WOB, 10-18, 2003.

[18] Wiesław Chmielnicki, Katarzyna Sta²Por , "Using The One–Versus–Rest Strategy With Samples Balancing To Improve Pair wise Coupling Classification", Int. J. Appl. Math. Computer. Sci., 2016, Vol. 26, No. 1, 191–201.

[19] Benjamin Oghojafor, Godson Mesike, "Discriminant Analysis of Factors Affecting Telecoms Customer Churn", International Journal of Business Administration Vol. 3, No. 2; March 2012.

[20] *https://en.wikipedia.org/wiki/Weka_(machine _learning)*

[21] *https://sourceforge.net/projects/ikvm/files*