A Hybrid System for Customer Churn Prediction and Retention Analysis via Supervised Learning in Telecom Industry

By

Soban Arshad

CIIT/FA16-RCS-030/ATK

MS Thesis

In

Computer Science

COMSATS University Islamabad
Attock Campus-Pakistan
FALL, 2019

COMSATS University Islamabad
Attock Campus

A Hybrid System for Customer Churn Prediction and Retention Analysis via Supervised Learning in Telecom Industry

A Thesis Presented to

COMSATS University Islamabad, Attock Campus

In partial fulfillment
of the requirement for the degree of

MS(CS)

By

Soban Arshad

CIIT/FA16-RCS-030/ATK

FALL, 2019

A Hybrid System for Customer Churn Prediction and Retention Analysis via Supervised Learning in Telecom Industry

A Post Graduate Thesis submitted to the Department of Computer Science as partial fulfillment of the requirement for the award of Degree of MS(CS).

Name	Registration Number
Soban Arshad	CIIT/FA16-RCS-030/ATK

Supervisor

Dr. Khalid Iqbal
Assistant Professor, Department of Computer Science
COMSATS University Islamabad, Attock Campus

February , 2020

Final Approval

This Thesis titled

A Hybrid System for Customer Churn Prediction and Retention Analysis via Supervised Learning in Telecom Industry

By

Soban Arshad

CIIT/FA16-RCS-030/ATK

Has been approved

For the COMSATS University Islamabad, Attock Campus

External Examiner: ______________________________

Dr……………………………………………

Supervisor: ______________________________

Dr. Khalid Iqbal
Department of Computer Science, COMSATS University Islamabad, Attock Campus

Co-Supervisor:______________________________

Dr. Rashid Ahmed
Department of Computer Science, COMSATS University Islamabad, Attock Campus

HoD: ______________________________

Dr. Khalid Mahmood Awan
Department of Computer Science, COMSATS University Islamabad, Attock Campus

Declaration

I Soban Arshad, Registration number: CIIT/FA16-RCS-030/ATK hereby declare that I have produced the work presented in this thesis, during the scheduled period of study. I also declare that I have not taken any material from any source except referred to wherever due that amount of plagiarism is within acceptable range. If a violation of HEC rules on research has occurred in this thesis, I shall be liable to punishable action under the plagiarism rules of the HEC.

Date: _________________

Signature of the student:

____

Soban Arshad

CIIT/FA16-RCS-030/ATK

Certificate

It is certified that Soban Arshad, Registration number: CIIT/FA16-RCS-030/ATK has carried out all the work related to this thesis under my supervision at the Department of Computer Science, COMSATS University Islamabad, Attock campus and the work fulfills the requirement for award of MS degree.

Date:______________

Supervisor: ____________________

Dr. Khalid Iqbal

Assistant Professor

Head of Department:

____

Dr. Khalid Mahmood Awan

Department of Computer Science

COMSATS University Islamabad , Attock Campus

To My Dear Family and Teachers

ACKNOWLEDGEMENTS

This thesis gave me confidence how to manage difficult task in a tough schedule of studies along with job. Many challenges has been came during this task but thanks to Allah, Almighty who made me strong enough to handle all the situations and made me comfort by completing my thesis work with high potential.

Specially I would appreciate and thanks from deep by my heart to my great supervisor, Dr. Khalid Iqbal, and co-supervisor Dr. Rashid Ahmed for their constructive counselling, giving respect, and calmness and provided me the study environment that every student wish to have.

It should may be difficult still to do my thesis without my friend's guidance, financial support from my brother and to my wife for making me strong to face any kind of problem with patience.

To my believe it will never be happened if I have no prays behind my success, love to my mother and father to be available for me throughout in my studies and specially for MS Degree.

Soban Arshad

CIIT/FA16-RCS-030/ATK

ABSTRACT

This study is an explorative piece of research was performed to enhance churn prediction in telecom industry. Telecommunication companies are well aware of the fact that a satisfactory relationship between customer and service provider results increase in company revenue. The term customer churn refers to customers those who leave companies services in near future. Precise and right time recognition of future churners in telecom industry requires churn prediction model to retain their customers. Retention not only contributes in profitability of an organization but also uphold the position in the competitive market. Also competitors are acknowledged that gaining new customers is more expensive and difficult task then retain their valued customers for long terms. In the past decade researchers proposed various different churn prediction models, however, Design an efficient and accurate model was a barrier in their studies on big datasets. Data sparsity, noisy data, and imbalanced nature of dataset are the main factors. To resolve the limitations addressed above in this era for an efficient performance, we proposed a hybrid model using Synthetic Minority Over-sampling Technique (SMOTE) and Particle Swarm Optimization (PSO) for removal of imbalance class issue and feature selection. Furthermore, Pre-Processing Stage was considered for data cleaning and normalization purpose. A huge dataset has been considered provided by a telecom service provider company named Orange. Substantial Experiments were done to test and validate the model on Random Forest (RF), Logistic Regression (LR), Naïve Bayes (NB) and XG-Boost. In this study, first most contribution in this study is to predict the churner. Performance evaluation shows that XG Boost has less error rate and greater Area under Receiver Operating Characteristic (ROC) of 98 % as compared with other methods.

Table of Contents

List of Figures

List of Tables

LIST OF ABBREVIATIONS

SMOTE	Synthetic Minority Over-sampling Technique
PSO	Particle Swarm Optimization
RF	Random Forest
LR	Logistic Regression
NB	Naïve Bayes
XG-Boost	eXtreme Gradient Boosting
AUROC	Area Under Receiver Operating Characteristic
KPI's	Key Performance Indicators
SVM	Support Vector Machine
AUC	Area Under Curve
KNN	K-nearest neighbor
DT	Decision Tree
ANN	Artificial Neural Network
ROS	Random Over Sampling
RUS	Random Under sampling
CLUS	Cluster based Under Sampling technique
ACO	Ant Colony Optimization
ABC	Artificial Bee Colony
OTD	Orange Telecom Dataset
PSOFS	Particle Swarm Optimization based Feature Selection
TPR	True Positive Rate
FPR	False Positive Rate

CHAPTER 1

Introduction

After 1990s the one major development area for an industrialized nations is the telecommunication sector. In [1] the author examined that increase in technology directly impact in increased number of competitors, hence competition arise. In this chapter general context along with motivation and previous work issues have been briefly discussed. Also contains a section of research questions that must be followed throughout the research to answer. Providing with a background study in telecom industry along with the aims and contribution done in our work.

1.1 Context

Customers are the directly influencing factor towards the company revenue. Customers are the assists of a company. As technology enhanced the number of competitors also saturated rapidly, therefore customers have plenty of choices to move between one another. Some identified reasons that make customer to churn are behavior of company representatives towards customers, low reachability services, high data rate price and lastly the easiness of portability between service providers without changing the cellular number. Considering all these factors companies are taking customers churn on serious note as it had directly influence on the profitability of the company. Telecom companies are well aware of that fact as they know it's the only easy way to gain profit with less loss as compare to gain new customers [2]. To predict churner factors must not be the same for each company.

1.2 Motivation

It has been noticed that a verity of work have already been done in the era of customer churn prediction. Researchers used different techniques and datasets to identify the customers to move from the current company to other service provider in near future. Taking in consideration the previous work in this era, however some factors made there model less efficient in churn prediction.

1.3 Background

Every organizations wish is to become successful and remain in business. Their major focus is to gain profit, profitability make them to survive long term. To make their wish complete companies focuses on increase in the number of customers. To accomplish that, companies target to acquire new customers or focus on retaining existing customers. In an analysis, it was noticed that companies pay five to twenty times more cost on gaining new customers as compared to the cost utilize on keeping the existing customers [3] [4].

In the previous work the validation was done on small benchmark datasets instead of working on big dataset like orange telecom data that contains large number of garbage , missing values and noisy data. As the researchers are well familiar with the fact that nature of dataset had a major impact on performance evaluation. Hence a much needed requirement to develop such a mechanism to predict customers churn on such large datasets in telecom industry along with retention mechanism [5] [6].

Multiple different technique are being considered to identify customer churn. Considerable research had been practiced in the field of churn prediction using statistical and datamining techniques in their time period. In different datamining techniques based on call detail records for churn prediction was considered. In [7] the author proposed churn prediction in subscription based organizations, by combining historical data along with Key Performance Indicators (KPI's). In [8] proposed a Questionnaire based data collection technique processed over EMOTE classifier. In [9, 10] Support Vector Machine (SVM) methodology using expert viewpoints for linearly separable and by Support Vector Data Description (SVDD) non-linearly problems was resolved respectively. In [11] the researcher suggested a feature selection mechanism using SVM. Bravo et al. in [12] proposed Relational and the non-relational learner's classifiers handling data sparsity by social networks analytics method. Gray et al. in [13] suggested Netlogo (agent-based model) by upgrading the cell network from 3G to 4G. Azeem et al. in [5, 6] predict future churners using fuzzy classifier along with retention campaign model that works only for Call Detail Records (CDR) dataset. Deng et al. in [14] proposed churn prediction in Big data by using RF classifier with Hive and Spark SQL for prediction and feature selection respectively.

1.4 Research Challenges

This section consist of some exceptional relevant research challenges which are not considered due to too limited resources and timeframe.

Right off the bat, we only examined uniquely dataset for feature selection in our study. Additionally, we have not tested our proposed model on multiple datasets due to inaccessibility of dataset from other telecom service providers.

Also, we have utilized statistical strategy Particle Swarm Optimization (PSO) to achieve feature selection goal and have not examined different methods of feature selection due to time constraints. Industrial desire is to check the impact of other feature selection techniques to predict high prediction accuracy.

However finally, previous work shows different performance evaluation metrices in there study. Hence we used Accuracy, Precision, Recall, Specificity, F1-Score and for visualization technique Area Under Curve (AUC) the commonly used method in this era was considered. So we used the most commonly used methods to our best knowledge to visualize the accuracy without considering the other performance evaluation methods.

1.5 Research Questions

RQ1. Does data pre-processing plays important role in accurate prediction ?

RQ2. What is the effect on performance after feature selection on large dataset ?

RQ3. What are the most popular methods in churn prediction problem ?

1.6 Research Contribution

In the title of our work shows comprehensive aim of this thesis. The primary aim is to design a model that predict customer churner accurately and efficiently for big dataset like orange dataset used in our study. Considerable main aims and objectives are as follows

- Moving from manual prediction techniques to automated solution to recognize the churner or non-churners by taking dataset from base and doing some computational methods to give final prediction analysis.

- Proposed a model covering major terminologies preprocessing, class imbalance, feature selection, cross validation and finally predict churner and non-churner.
- In the end model evaluation was done on most popular classifiers and compared classifiers performance to determine better performance in our proposed work.

1.7 Thesis Outline

To describe and achieve the issues and objectives in this era, this thesis is further divided in following sections:

Chapter 2 thoroughly describes the literature review of this research. Providing the previous work related to the customer churn prediction in the telecom industry. In this chapter Classifier based techniques, hybrid classifiers and some PSO based techniques have been considered in this chapter. Finally, in the end of literature review summary have been discussed.

Chapter 3 contains detail explanation about the proposed methodology. This chapter further divided in subsections. Section 3.1 contains dataset division 3.2 consists of pre-processing methods adopted. Section 3.3 tells about the PSO based feature selection procedure. Section 3.4 describes the cross-validation technique adopted. Section 3.5 demonstrate the performance evolution on multiple classifiers. At the end of this chapter summary of proposed methodology has been discussed.

Chapter 4 describes the implementation procedure implemented on the proposed methodology. Big dataset named as orange dataset taken from telecom industry had been considered for experiment. Dataset consists of 50,000 rows with 15,000 features along with class label. To maintain the security of customers features are not been labeled.

Chapter 5 shows comparison between researches have been proposed techniques introduced in past decade. Also this chapter contains comparison between different datasets and feature selection methods. Finally, results can be visualized through performance Metrix accuracy under curve AUC.

In the last chapter 6, summary of research work has been explained along with conclusion and future work directions should be considered after this research.

CHAPTER 2

Literature Review

In this chapter a detailed review conducted related with customer churn. Showing some bibliographic content found in this era. Customer churn prediction is a trending research in this time period. Our study main focus is to identify the techniques and methods in prediction of customer churn. Starting with exploring some areas in which customer churn had been noticed. Further moving towards the trending research area related to customer churn in telecommunication industry. Moreover, in this study along with some previous techniques we also examined in detail about some preprocessing methods used to resolve this problem. At the end summary of this chapter have been overviewed.

2.1 Customer Churn

As conducted in [15] by author, telecommunication field 'customer churn' is defined as a customer will switch from one service provider to another in near future. To retain their valuable customers telecom companies follow 'customer management' procedure.

2.2 Churn prediction in saturated market

Customer churn is a globally trending research in a highly saturated market. Saturated market is not only consists of telecom industry but had a lot many other industrial areas that are facing the same customer churn problem. Customer churn problem is also highly faced in retail food industry [16, 17], vehicles industry [18-20], gaming industry [21] and electronics industry [22]. Retaining the existing customers is cost effective technique rather than to gain new customers [23] Industrial need is to identify the accurate way to predict the churners on which they can apply some retention techniques to retain only those specific customers. Hence the major concern is to predict the churners accurately those who are the real one to churn.

Firstly in the food industry a comparative competition arisen in saturated market. Major factors noticed in the food industry multiple option for shopping and un predictable customers habits. Establishment of super stores make the food market competitors to think differently to retain their customers for remaining their position in industry [16]. Secondly in [18] vehicles industry author colovic and Mayrhofer examined some well-known Europe , Japan, and north American multinational companies. The factors considered in the study was globalization ,regional base and analysis on location wise selection of automobile. In their research study they had shown as automotive industry become in competition with as like other industries since 1980s. Becoming in competition auto makers start shifting from national industry to a multinational industry. Hence, auto maker faced challenges due to moving in global industry. In the saturated market manufacturers have to keenly consider some major factors to compete with other auto making companies. Factors discussed in the study was cost, quality and flexibility of auto mobile. To retain their customers manufacturers started to offer some productive services.

To maintain a reputative status in the saturated market of automobile the business administrators focused on the maintaining customers loyalty. Loyal customers are much productive for a company as they promote the brand through their words. If customers are loyal they give good remarks about the brand they are trusting and promote it in their circle of friends and family, also they will again purchase the same company product again and again [24].

Thirdly Gaming industry another industry in saturated market is a trending research now a days. To retain their valued customers this industry also worried for that. Researchers designed some players data driven approaches to identify churners [25]. And after on meta-analysis performed on machine learning techniques considered Wuzzit trouble games for churn prediction [26].

At last moving in the technology world electronics industry playing a repid competition between companies in saturated market. Electronics industry had multiple industries in it like mobile phones, televisions, computers and refrigerators. Same as in other saturated market competitors electronics industry also focuses on the customers retention rather than to gain new customers [27].

Churn prediction on Credit card was presented by [28] used hybrid classification technique based on supervised learning and rough clustering. The proposed model utilize the most commonly used machine learning algorithms like SVM, RF, Decision Tree (DT), Naïve Bayes and K-nearest neighbor (KNN).

2.3 Churn prediction in telecom Market

After considering customer churn in the above mentioned industries a more comitative market in saturated market in telecommunication industry. Telecom industry became a trending research in the current time period. Telecom service providers are more concerned to maintain their reputation in-between their competitors by showing power of their valued customers. This can only be happened if they can gain the new customers as well as by retaining their existing ones. Business analyst had a determined that cost on gaining new customers is much more then retaining the existing ones [3, 4]. Hence a question must arise in the analyst mind that " will this valued customer will churn or not ? " and the answer will be short in terms of yes or no. Therefore, Customer churn analysis is a binary classification problem.

To predict the customer churn researchers did a detailed work on that emerging issue held in telecommunication sector. For that different techniques had been introduced which can be categorized as classifier based, hybrid and PSO based.

2.3.1 Churn prediction using classifier based model

In [7] the author proposed churn prediction in subscription based organizations, by combining historical data along with Key Performance Indicators (KPI's). Babu et al. in [8] Questionnaire based data collection technique processed over EMOTE classifier. Dong et al. in [9], Maldonado et al. in [10] SVM methodology using expert viewpoints for linearly separable and by Support Vector Data Description (SVDD) non-linearly problems resolved. Flores et al. in [11] suggested a feature selection mechanism using SVM. Bravo et al. in [12] proposed Relational and the non-relational learner's classifiers handling data sparsity by social networks analytics method. Gray et al. in [13] suggested Netlogo (agent-based model) by upgrading the cell network from 3G to 4G. Azeem et al. in [5, 6] predict future churners using fuzzy classifier along with retention campaign model that works only for Call Detail Records (CDR) dataset. Deng et al. in [14] proposed churn prediction in Big data by using RF classifier with Hive and Spark SQL for prediction and feature selection respectively. In Table 2.1 it is shown the different techniques used by researches considering the classifiers based models.

Table 2. 1 Classifier based Churn Prediction Approaches

Reference Paper	Technique
Morabito et al. [7]	Key Performance Indicators (KPI's)
Babu et al. [8]	EMOTE
Dong et al. [9]	SVM
Maldonado et al. [10]	SVDD
Flores et al. [11]	FS using SVM
Bravo et al. [12]	Social networks analytics
Gray et al. [13]	Netlogo
Azeem et al. [5, 6]	fuzzy classifier
Deng et al. [14]	Random Forest

2.3.2 Chun prediction using hybrid model

Vijaya et al. in [29] investigated on hybrid data mining approach in telecom. K-Means used for groping the customers with similar features, then training and testing by DT or Naive Bayes (NB). Koen et al. in [30] proposed Hybrid algorithm logit leaf model, combination of DT and LR. Focused on built-in feature selection mechanism and selects the variables for group separately. Huang et al. in [31] combines a modified k-means clustering algorithm and a classic rule inductive technique (FOIL). Table 2.2 shows hybrid models techniques used by past researchers.

Table 2. 2 Hybrid Models for Churn Prediction

Reference Paper	Technique
Vijaya et al. [29]	K-Means , DT, NB
Koen et al. [30]	Logit leaf (DT,LR)
Huang et al. [31]	K-Means with FOIL

2.3.3 Churn prediction using PSO based model

Idris et al. in [32] handled imbalance data distribution by feature selection using PSO, moreover used RF for performance evaluation. Xue et al. in [33]and Lucija et al. [34] proposed BPSO+C4.5 based on the binary version of PSO for features selection. Vijaya et al. in [35] extracted features using PSO in telecom. Pre-processing and simulated annealing was done to resolve the imbalance and struck in local optima issues. In Table 2.3 shows previous work on churn prediction using PSO based model along with some classifiers techniques.

Table 2. 3 PSO based Feature Selection model

Reference Paper	Technique
Idris et al. in [32]	RF with PSO-FS
Xue et al. in [33] & Lucija et al. [34]	BPSO + C4.5
Vijaya et al. in [35]	PSO-FSSA

2.4 Architecture for churn prediction

Churn prediction involves a common structure after considering the previous studies held in this field. First and the most basic requirement is to get a suitable dataset for doing experiment and predict an effective output on it. After collecting the related dataset, if some purification stage on dataset is required then pass that dataset from preprocessing stage making in mind not to lose data quality. Preprocessing stage consists of data cleaning, normalization and resolving data imbalance issues. Feature selection is required to identify the effecting features that should be taken into consideration for predicting churner. At the last performance evaluation had to be done for analysis purpose.

2.4.1 Dataset

The first and basic requirement is collection of data. Dataset selection is initial stage in the architecture of churn prediction. It is necessary to have a proper dataset for getting good performance evaluation. Multiple bench mark and real world datasets have been examined by previous researcher. Zeng et al in [14] considered prepaid mobile operators in China for churn prediction, gained 0.96 precision on top 50000 predicted churners with 150 number of features . Sharma et al. in [36] experimented on a benchmark dataset using Artificial Neural Network (ANN) . Chen et al in [37] predicted churners by using SVM on 20 attributes. Buckley et al in [38] determined churn prediction on Ireland telecom company. Idris et al in [39] considered benchmarked telecom dataset of Duke university named as CelltoCell. Kim et al. in [40] used real world dataset with 29 attributes, used Random Forest algorithm and obtain 0.947 Area under ROC curve on South Asian GSM company. Moreover a benchmark dataset churn-bigml was also taken with reported result of 0.835 ROC. Reddy et al. in [41] shows experimental results after applying Random Forest with less error rate and greater accuracy of 91.66%. Table 2.4 demonstrate about the different telecom dataset have been considered by the researchers.

Table 2. 4 Different Datasets used in telecom sector

Author	Benchmark	Number of features	Numbers of instances
Zeng et al. [14]	No	150	50,000
Sharma et al. [36]	Yes	20	2,427
Chen et al. [37]	No	20	86,837
Buckley et al. [38]	No	738	8,27,124
Kim et al. [40]	No	29	19,213
Reddy et al. [41]	Yes	17	3,333

2.4.1.1 Orange Telecom dataset

In [42] international competition KDD Cup, 2009 dataset has been examined by the author. Different data mining algorithms were applied containing Decision table, j48, Naïve Bayes, K-Star, ADTree,IB1 and ADA Boost. The best accuracy of 89% was obtained.

In [43] firefly algorithm and hybrid firefly algorithm were been practiced with the experimental result of 86.36% and 86.38% of accuracy respectfully. While taking the comparison between both the techniques it was noticed that they have much less difference in accuracy. On contrary hybrid firefly showed less time latency then the firefly competitor on orange telecom dataset.

In [44] Genetic Programming (GP) and AdaBoost were integrated to get high-performance churn prediction. Class imbalance issue was resolved by Particle Swarm Optimization (PSO) under-sampling technique with performance of 0.86 AUC on orange dataset after implementation.

Idris et al. in [45] considered orange telecom dataset and predicted results on RotBoost based ensemble technique by obtaining 0.601 AUC. Vijaya et al. in [35] obtained accuracy of 90.65 on orange telecom data by just considering 1000 records out of 50,000 used a hybrid model of PSO based feature selection and simulated annealing techniques. Different methodologies implemented on OTD with performance evaluation can be visualized in Table 2.5.

Table 2. 5 Different Techniques utilized on OTD

Reference	Methodology	Performance
S. Sladojevic et al. [42]	DT, K-Star, NB, J48	89% Acc.
Maheswari et al. [43]	Hybrid Firefly	86.38% Acc.
A. Idris et al. [44]	GP + ADA Boost	0.86 AUC
Idris et al. [45]	Rotation Boost	0.761 AUC
Vijaya et al. [35]	PSO-FSSA	90.65 % Acc.

2.4.2 Class imbalance

In a reputative company churn ratio is to low and interesting to identify the churners in such scenarios. Telecom industry had a huge number of customers registered but few of them churn and majority of the customers are non-churners. Due to which class imbalance issue arises. Not handled class imbalance issue show the biased predictions towards non-churner (majority class) using well known algorithms Zhu et al. [46].

Number of different strategies have been proposed by different researchers to resolve the class imbalance problem.

We can divide the solutions of class imbalance problem into two main categories: data-level and algorithm level. Considering the first category the data-level solutions they resample the original data. On the other hand, second category the algorithm-level solutions use some predefined learning algorithms by adapting their learning capability considering the churner class (minority class). Draw back of the algorithm level solution is that it needs deep knowledge about the classifier and domain knowledge in which the application is going to be applied. Hence the data-level algorithms are the most favorite among the researchers as they had no concern of learning algorithms.

Considering the data-level solution two most commonly used random sampling techniques, one Random Oversampling (ROS) and the second Random Under sampling (RUS). ROS basically randomly replicate the minority class data from the original dataset provided. On the other hand RUS do opposite to that instead of increasing the minority class it randomly remove majority class instances from the original dataset to resolve the imbalance issue. Hence, both the sampling techniques have some advantages and disadvantages : both the techniques are easy to implement in any application domain but as RUS remove the instances it is hard to know that did RUS removed the unusful instances or deleted the useful data required for accurate prediction. And on contrary, ROS increases the number or minority class instances which may result in overfitting. Chawala et al. in [47] considered these issues and proposed an intelligent sampling technique Synthetic Minority Oversampling Technique (SMOTE) resolved those issues happened to the past researchers related to ROS and RUS. SMOTE architecture consist of randomly selection process from the minority class in an intelligent way by taking one or more than one nearest neighbors. After on it provide a new instances which lies in-between the original instance and the randomly selected nearest neighbors. Yen et al. in [48] proposed a Cluster based Under Sampling technique (CLUS) to resolve the class imbalance issue using data-level approach. Burez et al. in [49] discussed about the importance of churner class or minority class. Predicted a difference in churner and non-churner by visualization method. Used ROC and Area under ROC curve which can be visualized clearly as churner and non-churner is a binary class problem. Amin et al. in [50, 51] performed experiment on class imbalance by some effective sampling techniques to predict customers churn.

Secondly, algorithm-level solutions as stated early are concerned with the learning of classification algorithm. Kai et al. in [52] proposed cost-sensitive learning technique. This method assign a cost to misclassification, higher cost to the churner (minority class) than for the cost for misclassification of majority class. The purpose of this method is to minimize the total cost.

2.4.3 Feature selection

Feature selection can be defined as a procedure of selecting subset of features ($n \subseteq N$) from a pool of features (N). The prosses consists of identifying and then selecting the important and most relevant features from a provided dataset. The objective of doing this is it reduces the computational cost and also give good evaluation result at the end. Dealing with big datasets feature selection is an important phase in the architecture of churn prediction. Many researchers used different techniques to identify relevant feature in the literature.

Yi et al. in [53] proposed Recursive feature selection based on support vector machine (SVM-RFE) for selecting the relevant features. Gave better performance than other models such as ANN, Naïve Bayes and decision tree. In [54]experimented to find feature selection via Ant colony optimization (ACO) merged with rough set theory. Researchers did experiments on some swarm intelligence algorithms where Suguna et al. in [55] proposed Artificial Bee Colony (ABC) merged with rough set theory for feature selection.

Feature selection in telecom industry was less focused by the researchers. This factor was eliminated by the researchers due to having less features in the dataset so by selecting features researchers are afraid of losing some important features.

Azeem et al. in [56] Feature selection was considered in land-line telecom industry. New window technique was introduced giving model evaluation on ANN, SVM and DT with increase in performance of 3% to 5%.

2.5 Summary

Overviewing the past work it is noticed that churn prediction is a challenging task faced by different researchers. However their predicted model showed some limitations due to which they are lack in getting good performance. Due to our best knowledge PSO based feature selection and resolving class imbalance is the need of current time period. Along with calculating performance with best classifiers to predict accurate churner is need of the telecom industry. To resolve the challenges have been noticed in the past research, next chapter will discuss about the purposed methodology.

Chapter 3

Methodology

In the last chapter, the issues addressed in that chapter had been resolved by churn prediction model proposed in our study. In this chapter proposed model has been discussed in detail. Proposed model for an efficient and precise churn prediction has been shown in Figure 3.1. Furthermore one by one each module of the model has been explained. Big dataset of telecom provider company has been taken into consideration. Model will be work best for datasets having a large number of variables as Orange telecom dataset has. Minimize the variables by selecting the best ones for resolving specifically the churn prediction problems.

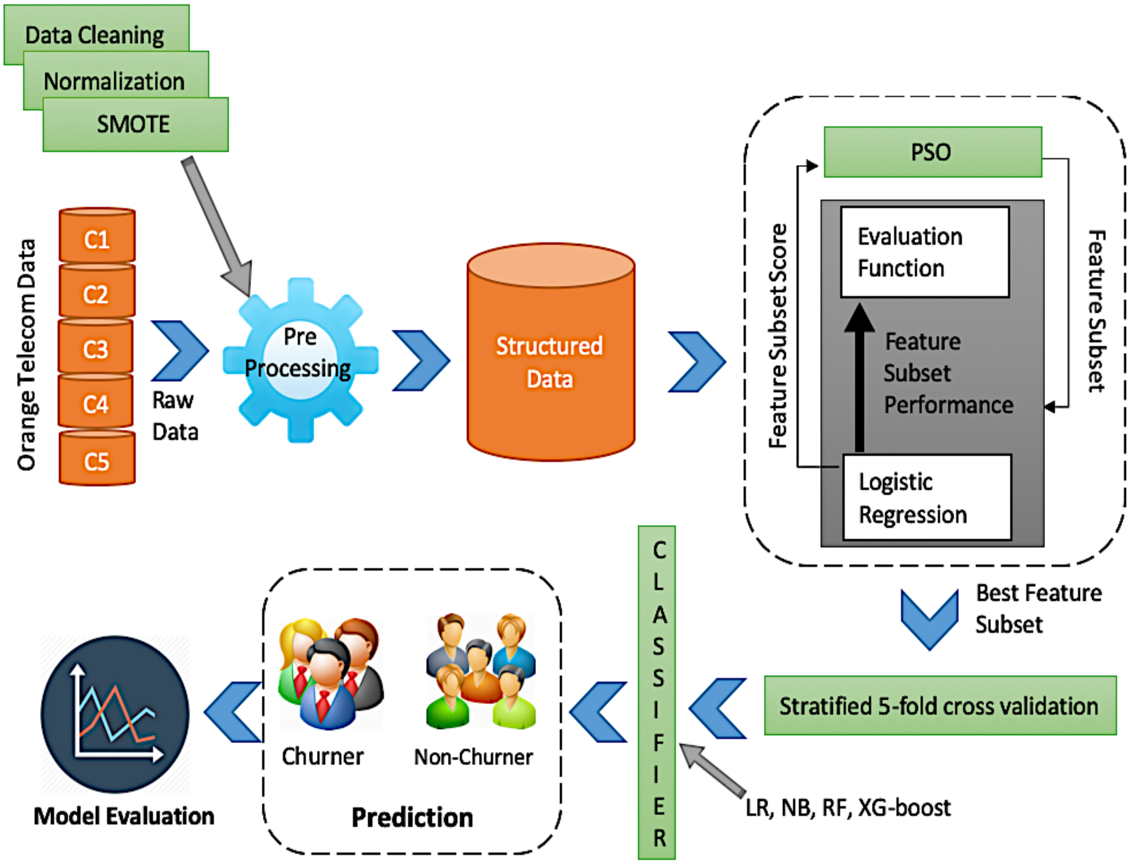


Figure 3. 1 Precise churn prediction proposed model

3.1 Dataset

To identify and determine the efficiency of the proposed model a relative dataset was required. Remembering the title of our work telecom dataset with a large number of variables should be taken for predicting churner and non-churner. Hence a publicly available dataset provided by French telecom company Orange was taken in our study.

Orange Telecom Dataset (OTD) was openly available dataset contains the data about the subscribers of the OTD and also was considered in KDD cup for customer churn prediction [14].

3.1.1 Orange Telecom Dataset

It is a big dataset consists of 15,000 variables and 50,000 instances; dataset was further divided into number of 5 chunks (C1,C2,C3,C4,C5) that contains equally number of samples (10,000 each). Furthermore out of 50,000 samples 3672 and 46328 samples were churners and non-churners respectively. Approximate percentage ratio between churner and non-churner in OTD is 7 : 93 respectively due to which class imbalance problem occurs in such dataset, Figure 3.2 shows a graphical representation between churner vs non churners. Orange telecom company maintains the privacy of their customers by veiled the variables which increased the difficulty to identify what the attribute actually means.

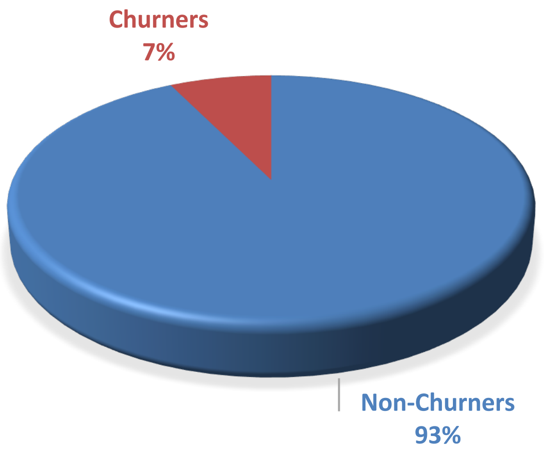


Figure 3. 2 Churner vs Non-Churner in OTD

OTD is a heterogenous dataset which consists of noisy data with variations in the measurement scale, features with null values, features with missing values, and data sparsity. Hence for which data preprocessing is a necessary requirement on such kind of datasets.

3.2 Pre-Processing

After selection of dataset we noticed that it is raw data must need some processing to get informational data. OTD consists of garbage , null values, data sparsity and inconsistent values which must be resolved. Preprocessing is necessary step to take for getting meaning and informative data.

3.2.1 Data Cleaning

Orange dataset have garbage, null values and variables having sum equals to zero. Data cleaning stage basically identify and try to resolve the highlighted issues to provide some cleaned data. Highlighted issues may leads towards inaccurate result. Most commonly issues lies in the big datasets like OTD it contains a lot of null values in shape of space or any special characters that needs to be resolved either by removing or by replacing relevant value. Null values are replaced by the mean value of each of its own variable and removed those columns with sum equals to zero. After removal of irrelevant columns and replacing the missing values, new data formation after cleaning stage is shown in Figure 3.6 with chunk-wise visualization.

3.2.2 Data Normalization

After cleaning stage it was noticed in the dataset that is out of measuring scale can be seen in Table 3.1 hence it is required to manage all the data into a range. Normalization can be described as adjusting values measured on different scale in to a common scale such that values in-between 0 and 1. Result prediction was not be an easy task on such dataset whose data range cannot be visualized. This kind of data may misleads the classifier.

Table 3. 1 Dataset before Normalization

Var 65	Var 66	Var 67	Var 68	Var 69	Var 70	Var 71	Churn
4664	0	659337	0	0	31005.63	598.36	-1
296	0	52941	0	0	-28620	1065.19	-1
9896	0	1.72E+07	0	0	-266031	962.85	-1
1800	0	1.84E+07	0	0	737358.3	1228.43	1
3432	0	1.85E+07	0	0	1452285	1082.41	-1
392	0	648932	0	0	-5198040	230.65	-1
3784	0	789593	0	0	-326838.6	728.7	-1

Orange dataset consist of data sparsity to remove that issue we normalized whole data between 0 and 1 up-to 3 decimal point using the Min-Max scalar normalization technique. Table 3.2 shows the dataset after normalization. Min-Max scalar makes the normalization phase easy and simple illustrated in the Equation 3.1. In which X_{min} and X_{max} are the minimum and maximum values and X is the observed value in the given set.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3.1}$$

Table 3. 2 Dataset after Normalization

Var 40	Var 41	Var 42	Var 43	Var 44	Var 45	Var 46	Churn
0.007	0	0	0	0.012	0	0.829	-1
0	0	0	0	0	0	0.829	-1
0.055	0	0	0	0.045	0	0.829	-1
0.036	0	0	0	0	0	0.829	1
0.011	0	0	0	0.018	0	0.829	-1
0	0.03	0	0	0	0	0.829	-1
0.029	0	0	0	0.014	0	0.829	-1

3.2.3 Class Imbalance

Orange telecom dataset it was clearly examined that ratio between churner and non-churner varies in large amount of only 7% of churners and 93% of non-churners due to which a class imbalance issue arise. The Figure 3.3 explains the exact situation of churner and non-churner on original orange telecom dataset after the division of dataset into five number of chunks. In each chunk it was noticed that the minority class (Churners) are very few in numbers then the majority class (Non-Churner). Imbalance data misleads and give biased prediction on applied classifiers in the final results.

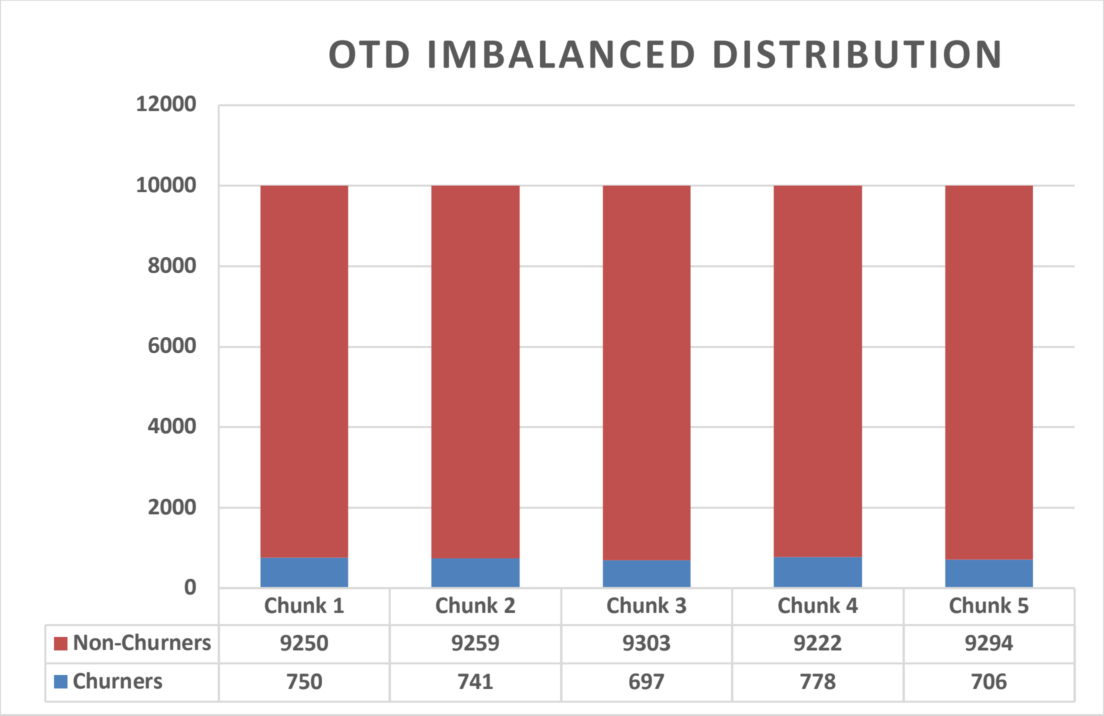
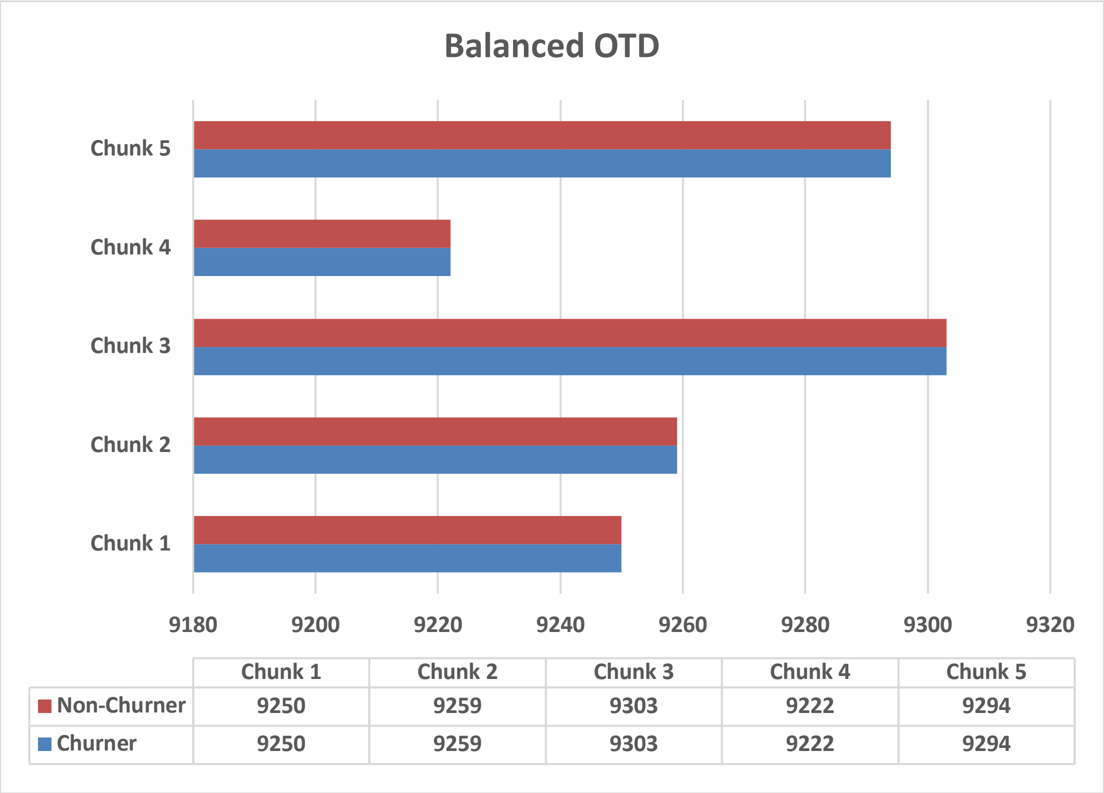


Figure 3. 3 Chunk wise OTD Churner vs Non-Churner

To resolve that issue the quantity of churners must have to be raised. We resolved this issue by using an advance technique known as SMOTE rather than considering the old simple RUS and ROS techniques. SMOTE increases the quantity of churners without any effect on majority class the non-churners. SMOTE generates the new instances by considering the old ones only of the churner class, starting with selecting some random points of churner class and generate new instance from the existing ones by selecting the neighbor via K nearest neighbor algorithm. At last we became in a position that ratio between churner and non-churner is 1 : 1 after gaining the objective of resolving the class imbalance issue can be visualized in Figure 3.4.



	Chunk 1	Chunk 2	Chunk 3	Chunk 4	Chunk 5
🟥 Non-Churner	9250	9259	9303	9222	9294
🟦 Churner	9250	9259	9303	9222	9294

Figure 3. 4 Balanced OTD Chunks after SMOTE

The Figure 3.5 shows the generic working of SMOTE creating new minority class instances from the existing ones. This oversampling technique does not create duplicate instances like Random Over Sampling (ROS).

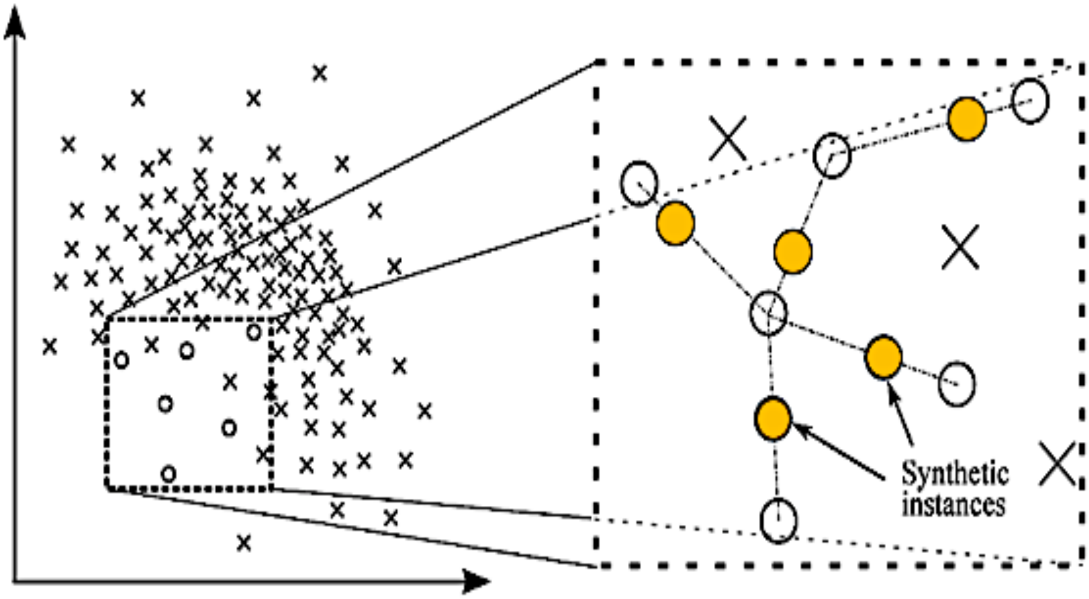


Figure 3. 5 Generic SMOTE synthetic churner instances creation

3.3 Feature Selection

Pre-processing stage only removed the columns having the null values, garbage values. Feature selection stage is required to only get those features that are the effecting one in accurate prediction of churner or non-churner. Firstly feature selection was done on some basic domain knowledge furthermore PSO based feature selection technique was used.

3.3.1 Particle Swarm Optimization (PSO) based Feature Selection

Particle Swarm Optimization based feature selection (PSOFS) process is also a part of data preprocessing which involves the elimination of features that are not effective in churn prediction. The advantage of PSOFS is it reduced the big orange telecom data, which contributed to make the churn prediction process speedy. Moreover due to sparsity nature of orange dataset there exists few useless features. Particle swarm optimization algorithm can easily recognize and remove those features effectively. Moreover to gain the accuracy it is required to only identify the relevant features that put up some good results of churn prediction.

Initializing with the first step it takes a set of features, and tries to find the relevant features for best churn prediction using a position-velocity update method. The velocity update rule is comparison between the particles present position with the positions of its neighbors. We set number of particles as 50 for better performance by taking dimensions equals to the total number of features in the dataset. And the position update rule will be decided if the output of the sigmoid function is less than a random number created between a threshold of 0 and 1. Furthermore the performance of the feature subset was determined and stored on the bases of Logistic Regression classifier. Lastly the global best selected features X = [$x_1,x_2,x_3, \ldots , x_i$] have been taken for further evaluation purposes. In Figure 3.6 final result shows the features selected after the implementation of PSO based feature selection algorithm, making the dataset purified from basic raw dataset.

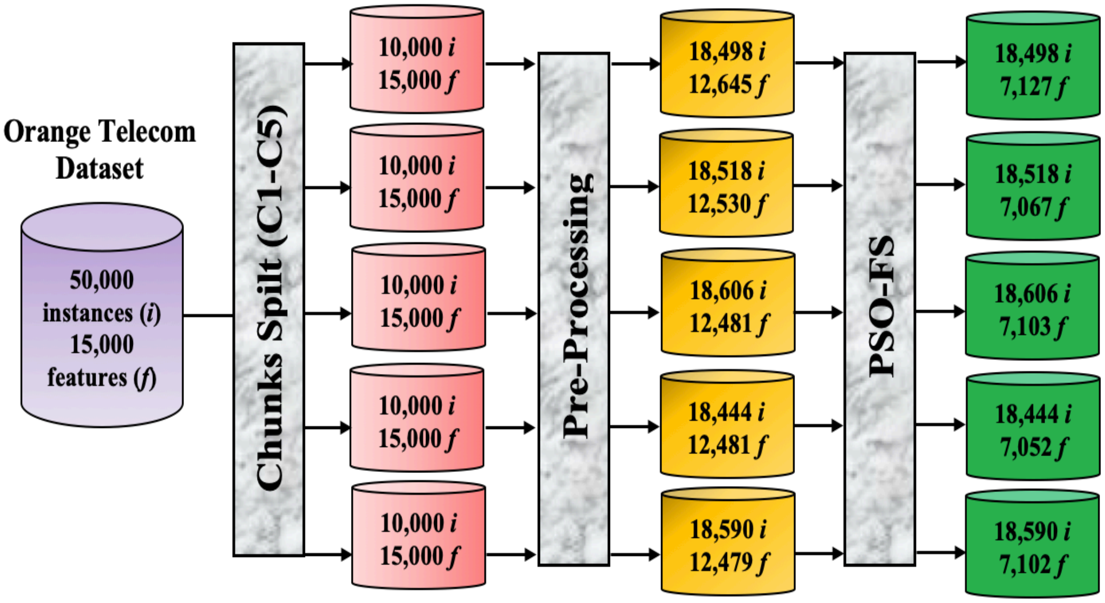


Figure 3. 6 Purified OTD

3.4 Churn Prediction Classification Models

Objective of this study is to utilize the accurate and efficient prediction results of churners using some relevant machine learning techniques, making easy for the service provider to retain their valuable customers. At this stage it is the most difficult task to select the only accurate classifier. We used multiple classifiers by not rely only on single classifier as evaluation results vary from classifier to classifier. Classifiers considered in our study are Naïve Bayes (NB), Logistic Regression (LR), Random Forest (RF) and XG-Boost.

3.4.1 Naïve Bayes

Naïve Bayes (NB) is a probability based model on the bases of observed data. Naïve Bayes determines the probabilities of both the classes $y_i \in \mathbb{C}$ on every single instance with features [$x_1, x_2, x_3, \dots, x_i$] taken by PSO based features selection technique discussed in Section 3.3. Furthermore weights to every instance has been assigned and the resultant will be multiplied with the real value. If the calculated value of probability is higher than 1 then it is said that the class is positive on the other hand if the condition was not fulfilled then it is said that the class is negative.

$$P(y_i|x_1, x_2, x_3, \dots, x_i) = \frac{P(y_i)P(x_1, x_2, x_3, \dots, x_i|y_i)}{P(x_1, x_2, x_3, \dots, x_i)} \qquad (3.2)$$

Performance of NB classifier depends on the linkage of the features and the feature dependency on the class label. In this study, features of orange telecom dataset was dependent on one another and also the large quantity of features are a challenging task towards the performance of Naïve Bayes.

3.4.2 Logistic Regression

Logistic Regression (LR) works when your class label suppose Y consists of binary values such that 0 or 1. Furthermore taking some features p of instance X. Eq. 3.2 shows formulated demonstration of LR.

$$\Pr(Y = 1|X) = p(X) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}} \qquad (3.3)$$

As our study is also a binary class problem that is churner or non-churner hence selection of logistic regression algorithm works in our problem. Performance of churn prediction was improved from Naïve Bayes method.

3.4.3 Random Forest

Random Forest (RF) a well-known ensemble technique in machine learning. Random Forest consists of huge numbers of individual decision trees. The reason of choosing RF is that it is considered a good performance developer classifier. A supervised learning method uses bagging approach for ensemble of DT. Firstly a training set was provided with features X_f = [x_1,x_2,x_3, ... , x_i] with classes $y_i \in \mathbb{C}$ where $\mathbb{C} = \{churner = 1, non - churner = -1\}$, Random Forest develops a number of decision trees with N_i set of features, The class label either churner or non-churner was decided on the bases of majority voting from all the individual trees. Figure 3.7 Presents the generic working of RF ensemble model on purified orange dataset with final predictions.

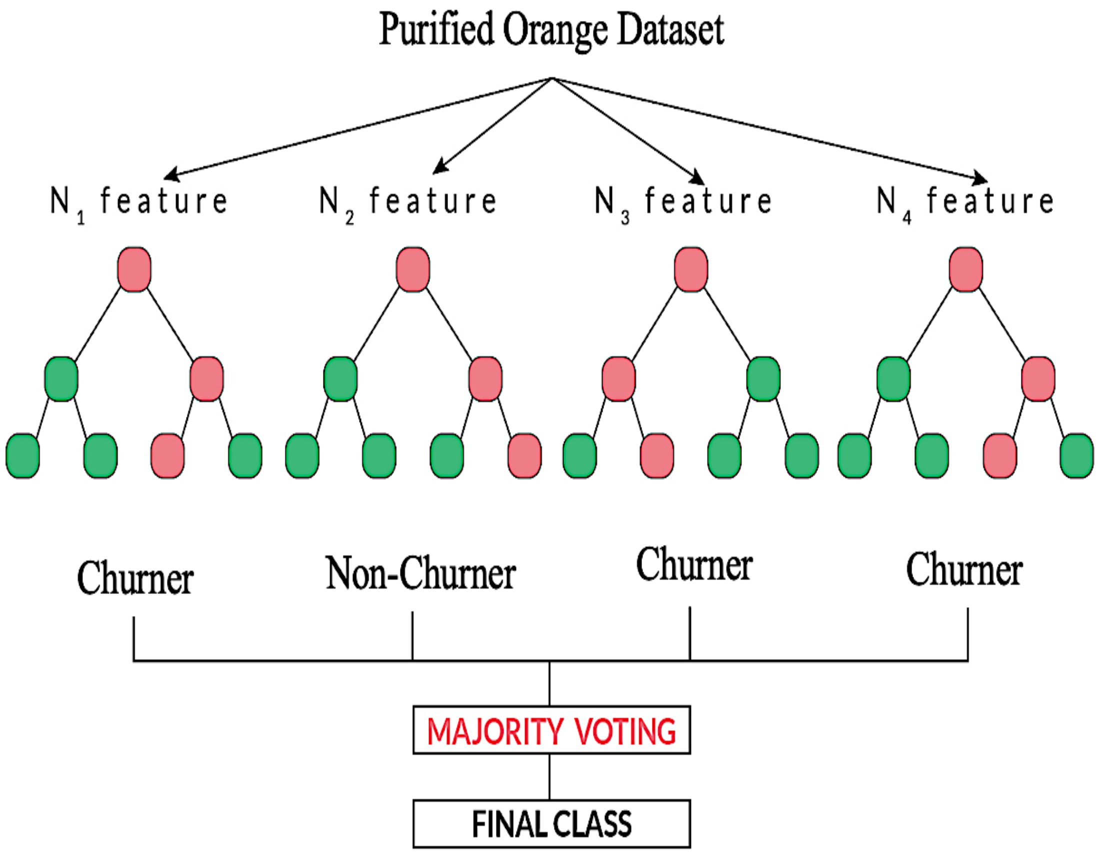


Figure 3. 7Generic overview of RF in Churn Prediction

3.4.4 XG Boost

After implementation on Random Forest it was required to get some results by applying a boosting technique for fast computation, therefore eXtreme Gradient Boosting (XG Boost) algorithm was utilized. XG Boost is also decision-tree-based ensemble algorithm that uses a gradient boosting framework. XG Boost was experimented on the orange telecom dataset to achieve the objective. It is famous due to the speed and performance factors. XG Boosting algorithm only takes numerical values which is a suitable technique to use on orange dataset. The major advantage of XG boost is that it can use multiple cores parallel and fasten the computation by combining the trees results along the way. XG Boost generates tree sequentially and the new generated tree learn from the errors made by the previous tree. Accuracy gained with XG Boost algorithm was better than all the previous methods.

3.5 Summary

Overviewing the proposed methodology chapter it has shown detailed and step by step proceedings of our proposed methodology. Firstly the selection of the dataset according to the problem statement than moving towards the preprocessing stage which was further consists of data cleaning, Normalization, Class imbalance steps. Secondly the feature selection technique namely PSO has been utilized in this work for getting the only meaningful and relevant features. Moreover stratified 5-fold cross validation has been performed along with multiple classification models such as Naïve Bayes (NB), Logistic Regression (LR), Random Forest (RF) and XG-Boost taken into discussion. At the end of this chapter evaluation metrics has been briefly discussed. To get the performance and evaluate our proposed methodology for churn prediction on big data, some experiments have been performed on orange telecom dataset. The next chapter will explain a detailed overview about the Experiments and the results generated on those experiments.

Chapter 4

Experimental Results

In this chapter experimental work on the proposed methodology along with the implementation work has to be discussed on the Big data provided by the telecom company orange. This section was divided into two subsections. In the first section results generated on the classification techniques that are Naïve Bayes (NB), Logistic Regression (LR), Random Forest (RF) and XG-Boost are presented. Moreover the performance evaluation metrics discussed earlier that are Accuracy, Recall or Sensitivity, Specificity, Precision, F-Measure, AUROC visualizes the performance of the proposed work. As it was discussed earlier that orange telecom dataset was a big data problem hence it was divided into five number of chunks, therefore the reported results are first reported as chunk wise. The second part of the study consists of the average results computed of all the chunks. And the comparison analysis between our best model and the previously used models on similar dataset.

4.1 Dataset

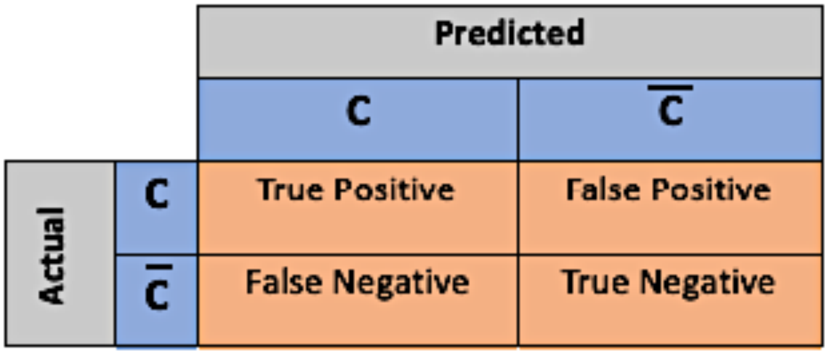
The dataset considered in this thesis is taken from a French telecommunication company Orange, dataset is publicly available at annual Data Mining and Knowledge Discovery competition (KDD Cup 2009) website. Dataset consists of 50,0000 records and was spilt into two datasets one with small and the one with large. Small dataset consists of 230 features whereas large dataset has 15000 features. We used large dataset due to the factor to minimize the features with some feature selection strategy. The orange large telecom dataset consist of large number of missing values, noisy features and data sparsity for this purpose data was taken into preprocessing stage before training on classifier. Section 3.2 shows the preprocessing on the orange telecom dataset.

4.2 Performance evaluation

After identification of the classifier the next step is to get the performance evaluation metrics. How a model perform and how to measure its performance to compare it with other research work. In literature some common performance metrices have been studied such as

Confusion Metrix A table consist of two rows and two columns which contains the number of True Positive (TP), False Positive (FP), False Negative (FN) and True

Negative (TN). It is the basic information required for analyzing churn prediction, Figure 4.1 explains Confusion Metrix diagrammatically .



		Predicted	
		C	$\bar{C}$
Actual	C	True Positive	False Positive
	$\bar{C}$	False Negative	True Negative

Figure 4. 1 Confusion Metrix

C : Churners , $\bar{C}$: Non-Churners

a) Accuracy

Accuracy (ACC) is computed as the sum of all correct predictions (TP and TN) divided by the sum of all prediction. The best accuracy is determined by 1.0 and the worst condition is 0.0. Formula of ACC is given below,

$$\text{ACC} = \frac{\sum(\mathit{Correct\ Predictions})}{\sum(\mathit{All\ Predictions})} \qquad (4.1)$$

b) Recall / Sensitivity

Recall / Sensitivity It is defined as number of true positive prediction divided by the sum of true positive and false Negative prediction,

$$\text{True Positive Rate} \ \ / \text{Recall} = \frac{TP}{TP+FN} \qquad (4.2)$$

c) Specificity

Specificity (also called the true negative rate) measures the proportion of actual negatives that are correctly identified

$$\text{True Negative Rate} \ = \frac{\mathit{Actual\ Negative\ Predictions}}{TN+FP} \qquad (4.3)$$

d) Precision

Precision can be calculated by the total number of true positive predictions divided by the sum of the total positive predictions with total false positive predictions formulated as follows.

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (4.4)$$

e) F1 Measure

F1-Measure can be calculated by knowing the values of precision and recall. Formulated as follows

$$\text{F1} - \text{Measure} = \frac{2 * \text{ Precision} * \text{ Recall}}{\text{Precision} + \text{Recall}} \qquad (4.5)$$

f) Area Under the Receiver Operating Characteristic curve (AUROC)

Researchers also used the standard scientific accuracy indicator Area Under the Receiver Operating Characteristic curve (AUROC) to evaluate the test data,

$$\text{AUROC} = \frac{\sum_{n \in true\ churners} Rank_n - \frac{P \text{ x } (\text{P}+1)}{2}}{P \text{ x N}} \qquad (4.6)$$

where P represent the number true churners and N shows the number of true non-churners. Arranging the churners in descending order, rank n is assigned to the highest probability customer, the next with rank n-1 and so on. Moving upward from base line in AUROC get the better performance. This is because of imbalance data of churner and non-churner AUROC is better metric for performance measure.

4.3 Experimental Setup

To get the accurate and efficient results on multiple classifiers, it was necessary to provide same environment and same kind of dataset stated in section 3.1. as the original dataset was not well organized and contained a lot of missing values, hence more computation work was required to gain the objective. Therefore implementation was conducted on MacBook Pro consists of 7 cores and 16 GB of memory. Python 3.7.3 was used for coding purpose on intelliJ idea tool. All the implementation was done chunk wise starting from the pre-processing stage till the evaluation stage.

4.4 Results

After obtaining purified and best features selection orange telecom dataset was passed, to obtain most relevant model for the given problem we applied multiple classifiers to identify which model outperforms we used graph visualization. Moreover implementation results are demonstrated on the bases of classifiers used.

4.4.1 Naïve Bayes

Naïve Bayes was implemented on the Orange Telecom Dataset. Dataset was first divided into five chunks to minimize the overload of computation by taking the whole dataset at once. Performance metrices used to predict the efficiency of the model chunk wise are Accuracy, Recall , Precision, F-Measure whose results can be visualized in Figure 4.2. It can be seen that chunk 4 shows high accuracy rate of 74.97%.

4.4.2 Logistic Regression

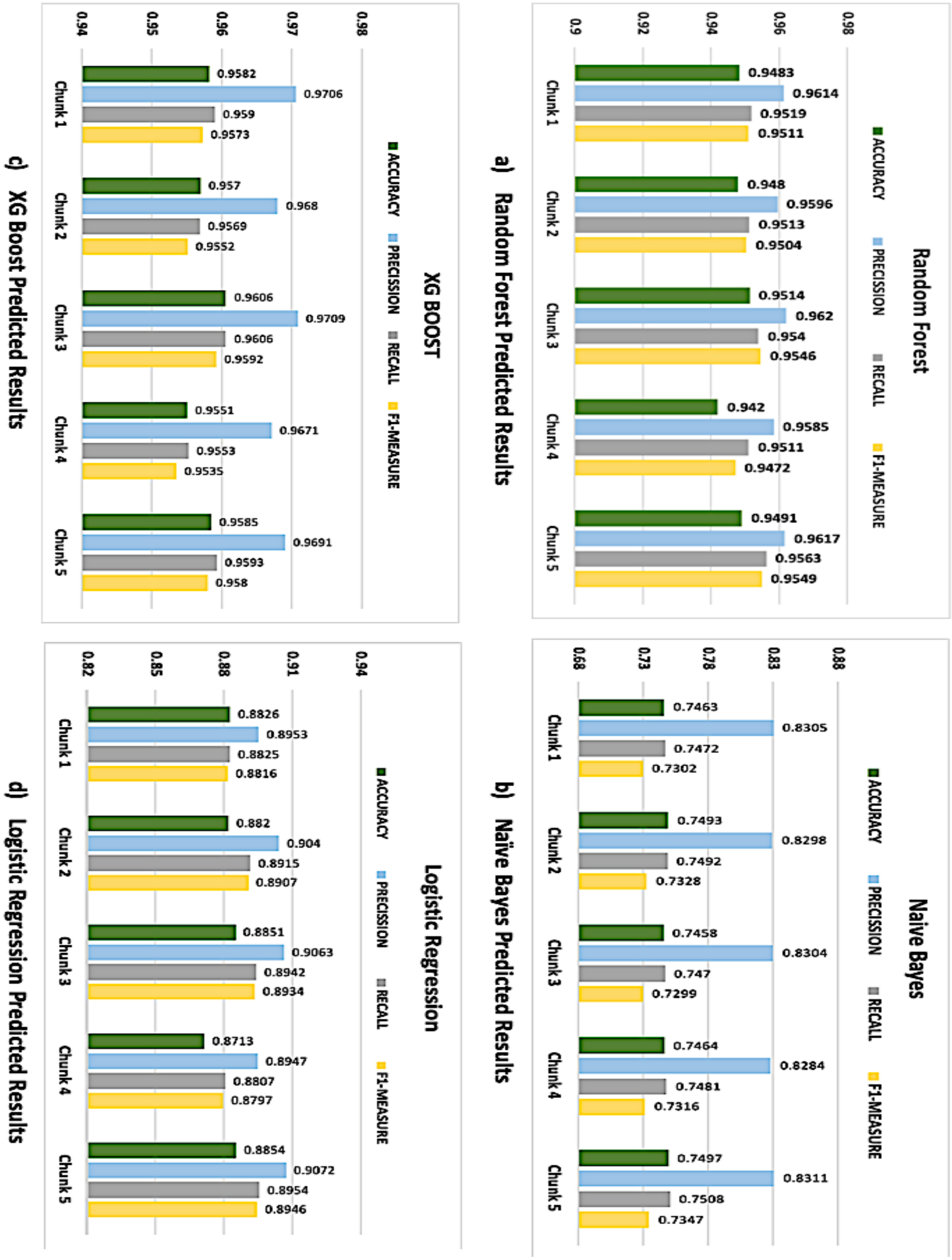
Logistic regression was implemented on the Orange Telecom Dataset due to the binary classification problem. The results shown in Figure 4.2 was not satisfactory so Logistic Regression was the second classifier taken into consideration. Similarly the same situation was taken for this model and same environment was been provided to take some better results. Performance metrices used to predict the efficiency of the model chunk wise are Accuracy, Recall , Precision, F-Measure whose results can be visualized in Figure 4.2. It can be seen that chunk 5 shows high accuracy rate of 88.54%.

4.4.3 Random Forest

Random Forest was experimented for prediction of churners on purified and structured Orange telecom dataset. Chunks data generated after selection of meaningful features through PSO was then delivered to Random Forest classifier for getting much better results than both the NB and LR algorithms. Random Forest proved by showing a major positive deviation from the previous two techniques discussed. Figure 4.2 describes the actual results generated after testing random forest on the similar dataset. Chunk 3 leads in accuracy race then other chunks with 95.14% reported result.

4.4.4 XG BOOST

Finally the last testing taken was on common boosting algorithm known as XG Boost. Similarly all the environment provided to the last three classifier was same for XG Boost. Boosting algorithm gave some excellent performance in churn prediction scenario by reporting good results as shown in Figure 4.2. In the race of accuracy between five chunks it can be noticed in Figure 4.2 that chunk 3 gave accuracy rate of 96.06%. Overall performance of XG Boost algorithm was the most efficient between the competitor taken in our study.



a) Random Forest Predicted Results

b) Naïve Bayes Predicted Results

c) XG Boost Predicted Results

d) Logistic Regression Predicted Results

Figure 4. 2 Classifiers Performance Evaluation on split OTD (Accuracy, Precision, Recall, F1-Measure)

4.5 Results Evaluation on ROC

In the last section, it can be noticed that different evaluation metrices has been taken. To show our results we used Receiver Operating Characteristic (ROC) curve an advanced technique to visualize results better. ROC is a curve representation of the proportion of True Positive Rate (TPR) on vertical axis and False Positive Rate (FPR) on Horizontal axis. Performance of classifier can be determined as more accurate if the result is close to 1 and less accurate if it is below 0.5. Stratified 5-fold cross validation was used on each of the chunk with reporting results of 5-folds and then mean of folds has also computed on the classifiers discussed in section 3.4.

In Figure 4.3 an overall view of ROC curve on multiple classifier has been visualized on split orange dataset along with mean AUC values. Step into Figure 4.3 Naïve Bayes classifier showed a mean AUC value of 0.75 after 5-fold cross validation similar for each chunk. To obtain more accurate results we then implemented Logistic Regression classifier and the results shown some good output. Similarly like the previous technique the environment was kept the same and the mean AUC value predicted was 0.93 on C1,C2 and C4 and mean AUC obtained on C3 and C5 was 0.94. Moreover to predict churner more accurate we also used an Ensemble technique namely Random Forest. ROC curve showed a much more required results with a mean AUC of 0.962 on each of chunk. In the game of obtaining more accurate results we used an boosting technique on the same split orange dataset. XG Boost was taken and the ROC curve gave the most accurate and efficient results then all the previous techniques, showing mean AUC value of 0.98 on C1,C2,C3,C4,C5 respectively.

Comparatively ROC curve results showed that Ensemble technique outperformed on the Orange dataset such that Random Forests and XG Boost leads on such on such type of datasets. Figure 4.4 demonstrate the graphical representation of mean AUC results reported by the classifiers used in our research.

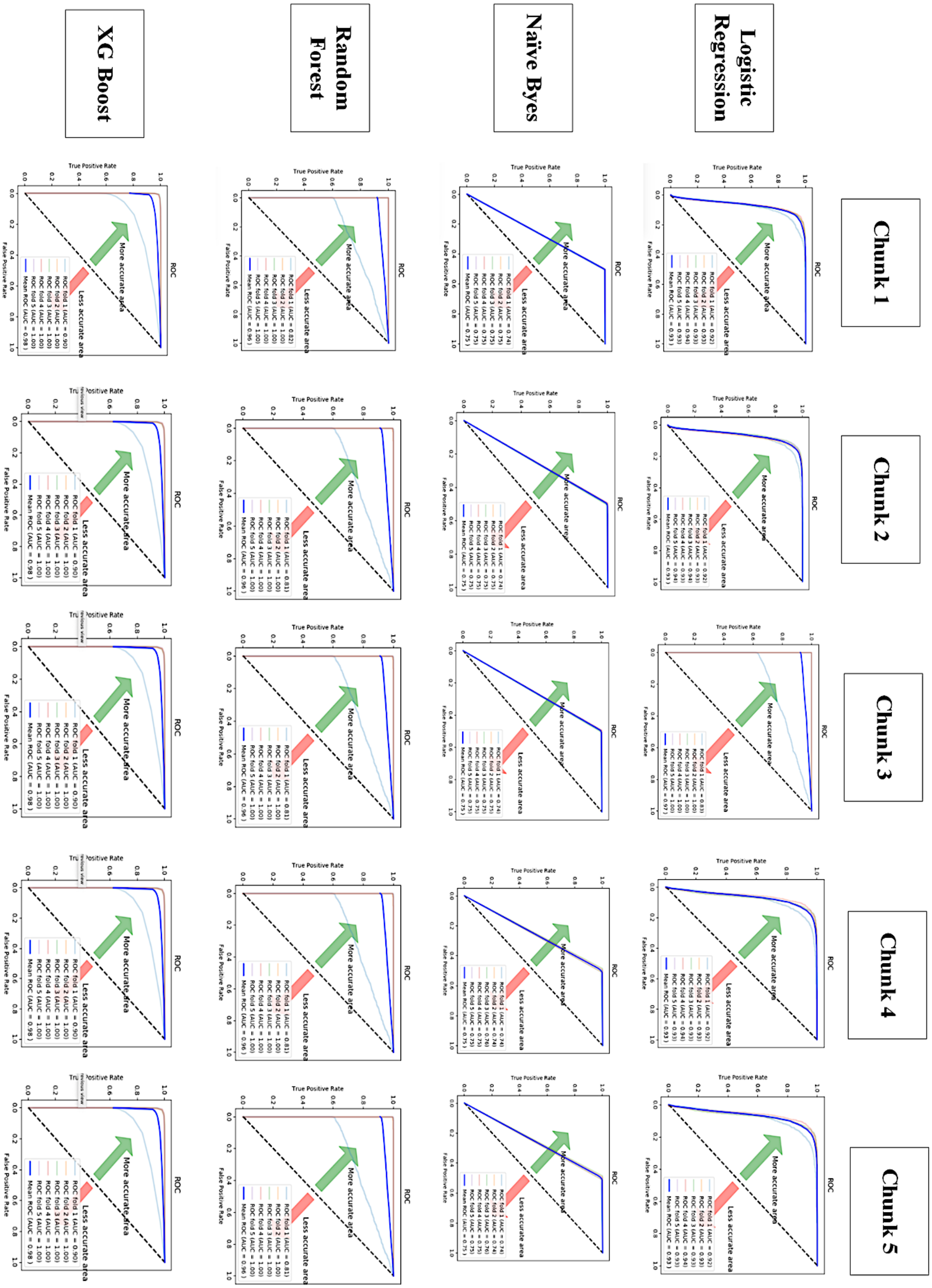


Figure 4. 3 Split OTD area under ROC curve visualization with multiple classifiers

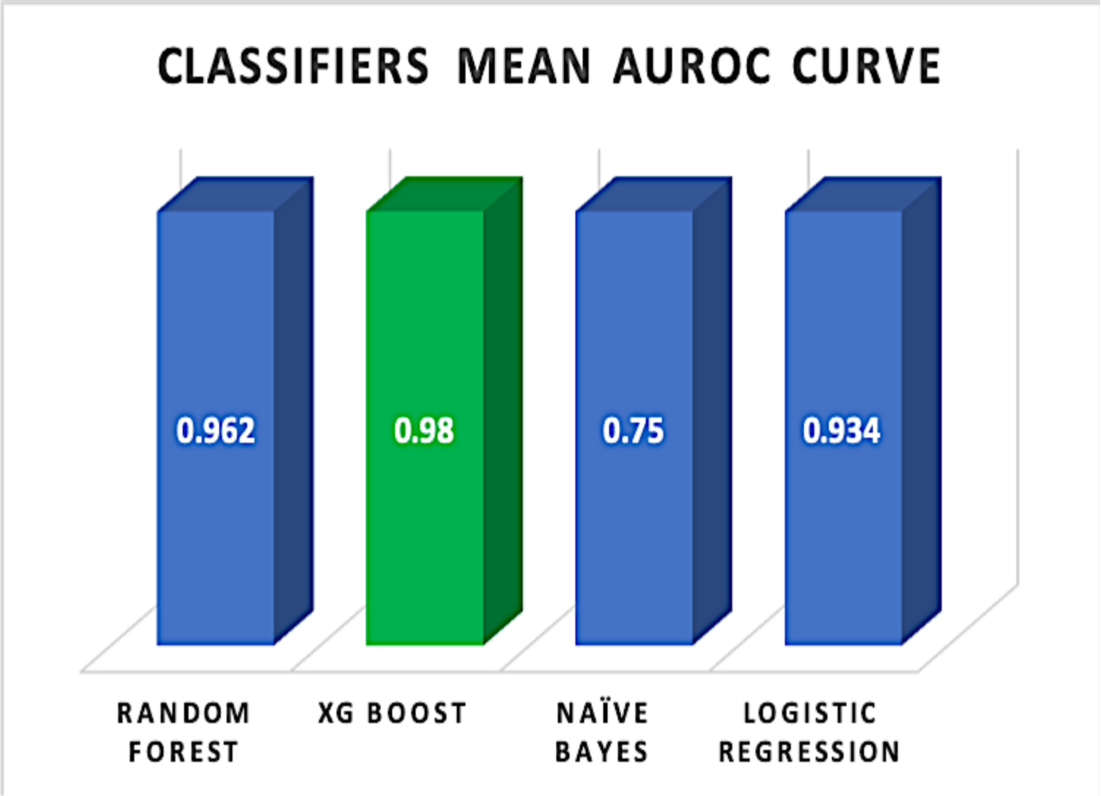


Figure 4. 4 Multiple classifiers mean AUROC curve

4.6 Comparison with existing approaches

A numerous approaches have been applied with different classifiers in the domain of churn prediction. The behavior of the classifiers relay on the dataset and the techniques that have been applied during preprocessing stage. Comparison can be divided into two parts the first one with the similar dataset used in our research and the second one with other telecom datasets used by different researchers. Comparison was taken on the bases of performance evaluation metrices have been taken commonly by the researchers discussed in Section 4.2.

Table 4.1 shows a result comparison between different methodologies used by researchers and results of our best model specifically on the orange telecom dataset (OTD). It was examined that XG Boost performed best by predicting efficient evaluation results on every considered metrics.

Table 4. 1Comparative analysis of XG Boost on OTD

Methodology	Accuracy	Precision	Recall	F1 Measure	ROC / AUC
DT, K-Star, NB, J48 **[42]**	89%	-	-	-	-
Hybrid Firefly **[43]**	86.38%	-	-	0.855	-
GP + ADA Boost **[44]**	-	-	0.835	-	0.86
Rotation Boost **[45]**	-	-	0.729	0.731	0.761
PSO-FSSA **[35]**	90.65%	95.98	92.92	94.43	-
Our Model (XG Boost) **Average Results of all Chunks**	**95.78%**	**96.91**	**95.82**	**95.66**	**0.98**

Table 4.2 demonstrate the comparison between the performance evaluation results taken on different datasets with the results taken by XG Boost used in our study related to churn prediction in telecom industry.

Table 4. 2 XG Boost Performance comparison with telecom sector datasets

Dataset	Methodology	Accuracy	Precision	Recall	F1 Measure	ROC
Churn-Bigml **[40]**	RF	-	0.891	0.896	0.876	0.835
South Asia GSM **[40]**	RF	-	0.893	0.888	0.882	0.947
Prepaid China **[14]**	RF	-	0.959	0.227	-	0.932
Churn-Bigml **[41]**	RF	91.66	0.831	0.988	0.952	-
Our Model (OTD)	**XG Boost**	**0.957**	**0.969**	**0.958**	**0.956**	**0.98**

Chapter 5

Conclusion & Future Research Directions

In this proposed thesis, dataset was taken of the famous French telecom company Orange. This work main focus was on the customer churn prediction in telecom industry. Hence Orange telecom company required to predict the customers which will be not be company users in the coming future. The obtained dataset was a large dataset which needs a heavy systems on which processing of data can be done. To resolve that issue dataset was divided into five chunks with equal number of instances in each. Furthermore it is a big problem in the large datasets that it has some garbage and null values in it, so pre-processing stage was the initial step in this journey.

Preprocessing stage was consisted of three major phases which resolved three different kind of problems in the orange dataset. Firstly data cleaning phase resolved the issue of elimination of empty columns and garbage data columns along with the replacement of values in the null cells with the mean of that particular column in which the null cell exists. Secondly the Normalization phase the orange dataset consists of out of measurement values it must have been to make all the dataset into a range of 0 and 1. Min-Max scalar normalization has done on the dataset to make it in some meaningful range. Thirdly it was a real fact that there are less customers in any company that may churn and the ratio of loyal customers vs the churners will be highly deviate. Due to which a class imbalance problem can be faced. Similarly in the orange telecom dataset it was examined that number of churners are only 7.34% and non-churners are 92.66% which proved that class imbalance exists here. Therefore SMOTE was used for making new instances only from the minor class (churner) to resolve the class imbalance problem. It was noticed that the performance was improved after preprocessing the raw orange telecom dataset.

A structured dataset was taken after the preprocessing stage. Still large data problem exists due to large number of features and needs heavy system requirement for computation. A feature selection phase was considered due to the factor that there are some features that are not relevant and had no impact on the final result or had a bad impact on the final result, hence those features should be identified and eliminate by the system. A PSO based Feature selection technique was used and the output result was obtained with the only global best features that had impact on the prediction. Moreover after following the pre-processing and feature selection steps we shrink the data by eliminating almost half of the useless features. Furthermore it was examined that gaining the meaningful features had a great impact on predicting the accurate churners and gave a better performance than .

After working on the multiple phases of implementation in this thesis. The last stage arrived in which the most commonly and advanced classifiers have been utilized to predict the performance evaluation on orange telecom dataset. Multiple classifiers have been tested passing through with stratified 5-fold cross validation procedure produced some good experimental results. The purified dataset was provided to the multiple classifiers for predicting results. Hence leading technique was XG Boost lies in ensemble category reported mean AUC results of 0.98. Orange dataset was numerical dataset and for fast computation we experimented a famous XG Boost algorithm a gradient boosting technique use the technology of parallel processing. The reported results gave the extra ordinary performance results. Whereas we used another ensemble technique Random Forest based on bagging algorithm gave minor difference from the Boosting technique with reported performance results on mean AUC 0.962. Two more classifiers have been experimented on the same environment and on same orange dataset for churners prediction. Logistic Regression and Naïve Bayes gave mean AUC results of 0.93 and 0.75 respectively. Obtained the objective of accurate and high rate prediction of churners with a simple model design.

For the future work we have planned to automate the retention mechanism on these prediction for the ease and now a days requirement of the telecom company. Furthermore to do more experiments on with increasing number of fold to 10-fold to compute some more accurate results.

Chapter 6

References

[1] T. J. Gerpott, W. Rams, and A. Schindler, "Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market," *Telecommunications policy,* vol. 25, no. 4, pp. 249-269, 2001.

[2] E. Ascarza, R. Iyengar, and M. Schleicher, "The perils of proactive churn prevention using plan recommendations: Evidence from a field experiment," *Journal of Marketing Research,* vol. 53, no. 1, pp. 46-60, 2016.

[3] S. Longo, "The Cost of Customer Acquisition vs Customer Retention," *Kapost Blog. URL https://marketeer. kapost. com/customer-acquisition-versus-customer-retention/(accessed 3.16. 19),* 2016.

[4] Ö. G. Ali and U. Arıtürk, "Dynamic churn prediction framework with more effective use of rare event data: The case of private banking," *Expert Systems with Applications,* vol. 41, no. 17, pp. 7889-7903, 2014.

[5] M. Azeem and M. Usman, "A fuzzy based churn prediction and retention model for prepaid customers in telecom industry," *International Journal of Computational Intelligence Systems,* vol. 11, no. 1, pp. 66-78, 2018.

[6] M. Azeem, M. Usman, and A. C. M. Fong, "A churn prediction model for prepaid customers in telecom using fuzzy classifiers," *Telecommunication Systems,* vol. 66, no. 4, pp. 603-614, 2017.

[7] L. Katelaris and M. Themistocleous, "Predicting Customer Churn: Customer Behavior Forecasting for Subscription-Based Organizations," in *European, Mediterranean, and Middle Eastern Conference on Information Systems*, 2017: Springer, pp. 128-135.

[8] S. Babu and N. Ananthanarayanan, "Enhanced Prediction Model for Customer Churn in Telecommunication Using EMOTE," in *International Conference on Intelligent Computing and Applications*, 2018: Springer, pp. 465-475.

[9] R. Dong, F. Su, S. Yang, X. Cheng, and W. Chen, "Customer Churn Analysis for Telecom Operators Based on SVM," in *International Conference On Signal And Information Processing, Networking And Computers*, 2017: Springer, pp. 327-333.

[10] S. Maldonado and C. Montecinos, "Robust classification of imbalanced data using one-class and two-class SVM-based multiclassifiers," *Intelligent Data Analysis,* vol. 18, no. 1, pp. 95-112, 2014.

[11] S. Maldonado, Á. Flores, T. Verbraken, B. Baesens, and R. Weber, "Profit-based feature selection using support vector machines–General framework and an application for customer retention," *Applied Soft Computing,* vol. 35, pp. 740-748, 2015.

[12] M. Óskarsdóttir, C. Bravo, W. Verbeke, C. Sarraute, B. Baesens, and J. Vanthienen, "Social network analytics for churn prediction in telco: Model building, evaluation and network architecture," *Expert Systems with Applications,* vol. 85, pp. 204-220, 2017.

[13] S. D'Alessandro, L. Johnson, D. Gray, and L. Carter, "Consumer satisfaction versus churn in the case of upgrades of 3G to 4G cell networks," *Marketing Letters,* vol. 26, no. 4, pp. 489-500, 2015.

[14] Y. Huang *et al.*, "Telco churn prediction with big data," in *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, 2015: ACM, pp. 607-618.

[15] S.-Y. Hung, D. C. Yen, and H.-Y. Wang, "Applying data mining to telecom churn management," *Expert Systems with Applications,* vol. 31, no. 3, pp. 515-524, 2006.

[16] T. A. Park and J. Sauer, "Evaluating food retailers using dual elasticities of substitution," *Journal of Productivity Analysis,* vol. 39, no. 2, pp. 111-122, 2013.

[17] S. Wood and D. McCarthy, "The UK food retail 'race for space'and market saturation: A contemporary review," *The international review of retail, distribution and consumer research,* vol. 24, no. 2, pp. 121-144, 2014.

[18] A. Colovic and U. Mayrhofer, "Optimizing the location of R&D and production activities: trends in the automotive industry," *European Planning Studies,* vol. 19, no. 8, pp. 1481-1498, 2011.

[19] W. C. Lucato, M. V. Júnior, R. M. Vanalle, and J. A. A. Salles, "Model to measure the degree of competitiveness for auto parts manufacturing companies," *International Journal of Production Research,* vol. 50, no. 19, pp. 5508-5522, 2012.

[20] A. Mittal, P. Mitra Mukherjee, and D. Roy, "Global Competitiveness: World Passenger Car Industry," *SCMS Journal of Indian Management,* vol. 10, no. 4, 2013.

[21] W. Yang *et al.*, "Mining Player In-game Time Spending Regularity for Churn Prediction in Free Online Games," in *2019 IEEE Conference on Games (CoG)*, 2019: IEEE, pp. 1-8.

[22] J. Kim, "Manufacturing or service? Market saturation and cycles of over-investment as a clue to future of service economies," *Technological Forecasting and Social Change,* vol. 78, no. 8, pp. 1345-1355, 2011.

[23] E. Ascarza *et al.*, "In pursuit of enhanced customer retention management: Review, key issues, and future directions," *Customer Needs and Solutions,* vol. 5, no. 1-2, pp. 65-81, 2018.

[24] M. Fraering and M. S. Minor, "Beyond loyalty: customer satisfaction, loyalty, and fortitude," *Journal of Services Marketing,* vol. 27, no. 4, pp. 334-344, 2013.

[25] A. Drachen *et al.*, "Rapid prediction of player retention in free-to-play mobile games," in *Twelfth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2016.

[26] Q. Wang, "Design the Churn Analysis on Games A Review on Techniques for Churn Analysis," Northeastern University, 2018.

[27] R. Boohene, G. K. Agyapong, and E. Gonu, "Factors influencing the retention of customers of Ghana Commercial Bank within the Agona Swedru Municipality," *International Journal of Marketing Studies,* vol. 5, no. 4, p. 82, 2013.

[28] R. Rajamohamed and J. Manokaran, "Improved credit card churn prediction based on rough clustering and supervised learning techniques," *Cluster Computing,* vol. 21, no. 1, pp. 65-77, 2018.

[29] E. Sivasankar and J. Vijaya, "Customer Segmentation by Various Clustering Approaches and Building an Effective Hybrid Learning System on Churn Prediction Dataset," in *Computational Intelligence in Data Mining*: Springer, 2017, pp. 181-191.

[30] A. De Caigny, K. Coussement, and K. W. De Bock, "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees," *European Journal of Operational Research,* vol. 269, no. 2, pp. 760-772, 2018.

[31] Y. Huang and T. Kechadi, "An effective hybrid learning system for telecommunication churn prediction," *Expert Systems with Applications,* vol. 40, no. 14, pp. 5635-5647, 2013.

[32] A. Idris, M. Rizwan, and A. Khan, "Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies," *Computers & Electrical Engineering,* vol. 38, no. 6, pp. 1808-1819, 2012.

[33] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimization for feature selection in classification: A multi-objective approach," *IEEE transactions on cybernetics,* vol. 43, no. 6, pp. 1656-1671, 2012.

[34] L. Brezočnik, "Feature selection for classification using particle swarm optimization," in *IEEE EUROCON 2017-17th International Conference on Smart Technologies*, 2017: IEEE, pp. 966-971.

[35] J. Vijaya and E. Sivasankar, "An efficient system for customer churn prediction through particle swarm optimization based feature selection model with simulated annealing," *Cluster Computing,* vol. 22, no. 5, pp. 10757-10768, 2019.

[36] A. Sharma, D. Panigrahi, and P. Kumar, "A neural network based approach for predicting customer churn in cellular network services," *arXiv preprint arXiv:1309.3945,* 2013.

[37] Z.-Y. Chen, Z.-P. Fan, and M. Sun, "A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data," *European Journal of operational research,* vol. 223, no. 2, pp. 461-472, 2012.

[38] A. Karahoca and D. Karahoca, "GSM churn management by using fuzzy c-means clustering and adaptive neuro fuzzy inference system," *Expert Systems with Applications,* vol. 38, no. 3, pp. 1814-1822, 2011.

[39] A. Idris, A. Khan, and Y. S. Lee, "Genetic programming and adaboosting based churn prediction for telecom," in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2012: IEEE, pp. 1328-1332.

[40] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," *IEEE Access,* vol. 7, pp. 60134-60149, 2019.

[41] A. Mishra and U. S. Reddy, "A comparative study of customer churn prediction in telecom industry using ensemble based classifiers," in *2017 International Conference on Inventive Computing and Informatics (ICICI)*, 2017: IEEE, pp. 721-725.

[42] S. M. Sladojevic, D. R. Culibrk, and V. S. Crnojevic, "Predicting the churn of telecommunication service users using open source data mining tools," in *2011 10th International Conference on Telecommunication in Modern Satellite Cable and Broadcasting Services (TELSIKS)*, 2011, vol. 2: IEEE, pp. 749-752.

[43] A. A. Ahmed and D. Maheswari, "Churn prediction on huge telecom data using hybrid firefly based classification," *Egyptian Informatics Journal,* vol. 18, no. 3, pp. 215-220, 2017.

[44] A. Idris, A. Iftikhar, and Z. ur Rehman, "Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO undersampling," *Cluster Computing,* vol. 22, no. 3, pp. 7241-7255, 2019.

[45] A. Idris, A. Khan, and Y. S. Lee, "Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification," *Applied intelligence,* vol. 39, no. 3, pp. 659-672, 2013.

[46] B. Zhu, B. Baesens, and S. K. vanden Broucke, "An empirical comparison of techniques for the class imbalance problem in churn prediction," *Information sciences,* vol. 408, pp. 84-99, 2017.

[47] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research,* vol. 16, pp. 321-357, 2002.

[48] S.-J. Yen and Y.-S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," *Expert Systems with Applications,* vol. 36, no. 3, pp. 5718-5727, 2009.

[49] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," *Expert Systems with Applications,* vol. 36, no. 3, pp. 4626-4636, 2009.

[50] A. Amin *et al.*, "Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study," *IEEE Access,* vol. 4, pp. 7940-7957, 2016.

[51] A. Amin, F. Rahim, I. Ali, C. Khan, and S. Anwar, "A comparison of two oversampling techniques (smote vs mtdf) for handling class imbalance problem: A case study of customer churn prediction," in *New Contributions in Information Systems and Technologies*: Springer, 2015, pp. 215-225.

[52] K. M. Ting, "An instance-weighting method to induce cost-sensitive trees," *IEEE Transactions on Knowledge & Data Engineering,* no. 3, pp. 659-665, 2002.

[53] H.-x. WU, Y. WU, Z.-t. LIU, and Z. LEI, "Combined SNP feature selection based on relief and SVM-RFE," *Application Research of Computers,* vol. 6, 2012.

[54] Y. Chen, D. Miao, and R. Wang, "A rough set approach to feature selection based on ant colony optimization," *Pattern Recognition Letters,* vol. 31, no. 3, pp. 226-233, 2010.

[55] N. Suguna and K. Thanushkodi, "A novel rough set reduct algorithm for medical domain based on bee colony optimization," *arXiv preprint arXiv:1006.4540,* 2010.

[56] B. Q. Huang, T.-M. Kechadi, B. Buckley, G. Kiernan, E. Keogh, and T. Rashid, "A new feature set with new window techniques for customer churn prediction in land-line telecommunications," *Expert Systems with Applications,* vol. 37, no. 5, pp. 3657-3665, 2010.