A Machine Learning Approach for Churn Prediction in Telecommunication

A Thesis
submitted in partial fulfillment of the requirements for the award of the degree of

Master of Engineering

in

Computer Science and Engineering

Submitted By

Kriti Mishra
(Roll No: 801532027)

Under the supervision of

Dr. Rinkle Rani
Associate Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA – 147004
July 2017

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "A Machine Learning Approach for Churn Prediction in Telecommunication", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Rinkle Rani* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Kriti Mishra

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

14/7/17

Dr. Rinkle Rani
Associate Professor
Computer Science and Engineering Department

Acknowledgement

I wish to express my earnest gratitude to my supervisor Dr. Rinkle Rani for her invaluable advice and encouragement at every step of my ME program. Without her unfailing support and belief in me this thesis would not have been possible. I am truly grateful for this. Their contribution to this thesis goes well beyond their role as an academic supervisor and includes constant support on a personal level without which this journey may never have been completed and for this I am truly grateful.

I would also like to express my sincere and deep gratitude towards my friends and family members for extending their kind support, encouragement and belief in me throughout my work. They have been a source of motivation for me.

Kriti Mishra

Abstract

The telecommunication industry always has a tough competition with its competitors to retain customers, and therefore has become one of the research sectors in machine learning and data mining. Since the customers' churn behavior is to be monitored closely and efficiently it requires for a methodical churn prediction model to monitor the customers' churn. The main setbacks in achieving the desired performances in a classifier are the enormous datasets, large feature space and imbalanced class distribution. In this work, we explore the implication of Synthetic Minority Over-sampling TEchnique (SMOTE) to reduce the imbalance in data in collaboration with different feature reduction techniques such as Co-relation feature extraction, Gain ratio, Information gain and OneR feature evaluation method. Classification and Regression Trees (CART), Bagged CART and Partial Decision Trees (PART) classifiers are trained to analyze the performance on balanced and reduced feature space dataset. Prediction performance of the classifiers is evaluated through measures such as Area Under the Curve (AUC), sensitivity and specificity. Finally, it is concluded through simulations that our proposed method based on SMOTE, co-relation, and ensemble approach performs well for predicting churners as against simply applying learners on the unrefined dataset. Therefore, this methodology can be helpful for the telecommunication industry to predict churn.

Keywords—Data Mining, Machine Learning, Churn Prediction, Data balancing, Decision tree, Ensemble

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ANN	Artificial Neural Network
ANOVA	Analysis of Variance
AUC	Area Under the Curve
BPN	Back Propagation Neural Network
CART	Classification and Regression Trees
CRM	Customer Relationship Management
DSS	Decision Support System
DT	Decision Tree
ID3	Iterative Dichotomizer 3
IG	Information Gain
KNN	K-Nearest Neighbor
LASSO	Least Absolute Shrinkage and Selector Operator
LDA	Least Discriminant Analysis
LR	Logistic Regression
OLTP	Online Transaction Processing
PART	Projective Adaptive Resonance Theory
PCA	Principal Component Analysis
PSO	Particle Swarm Optimization
RF	Random Forest
RIPPER	Repeated Incremental Pruning to Produce Error Reduction
RUS	Random Under Sampling
ROS	Random Over Sampling
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine
SMOTE	Synthetic Minority Over sampling Technique
SOM	Self- Organizing Maps

Chapter 1

Introduction

Acquisition and retention of customers are two major challenges for businesses. While emerging companies try to acquire new clients, the mature ones focus on retention of existing ones, the objective being providing them with the opportunity of cross-selling. Keeping customers for a longer periods is a significant way of increasing customers' value (Freeman, 1999).

The new era of electronic commerce has led to boost in the information available. According to [1], customers have been empowered by the internet, as they are no longer attached to one company and this has led to aggravated competition. This implies that the attrition rate of clients of a company would grow due to customer empowerment as the competitors are only a "click away". Due to this threat that companies have been facing, they need to be armed and equipped with effective and efficient methods of examining clients' behavior predicting their possible loyalty switching.

This study is aimed at minimizing customer churn in the telecommunication sector, particularly mobile telephony market by finding an efficient and effective predictive model for by utilizing machine learning techniques.

1.1 What is Data Mining?

Progress and innovation is no longer hindered by the ability to collect data. Data generation and collection is not a challenge these days. The challenge is to acquire, analyze, compile, summarize, visualize and discover knowledge from the data under question in a scalable and timely manner.

Data mining is a combination of techniques from machine learning [1] and statistics [2] to dig out previously unknown, useful, novel and valid patterns in data warehouses. The patterns should be applicative in order to be able to use them in an organization's decision making. Data mining is widely used across enterprises to handle large amounts

of information with the help of advanced analytical tools and techniques. The techniques may help in finding novel patterns in data that may further help organizations in decision making.

Data mining has a capability to help companies explore and discover the most useful and relevant information in their data warehouses that can be converted into knowledge to produce knowledge patterns. Data mining technologies have tools and statistical techniques that can predict future trends and have the potential to make dedicated, knowledge-driven decisions based on past behaviors and patterns. The retrospective tools provide automated analysis for the given information. Data mining tools solve questions related to business operations that are too tedious to answer with traditional methods. They thoroughly search databases for patterns and may come out with predictions that may be entirely unexpected even for experts to estimate. Most companies collect large amounts of data from their operational sources and integrate them to achieve a comprehensive view of the data.

The cycle shown in figure 1.1 gives an overview of the process of exploration and business discovery followed in business organizations.

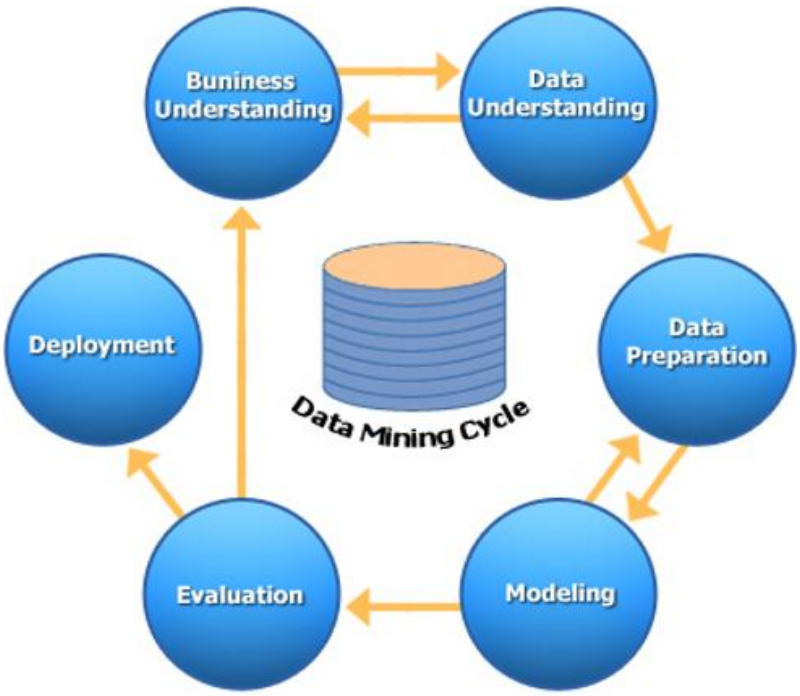

Figure 1.1: A typical Business Cycle using Data Mining

1.2 Why Data Mining?

There are many reasons for the recent growth in data mining. Some of these are:

- Explosion in data

  As is previously noted, there has been an enormous growth in data in the past few decades. Data is being generated from transactions [3], remote devices, websites and user generated content. Growing use of mobile phones and credit cards is one of the major sources of data growth. E-commerce developments have also resulted in spurt of data as visitors to websites are recorded which can be a source of important information to some enterprises.
- Low processing cost

  During the past few decades, the reduction in the cost of processing of data and the increase in power of processing of data by the digital technologies has been remarkable. Inexpensive hardware makes it affordable in terms of cost and time to attempt intensive analysis on data.
- Growth in data storage capacity

Modern technologies have been able to produce disks that can store massive amounts of data.

- Competitive businesses

  Due to the increasing globalization, businesses have become increasingly competitive. There is fierce competition in all business areas, for example in banking and finance industry [7], retail industry [8], telecommunication industry [9], healthcare industry [10], e-commerce, etc. These industries create new knowledge and need to transfer it across various levels of the organization in order to work effectively. As knowledge is prime component in formulating policies and strategies and their implementation, knowledge management is chief task for industry experts to grow and succeed in today's complicated business environment.

- Availability of data mining software

  Many companies in the past have come up with data mining software and are still coming out with new products. Companies like SAS and SPSS have developed statistical tools. Many other software packages like GhostMiner and Quadstone are available.

1.3 The Data Mining Process

For data mining to be effective, several steps are needed to be taken before the actual exploration task. These involve selection, cleaning, transformation and separate storage of data appropriate for data mining. This separate storage is known as data warehouse [13] that stores data needed by an enterprise's decision makers. Data warehouse consists of not just current but also historical data.

A typical data mining process includes the following six steps based on a conventional software approach.

Step 1: Requirements Analysis

In the first step, decision makers formulate goals that they intend to achieve through the data mining process. It includes clearly defining the business problem. It is not a good

idea to proceed without a clear objective. If objectives are well-defined, it becomes easier to evaluate results of a project. A document describing the result of the analysis is produced.

Step 2: Data Collection and Selection

This step comprises of finding the most appropriate data sources for the process. A data warehouse, if implemented may be used for this purpose as it contains most of the data. If not, then source OLTP [14] systems need to be identified and relevant information collected and stored in a reliable provisional storage. A document is produced which gives details of the data that is to be used.

Step 3: Cleaning and Preparing Data

As noted earlier, a data store may be created to integrate data from all sources. While integrating data, we may come across problems such as data conflicts, identifying data, ambiguity, and missing data [16]. An ETL tool is used to assist with these problems. The cleaning and data preparation carried put is documented in a report.

Step 4: Data Mining Exploration and Validation

In this step, the user starts with the exploration of tools and techniques in data mining and constructs models based upon the enterprise's needs. They may take sample data and apply a number of techniques. Afterwards, results are evaluated to interpret and decode their meaning and significance. This is an iterative process which leads to selection of some techniques and discarding of some other unhelpful techniques. Some techniques are considered for further exploration. The selected techniques are finally tested and validated. The details of exploration and conclusions drawn should be described in a report.

Step 5: Implementation, Evaluation and Monitoring

Once a technique has been selected and validated, the decision makers can implement the model for use. It is possible that more than one technique proves to be effective for the problem under consideration. It is then required to evaluate and compare these techniques

on their results and select one technique. Furthermore, it is vital for an enterprise to continually monitor the performance of the models that have been implemented. With the evolving enterprise from time to time, it is important that the data mining system also evolves. Therefore, monitoring is important to refine the tools and techniques that have been implemented. Details of implementation and evaluation are documented in a report. Also, instructions for monitoring are also described in a separate report.

Step 6: Results Visualization

This step involves explaining the results and inferences drawn from the studies of the data mining process to the decision makers. Data mining tools contain visualization [18] modules and these tools are necessary in disseminating results to the managers. Those tools that have been tried and tested are used if found powerful for a specific problem.

The diagram in figure 1.2 shows the various steps explained above.

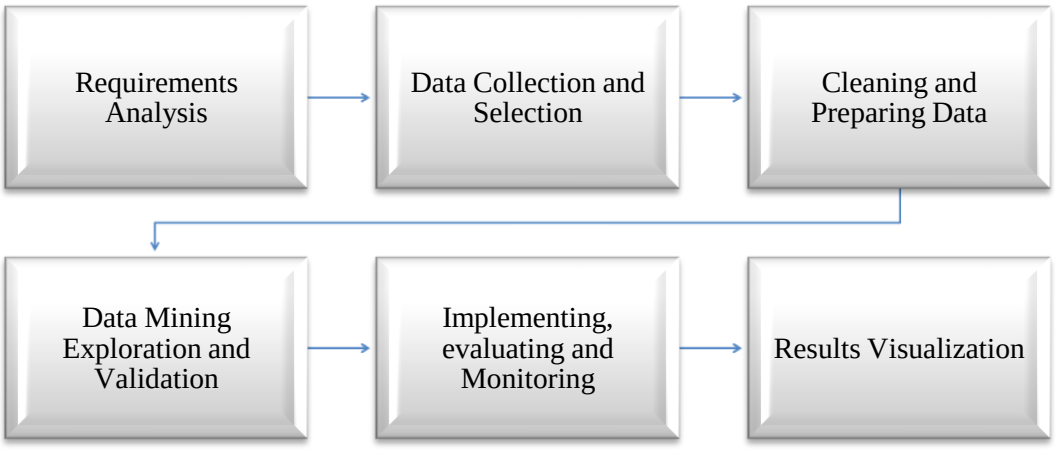

Figure 1.2: A typical Data Mining Process- Software Development Approach

1.4 Applications of Data Mining

Some of the applications of data mining have been listed below.

1. Prediction and Description

It is seemingly important for an enterprise to predict some aspects of business to be able to make appropriate decisions in the future. For example, the number of students that may enroll in the coming year in a university. It may also be important to describe particular data.

2. Relationship Marketing

   Data mining helps in formulating convincing methods to maintain congenial relationships with customers. It segregates customers into sections such that customers within same segment tend to be similar in their buying behaviors also. This is useful in estimating which customer should be subjected to which policies. This is important for customer retention and ensuring customer loyalty.

3. Customer Profiling

   Customer profiling is the process of discriminating customers based on a number of features and past activities to describe their purchasing decisions. Profiling helps a business enterprise in identifying its most valuable customers so that they know their needs. It also helps in increasing lifetime value of customers.

4. Customer Segmentation

   It is the process of finding sub-groups of similar people in a dataset and can be useful in marketing. It is used to assess and serve individuals based on their needs and status. It helps in understanding customer behavior, develop new services or products and to effectively market them.

5. Outliers Identification and Detecting Fraud

   Data mining is extremely helpful in detecting unusual activity and therefore frauds. Frauds in sectors such as banking [19], insurance, social welfare, taxes, healthcare, telecommunications and customs can be effectively revealed using data mining.

6. Website Design and Promotion

Data mining can help in discovering how users navigate through a website and their results can aid in developing effective sites and making them more visible on the internet. It may also be exploited in cross-selling by giving suggestions to customers based on their buying and search patterns.

The chart in figure 1.3 below shows the most common fields using data mining.

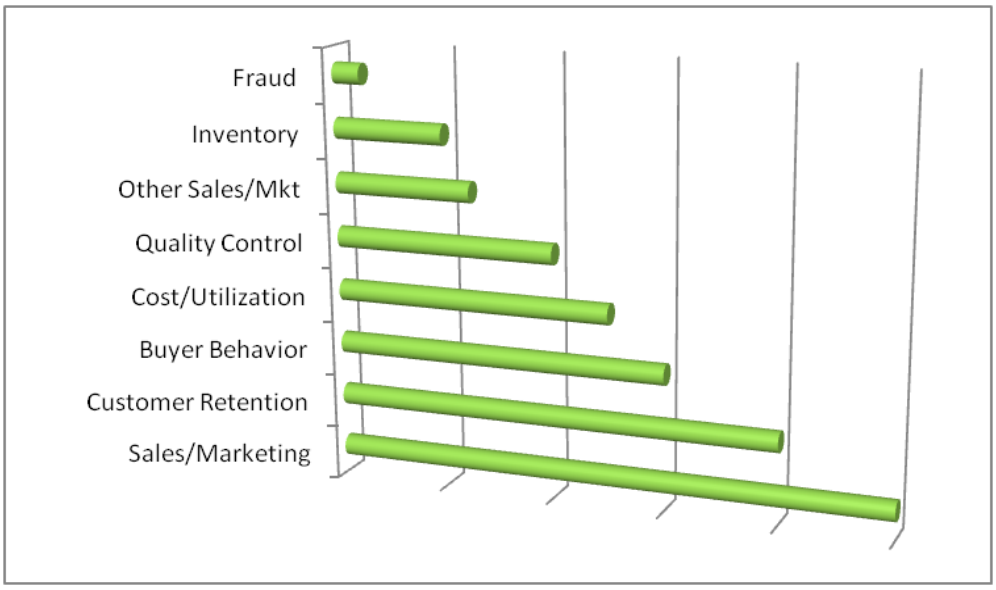

Figure 1.3: Common Fields using Data Mining

Some practical examples of data mining are:

- Astronomy
- Banking and Finance
- Business
- Crime Prevention
- Education
- Government
- Healthcare
- Manufacturing
- Telecommunications
- Transportation

The diagram in figure 1.4 shows popular fields extensively using data mining techniques.

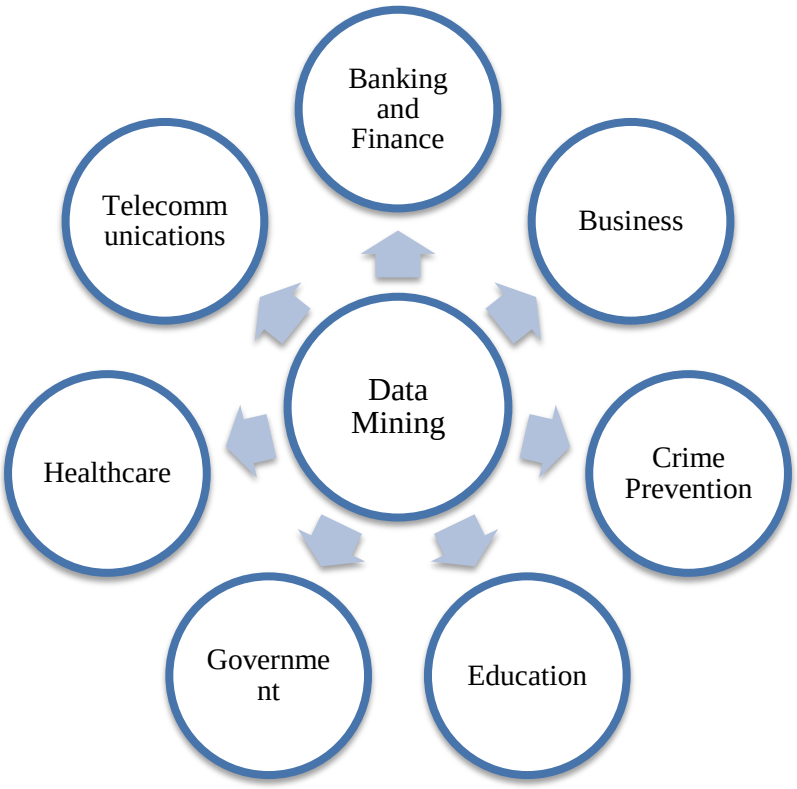

Figure 1.4: Practical Examples of Data Mining

1.5 Machine Learning

Machine Learning is a methodology of data analysis [20] to automate analytical model building and a component of data mining. It learns from data provided to it, termed as training data in an iterative mode by using statistical algorithms allowing computers to find hidden insights about certain data without the need of explicit programming. The algorithms can be either predictive statistical [21] algorithms or conventional statistical algorithms. Machine learning is more about predictive modeling. The models learn from data, so it comes out that process of exposing a model to data iteratively is extremely important. Models adapt independently and develop a capability to predict similar data if exposed to training data appropriately as they learn more by exposure to data. They learn from known information and facts to present reliable decisions and results for the future.

Machine Learning is becoming extremely popular due to the same reasons as those of data mining and Bayesian [22] analysis. These factors are growing volume and varieties of available data, affordable data storage and more powerful and cheaper computational processing.

This means that machine learning is capable of quick and automatic model building to analyze big and complex data to provide with results that are faster and more accurate. Conclusively, we can say that by building precise models an organization escalates its chances of establishing profitable opportunities and avoiding risks.

1.5.1 Popular Machine Learning Methods

Supervised Learning algorithms learn or, train through labeled examples. For example, parts of equipments in a factory could be example data and labeled either “R” (runs) or “F” (failed). The learning algorithm receives input data and corresponding correct output labels and compares actual outputs to correct labels and finds errors to modify the algorithm accordingly. Supervised learners include classification, regression and gradient boosting methods [23]. It uses learned patterns to label non-learned data that are its predictions. It is often useful in areas where historical events have a strong correlation with future events and therefore the probability of several future events can be evaluated with this type of learning algorithms. For example, you can determine whether or not a customer can be approved with bank credit by looking for past records and details of defaulters, if a credit card transaction is likely to be fraudulent or if an insurance client is likely to file a claim. The diagram in figure 1.5 explains the process of a supervised learning algorithm.

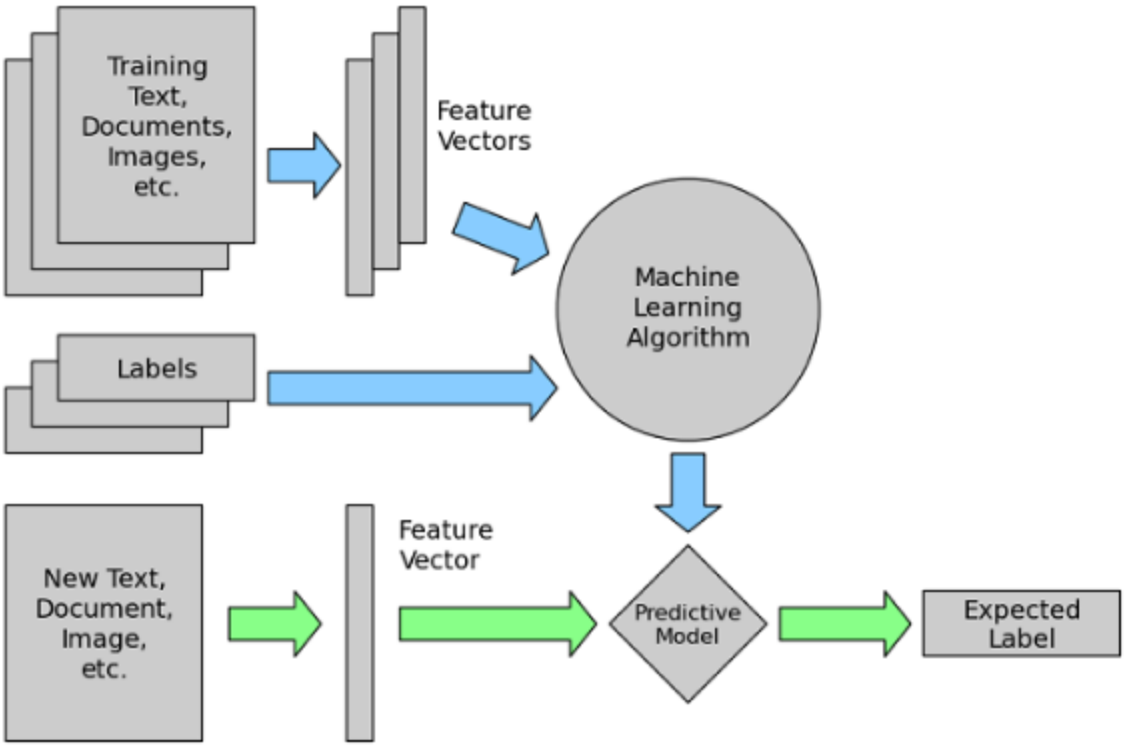

Figure 1.5: Supervised Learning

Unsupervised Learning algorithms work on unlabelled data. It does not always happen that the data to perform learning upon will have class labels. We often come across data wherein no class labels have been given but only the features of the data points. Examples of such kind of data include bioinformatics sequences data and medical imaging data. In such cases, unsupervised learning algorithms are used. In this, the system must figure out patterns in data. It explores data to find structures within. Popular techniques include mapping of nearest neighbors [24], k-means clustering [25], self-organizing maps [26] and singular value decomposition [27]. The most popular application of such algorithms are found in bioinformatics for genetic clustering and sequence analysis, data mining for sequence and pattern mining, medical images for segmentation of images and computer vision for recognition of objects. These algorithms are also used in areas such as identifying data outliers, recommending items and segmenting text topics. They are used in identifying segments of customers having

identical attributes to be able to manage them in similar ways in marketing campaigns. Figure 1.6 explains the workflow of an unsupervised learning algorithm.

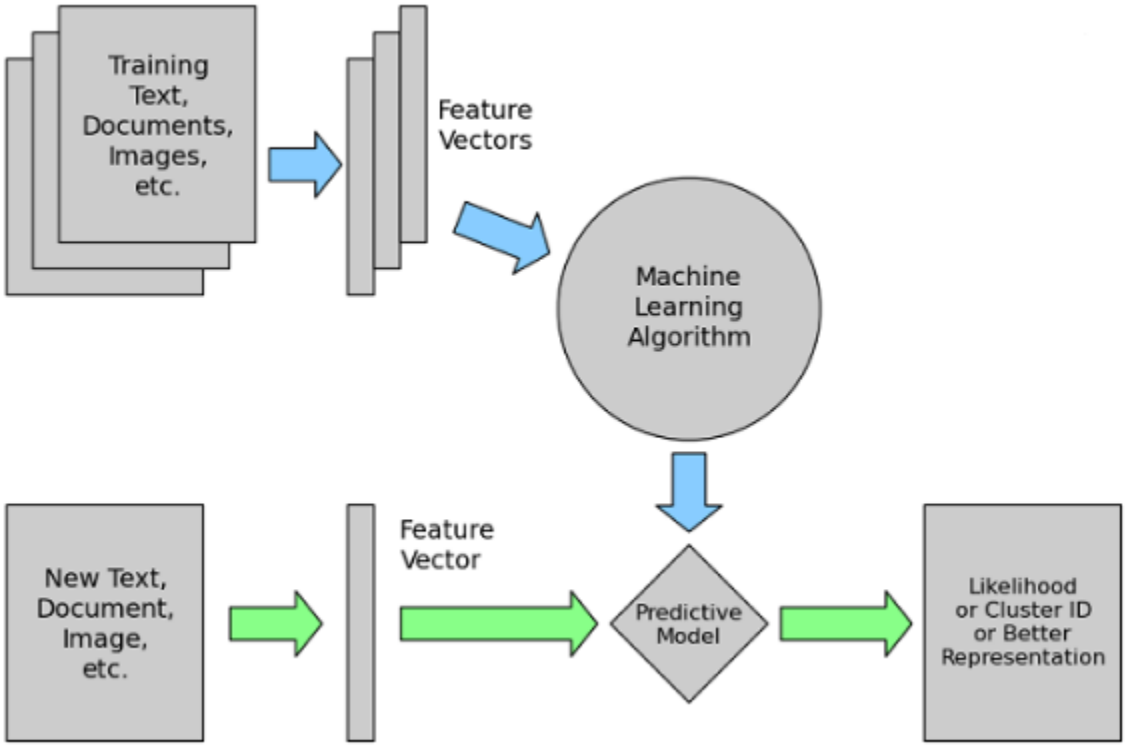

Figure 1.6: Unsupervised Learning

Semi-supervised learning learns from labeled and unlabeled training data. The amount of classified data used in this technique is very small compared to the unclassified data since the cost of the classified data in such applications is extremely high. It can be used with prediction, classification and regression methods. Semi-supervised learning is preferred in applications where the cost of acquiring labeled data is extremely high to allow for a training process that is fully labeled. Identification of facial attributes of a person on a web cam is a common example of this type of learning. Figure 1.7 shows the working principle of semi-supervised learning algorithms.

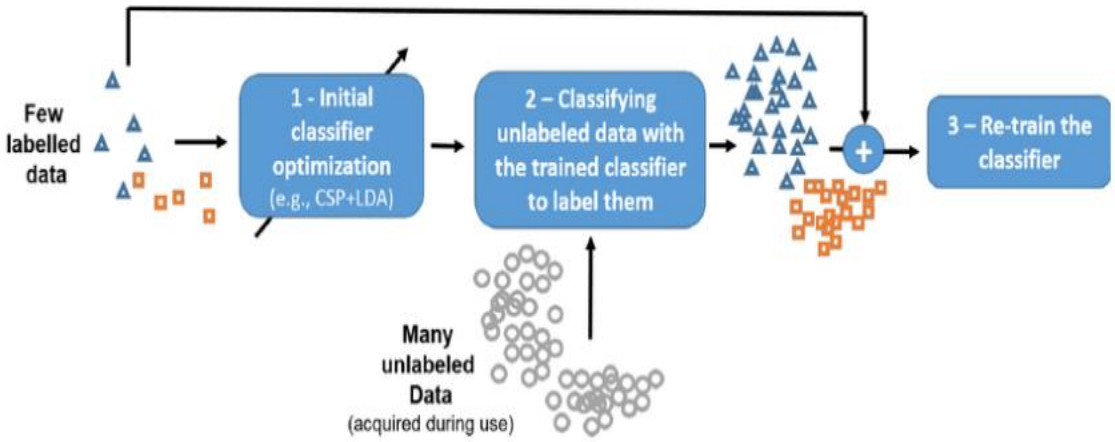

Figure 1.7: Semi-supervised Learning

Reinforcement learning is used in games, navigation and robotics. In reinforcement learning, there are three primary components: agent, environment, and actions. The agent is the apprentice or the decision maker, the environment is all that the agent interacts with and the actions are the decisions made by the agent.

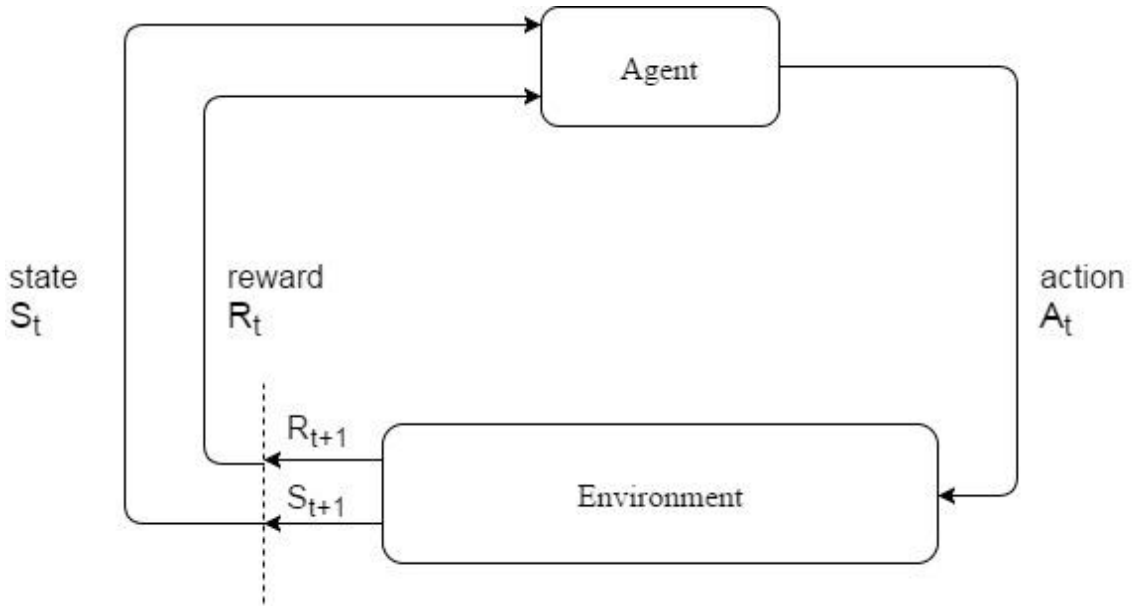

Figure 1.8: Reinforcement Learning

This type of algorithms gain experience through trial and error to learn about the actions that yield the greatest rewards. Therefore, the objective of this type of learning is to maximize the expected reward within a stipulated time period. If the agent follows a good policy, it reaches the goal faster. So, the goal in reinforcement learning is to learn the best policy. Figure 1.8 explains the phenomena of choosing the best action on which reinforcement learning is based.

1.6 The Telecommunication Industry

There is tough competition in telecom industry because of the large number of companies in the market. It is very easy for customers to switch their services from one service provider to another. This activity of the customers of switching from one company to other is referred to as "churning". The churning of customers is very costly for the telecommunication industry. It is easier and cheaper for the companies to retain existing customers and make sure that their complaints are addressed than trying to acquire new customers. Acquiring new customers in telecommunication industry is more taxing both in terms of expenses and efforts than retaining existing customers. Telecommunication companies face a continual threat of losing customer loyalty and in turn loss of revenue caused by customers who may leave services of the company. If such customers increase, then it becomes a crucial problem for the company to find out who those customers are and find patterns and possible reasons for their leaving the services. Therefore, it becomes important to have a decision support system that enables the CRM to figure out customers who are likely to switch their service providers. Companies maintain detailed call records of their customers in order to use these records in predicting potential churners. Churn analysis is useful in many sectors and particularly gaining notice in the telecommunication sector. The data mining techniques help in predicting churn by collecting old records of customers who had churned and those who did not in a particular time period. Data mining and machine learning techniques learn from records provided to them which is termed as training. During training they are able to find patterns in data. These patterns learnt during the training are applied to some other dataset which is currently under study in the company. The DSS then predicts for each customer of the company, whether it has the tendency to churn or not. Though, it does

not ensure that the predictions will be completely true. Also it never happens and is not possible for any decision support system to give 100% true results.

Telecommunication companies use machine learning and data mining in several applications including financial management, sales, marketing, future prediction, etc. Companies may collect customer data through many techniques like business intelligence. Applying data mining to this data helps in formulating new policies, tariffs, and developing campaigns for existing customers. Programs can be built to increase customer loyalty. All this may lead to greater customer satisfaction and benefits from reduced costs and risks. Apart from predicting customer churn tendencies, companies can analyze as to why there is a tendency in some customers to switch providers. The likely reasons for churn can be figured out through machine learning. Once it is known why customers churn, it is all the easier for companies to assess where the rules and policies related to services can be improvised to match up to the customers' expectations so that they may rethink before switching to other companies. This is what happens, when a certain company reduces its tariff, other companies also follow suit and reduce their call rates. Customers' churn behavior is the biggest reason why telecommunication companies lose out on massive amounts of revenue. Such customers create financial strain on the companies. These behaviors cause extreme sickness to the company. Thus, it is of utmost importance for the companies to detect such customers who could be having a tendency to move to a competitor so as to avoid financial glitch to the companies.

1.6.1 Data Mining and Machine Learning in Telecommunication

As noted earlier data mining can be applied in telecom industry in various spheres. Some of these are discussed below:

1. Churn Prediction

   Predicting customers who are having a tendency to leave a company and shift to a competitor due to some or the other reasons is called as churn prediction in telecom. Predicting such customers and tracking their activities well in advance is important so that they may not lose their customers as it is easier to hold on to prevailing customers than to secure new ones.

2. Insolvency Prediction

   Customers who decline to pay bills are called insolvent customers. Insolvency [28] is a big problem for telecom companies. To detect customers who may decline to pay bills, data mining techniques can be used. These techniques will help to detect such customers well in advance.

3. Fraud Detection

   Fraud is a major concern for telecom companies. This problem leads to damages to the finances of a company since fraudsters "leech" the revenue of the operators. The main telecommunication fraud is the type where a third party uses the operator infrastructure without the intention of paying them for the services used. This type of fraud corresponds to abusive usage. Other frauds include using the identity of legitimate clients in order to commit fraud.

This results into legitimate clients being framed for fraud. This does not go well with clients and results into the company losing out on a customers' confidence in their carrier and in turn giving them reasons to leave. Besides, some clients simply refrain from being attached to a company that has been a victim of fraud attacks and since multiple carriers offer the same services, the clients switch their carrier.

The problems listed above cause the telecommunication companies to take action against such misuse and fraud of their carrier and resort to bring data mining into picture to analyze data for any such possibilities. Data mining techniques have the most viable solutions against such practices since they allow identifying fraudulent activities with certain degree of confidence. They also work with large amounts of data which is an imminent characteristic of data generated by telecommunication companies.

Apart from these, there are many other applications of data mining and machine learning in telecommunication sector as shown in figure 1.9.

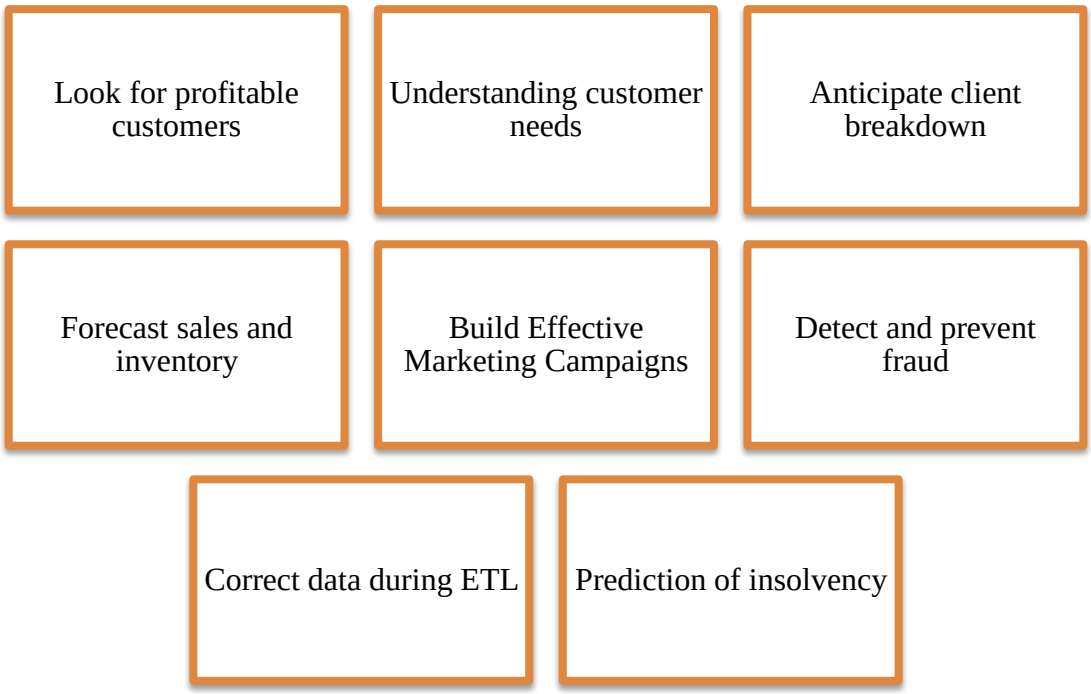

Figure 1.9: Applications of Data Mining in Telecommunication

1.6.2 Churn in Telecommunication

A major concern for CRM [29] in telecommunication companies is the customer activity called “churning”, especially because of the low costs and ease associated with switching service providers in telecommunication industry and the difficulty and high cost of acquiring new customers. Thus, churning is a costly activity for companies.

As shown in Figure 1.10, customers use the services of the carrier for a period of one month. This period is called the billing period. The company issues the invoice one week after the end of the billing period. The payment due date is usually two weeks after the invoice is issued. If the invoice is not paid during this period, the company takes action against the connection of such customers. The company disconnects the channel in a way, which means that the customer can no longer make outgoing calls and can only receive incoming calls until the customer makes the payment.

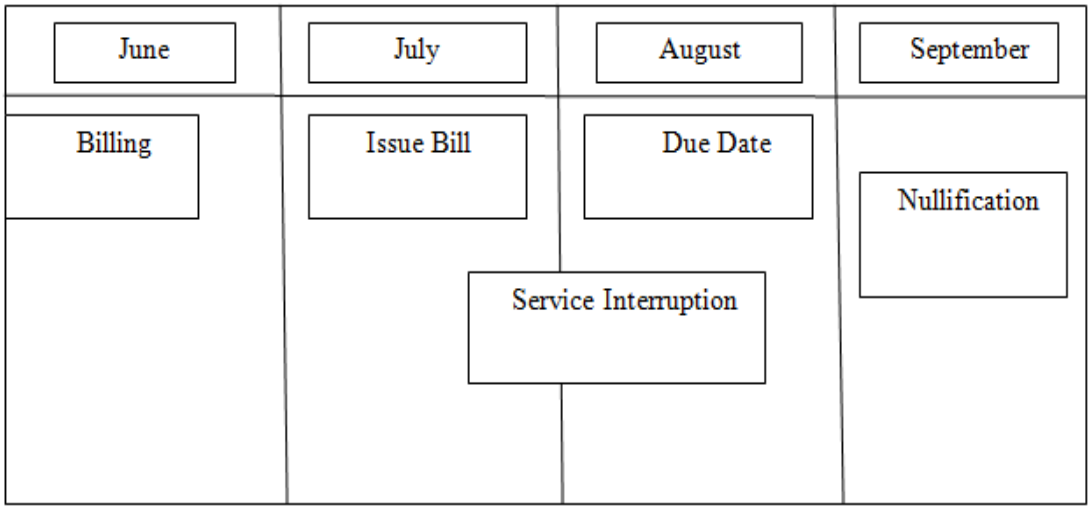

Figure 1.10: Stages of services and billing in a telecommunication company

If the customer pays the bill, the connection is reset. If the customer does not pay the invoice within 30 days after the disconnection of the channel, the company rescinds the client's contract. The amount that the customer owes is transferred to the bad debts and the company considers the money most likely lost. Due to these types of clients, telecommunications companies support huge amounts of debt. Detection of these clients are likely to become the goal of this work.

1.6.3 Defining Brand Switching

Customers are the key players of the market. The affluence of a company greatly depends upon their customers. A company which has a loyal customer base is likely to outperform the other players of the same market. Organizations look forward to achieve strong hold on their customers to achieve long-term success. Companies can achieve long-term success if they work along with all their stake-holders including the customers as they are the most important stake-holders of a company. For different service-providers, customers have different attitudes, behaviors and perceptions. There are often a number of reasons for customers switching from one service provider to another. Customers' shift from one company to the other is termed as brand switching behavior of customers.

1.6.4 Types of Brand Switching

There are two types of brand change: one temporary and one permanent. The temporary change of brand is when the customer changes from brand "A" to brand "B" due to non-availability of brand "A". Therefore, the person may resort to a different brand to fulfill his requirements for some period of time. Brand-switching can last long if customer shifts brands permanently. Brand-switching is a process where customers shift from a product or services of one brand to another brand. Permanent brand-switching is harder to change.

There are two more categories in which brand change can be categorized. They are aggressive and defensive brand switching. Aggressive brand-switching is the type where the customers are influenced by the advertisements of various companies to switch from other brands to their brands. They use lucrative offers to attract customers. Defensive brand-switching on the other hand, is where companies use different ways to convince existing customers to buy again product/services from their brand. Here also, companies use advertisements and promotions. Defensive brand-switching is to induce loyalty in customers.

1.6.5 Factors inducing Brand Switching

The advertisement plays an important role in the change of brand. They influence customers to change brands. There are many strategies used by companies to attract competitors' customers. Ads encourage customers to change their loyalty to a company. This in turn improves sales of a company. All companies in the market advertise to grow customer base. In this way, total effect of advertisement is equalized. Advertisements are a motivational factor in persuading customers to switch to their brands. Therefore, all companies advertise in order to convince customers to follow their brands.

A number of factors induce brand change in customers. If companies are not able to meet customer expectations, then the customer always has an alternative and can easily switch to other brands, as there are many competitors for any product or service in the market. The market is competitive for any kind of products and services and each player in the market strives to acquire market share. Companies attract customers using various offers

such as lower prices, discounts, higher quality and promotions with the objective of acquiring market share. In today's market, business success depends on the customers' buying intentions. The success of a business may be in jeopardy if it is not able to control the switching behaviors of customers.

1.6.6 Brand Switching in Telecommunication Industry

Brand change behavior of customers varies from industry to industry. Industries can be characterized by their switching costs. Some industries may have high switching costs while some others may have moderate to low switching costs. Industries that have high switching costs seem to have low competition and generally have loyal customer base. It is the industries with low switching costs that suffer with the problem of customer switching to their competitors which is the reason for high competition in such industries. The prime example of such industries is the telecommunication industry. It faces fierce competition due to both large number of companies in the market and extremely low switching costs. These are some reasons that cause the telecommunication company to have high rates of switching. Customers of mobile service providers switch brands easily if they are offered better services elsewhere at similar prices. Due to the low switching costs in the telecom industry, customers find it easy and affordable to change brands and frequently switch from one service provider to another. Therefore, it has become crucial for service providers to investigate the switching behavior of customers.

1.6.7 Churn Magnitude in Telecommunication Industry

The mobile services market segment of the telecommunications sector is one of the fastest growing where potentially more than 75% of all phone calls can be made with mobile phones. Like any other market, the whole competition approach has shifted from customer acquisition to retention (Kim & Yoon, 2004). We have collected statistics for churn magnitude and costs related to it to understand why it is necessary and useful to research under this area and to gain useful insights about the field. These are beneficial to picture a mental image of the problem. The facts are as stated under.

- SAS (2000) has reported that the telecommunications sector undergoes an annual turnover rate of between 25 and 30 per cent and that it still has the potential to increase in correlation with market growth.
- Annual churn costs of European and US telecoms companies have been estimated at US $ 4 billion (SAS Institute, 2000).
- The ratio (cost of customer acquisition / customer retention or satisfaction costs) would be 1: 8 for wireless companies (SAS Institute, 2000).

According to Groth (1999) and SAS Institute (2000), the annual churn rate of customers in the telecommunications sector has been estimated at about 30 percent. In addition, annual customer retention costs for the telecommunications sector in Europe and the United States are estimated at $ 4 billion. Therefore, it seems reasonable why mature companies should invest more in churn management rather than in procurement management, especially considering that the cost of acquiring a new customer is eight times more than retaining an existing one (SAS Institute, 2000).

1.7 Thesis Organization

The thesis is organized into following 6 chapters:

- Chapter 1: This chapter describes data mining and its applications. It gives an insight on the emerging field of data mining and how it is coming across as an important field in various industries. It also describes how telecommunication industry functions and the issues and challenges faced by it.
- Chapter 2: This chapter provides a survey of literature in the related area.
- Chapter 3: In this section, motivation behind the work, problem description and objectives are stated.
- Chapter 4: This chapter explains the proposed method and technologies used.
- Chapter 5: This chapter discusses the results obtained from the proposed methodology. Performances of all techniques after each step are discussed and evaluated in this chapter.

- Chapter 6: Finally, chapter 6 states the conclusions and also gives an idea about the possibilities of future work that can be extended in this field.

Chapter 2

Related Work

Churn prediction has become an important activity in telecom these days and there is a high demand for models that can predict churn as accurately as possible. Therefore much research has been performed in this area. Some related work in this field is described below.

Adnan Idris et al. (2012) [30] discusses churn prediction of telecom using Random forest and KNN method. In order to handle the imbalance in data distribution, PSO is used for under-sampling the dataset, along with feature reduction techniques like Minimum Redundancy and Maximum Relevance, Fisher's ratio, PCA and F score. Performance of the modeling methods is evaluated using specificity, sensitivity and AUC parameters. These simulations were found to be successful in positive prediction of churn.

Yeon Soo Lee et al. (2012) [31] have also worked on the churn prediction for customer retention using Genetic algorithm approach. For each class, various programs were generated using Adaboost method. These programs were used for predictions using the highest performance, from the weighted sum of the results of the programs by class. A 10-fold cross-validation technique was used to verify the accuracy of the prediction and an area under the curve of 0.89 was found.

Javed Basiri et al. (2010) [32] used OWA (Ordered Weighted Average) to fuse the output of each learned classifier to introduce a hybrid approach to improve the accuracy of the results obtained. In this study, different number of features were generated to study various algorithms like bagging, boosting and LOLIMOT algorithm are learned. The results generated showed that the approach was good enough than some well-known classifiers.

Chih - Ping Wei et al. (2002) [33] proposed a technique to identify potential churners at the contract level for a specific prediction period. In their technique, they addressed the

problem of biased distribution of data by adopting the class combiner approach of several classifiers. The results of the empirical evaluation showed that the multi-class classifier approach was better than the single classifier approach. The proposed technique gave more accurate predictions when more recent call details were used for the construction of the churn prediction model. In addition, the proposed technique was able to produce more efficient predictive powers when there was a one-month interval between model construction and churn prediction.

Shin –Yuan Hung et al. (2006) worked with two approaches to evaluate the performances of decision tree (C4.5) and BPN. In the first approach, they used the K-means grouping method to segment customers into 5 groups according to their tenure months, billing amount, and payment behaviors. Then, a decision tree was created in each group to evaluate whether the churn behavior is different in different value loyalty segments. In the second approach, the neural network has been used to segment clients followed by decision tree classification in order to test whether BPN could improve decision tree accuracy. The evaluation measures used were the ratio of stroke and elevation. In addition, they included customer complaint logging and customer service for modeling, as suggested in a previous research by Wei and Chiu (2002). In general, they found that DT without segmentation performed better than DT with segmentation and that LBP exceeded DT without segmentation.

Kristof Coussement et al. (2006) [34] used a newspaper subscription context and applied SVM to it to build a highly accurate churn model. They apply two parameter-selection techniques to SVM, where both the techniques are based on grid search and cross-validation. The parameter selection techniques are a major factor in improving the performance of SVM. Afterwards, both SVMs are benchmarked against Logistic Regression and Random Forest. The study shows that SVMs perform better when applied to noisy marketing data. It has been shown that RF outperforms both LR and SVM. SVM is better than RF only after the application of optimal parameter-selection technique. This study shows that the most important parameters are from the group that described the subscription services. It says that unlike many researches, monetary value and frequency are not among the most important churn predictors.

Yaya Xie et al. (2008) [35] used IBRF to test its effect on churn data. IBRF (Improved Balanced Random Forests) is a novel method proposed by them. They integrated cost-sensitive sampling and learning techniques into the standard random-forest approach to study their effect on churn data. They applied this data set to a churn data set of bank customers. It was found that IBRF worked better than many algorithms such as artificial neural networks, decision trees and support vector machines. IBRF also behaved better against other versions of random forests such as balanced random forests and weighted random forests.

Luo Bin et al. [36] have worked on Personal Handyphone System Service (PHSS) to overcome its limitations of lack of information of customers of PHSS of China Telecom. They have presented three experiments: changing the subperiods to form data sets, changing the cost of misclassification in the rejection model, and changing the sampling methods to form data sets to construct an accurate and predictive churn prediction model. All of these experiments were performed on model decision tree to improve prediction performance. It was observed that the optimal parameters that gave the best results are: - 10 days subperiod time, 1: 5 misclassification cost and random sample method for the train set. The results of the evaluation show that the proposed churn models have excellent performance given that there is less information and that the class distribution is skewed. The study says research is useful not only in predicting churn but also in other applications that have similar characteristics.

Essam Shaaban et al. (2012) [37] proposed a new churn prediction model which consists of six steps. They are identifying problem domain, data selection, research dataset, classification and clustering and knowledge use. A data set of 23 attributes and 5000 instances was selected. 4000 instances were used in the formation of the model and 1000 cases were used as set of tests. The planned churners were classified into three categories. The algorithms used in this research are Decision Trees, Support Vector Machines and Neural Networks. The simple K-Means technique was used to perform clustering. It turned out that the SVM outperformed the DT and ANN algorithms. In

future work, the author plans to develop a prediction model of churn by combining predicted data and grouped data to propose a new technique that is capable of suggesting retention schemes to each type of churner.

V. Umayaparvathi et al. (2012) [38] explore the applications of data miming in predicting customer churn. They study the impact of attribute selection in model building. They compare the efficiency of Decision Tree and Neural Networks classifiers. The evaluation parameters used in this study are the accuracy, false positive rate and false negative rate. Appropriate attributes were selected and were grouped under 4 categories namely- Customer Demography, Bill and Payment, Call Detail Record and Customer Care Service. It was observed through experiments that decision tree had a predictive accuracy of 98.88% and that of neural network was 98.43%. Clearly, the decision tree model outperforms the neural network model. In future research, the author may focus on retention policies to be framed by selecting the appropriate variables from the dataset.

Afaq Alam Khan et al. (2010) [39] work on three algorithms to predict churn- Decision Tree, Logistic Regression and Neural Network. They study to identify the best churn predictors and also the importance of clustering by incorporating the cluster membership information in classification models. They found that all the major features were either demographic, billing or usage features. The purpose of their second objective in this study is to find out the importance of these three types of features. By experimentation, they realized that demographic features were the least important in predicting churn. The billing and usage features were found to be equally important. Cluster membership incorporation was used in all the three algorithms and it was noticed that not including the cluster membership information degraded the performance of the models considerably. It was found that Logistic Regression and Neural Network showed the accuracy of 89.01% and 89.08% respectively which was a slightly higher than that of decision tree which was 87.74% . Therefore, in this study LR and ANN performed better than DT.

Marcin Owczarczuk (2009) [40] tests the usefulness of regression and decision trees approach on Polish cellular telecom company data to predict the churn of customers. His work is novel in three ways. They are 1) prediction is done on prepaid clients 2) use 1381 variables in this study and 3) test the stability of the models across time for all the percentiles of the lift curve. They test the data after 6 months of model estimation. They learnt through experiments that linear models, especially logistic regression model works best in case of churn prediction of prepaid customers. Decision trees showed poor performance in high percentiles of the lift curve. Apart from this, they also showed that churn prediction models work better when trained on large data marts. In future, they propose to work on prediction models for mix sector which has the attributes of both the postpaid as well as prepaid sectors.

Ning Lu et al. (2014) [41] propose the use of boosting as a method to predict customer churn. In this research, boosting is applied not only to boost the accuracy of the predictions like most other researches but segments the customers into two different clusters based on the weights assigned by the boosting algorithm to the instances. As a result, higher risk customers have been identified. Logistic regression was used as a base learner in this study and a churn prediction model was built on each cluster and results compared with single model. Their findings have led to the conclusion that boosting provided a good segmentation of the customers which led to better results of churn prediction thus suggesting boosting for churn prediction analysis.

Palupi D. Kusuma et al. (2013) [42] investigate the value of combining regular tabular data mining with social-network mining. They combine classic churn datasets with social network neighborhood predictors. In a second approach, they extend the traditional models of activation of the social network with information of the classic models of churn tabular. Their results said that in the second approach, the combination of social media mining improved results, but overall tabular data mining functioned better.

Chih-Fong Tsai et al. (2009) [43] considered two hybrid models combining two different neural network techniques that are artificial neural networks of retropropagation

(RNA) and self-organizing maps (SOM). The hybrid models are ANN + ANN and SOM + ANN. Here, the first algorithm performs the task of reducing data by filtering non-representative training data. The second algorithm is used to construct the model in the training data output by the first technique. The test is carried out on three types of general test dataset tests and two fuzzy sets of filtering action by the first technique in both hybrid models (ANN and SOM, respectively). The experimental results showed that the hybrid models behaved better than the single basal RNA model. In addition, the ANN + ANN model gave significantly better results than the other hybrid model, ie SOM + ANN.

Wouter Verbeke et al. (2010) [44] studied the combination of induction techniques of comprehensible rules like C4.5 and RIPPER with more precise techniques. The AntMiner + technique was used to include domain knowledge by applying monotonicity constraints on the final rule set. AntMiner + is a high-performance mining technique based on Ant Colony Optimization. The author uses the ALBA technique that combines the high predictive accuracy of the nonlinear support vector machine model with the intuition of the rule set format. Here, they benefited from the precision of the SVM model and the comprehensibility of the rule set format. ALBA improves the learning of classifiers and with a good understanding and performance. The results are compared with C4.5, RIPPER, SVM and logistic regression. The greater precision is obtained when ALBA is combined with C4.5 or RIPPER. Sensitivity is highest when using C4.5 or RIPPER in oversampled data. AntMiner + does not produce highly sensitive rules, but is able to incorporate domain knowledge. It gives understandable rule sets that are much smaller than those given by C4.5. RIPPER also offers small, understandable rule sets, but leads to non-intuitive models.

Chapter 3

Problem Statement

3.1 Motivation

In today's times of globalization and competitive business environments, it is quite natural for enterprises to get involved and invest in areas such as business intelligence and business analytics. These are the fields that assist organizations into better functioning and hold a good base in the market to be able to sustain and emerge as a key player in the market. For this, they need to understand the practice and process of CRM (Customer Relationship Management) and be able to protect themselves from competitors.

Business intelligence [45] is a broad category of computer software solutions consisting of tools and techniques such as reporting and analysis tools to enable a company to gain insights into its operations. BI applications consist of a wide variety of applications such as spreadsheets, tabular reports, charts and dashboards. A company has many "information assets" such as supply chain information, customer databases, personal data, product data, manufacturing and sales and marketing data. Well-designed BI applications give useful insights into these information assets of a company to enable it to function more effectively and take better decisions for the company. They help in understanding such sources and any other critical information as well.

Business analytics [46] is a combination of BI and deeper statistical analysis. It is a modern and emerging analysis technique to see into the depth of a business model of an organization. It is a way to explore and analyze the business policies and improve upon them to attain the best possible growth in various aspects. BA makes use of statistical techniques to predict future trends and possibilities, called predictive analysis. It may use historical or current data to predict the future performance of a product or service. It may also use advanced analytics techniques like cluster analysis [47] which may be helpful in certain applications such as customer segmentation for targeted marketing campaigns. It has applications, tools and practices that enable an organization to make the optimal decisions about their businesses. As discussed in the previous section, churn management and customer retention are important issues for telecommunication companies. Being one of the major challenges in telecommunication sector, churn management has the need to be handled intelligently and efficiently. Earlier, simple statistical techniques were used

for this purpose which had their own limitations. In order to make better decisions, modern tools and methods should be explored which can be done through data mining and machine learning. Therefore, for addressing this issue with new method, we have proposed a novel churn prediction mechanism.

3.2 Problem Description

Customer churn is a focal concern for industries that have low switching cost. This immediately brings in notice the telecommunication companies as these are the companies that face the challenge of customer churn the most due to negligible switching costs. The telecommunication industry suffers the most in this regard amongst all industries with an approximate annual churn rate of 30 per cent. This implies wastage of efforts and money. Consequently, to tackle this problem, development of models and techniques has become necessary to enable companies to determine churn activities of customers if any. These methods must be capable of identifying customers who may possibly churn in near future. In cases of prepaid mobile customers, it is difficult to not only trace customer churn but also to define churn due to the non contract-based nature of prepaid mobile telephony. Thus, in such cases, building a predictive model is of high complexity. Furthermore, in machine learning it is important to handle class imbalanced data as is the nature of churn datasets. Churning is rare phenomena in large churn datasets so these datasets almost invariably suffer from class imbalance problem against the churn class.

3.3 Objectives

The main objectives are:

- To preprocess the unorganized and incomplete data in an efficient manner so as to prepare data for analysis.
- To design an efficient prediction model for the telecommunication churn
- To test the effectiveness of the prediction model using real data.

Chapter 4

Proposed Method and Technologies Used

4.1 Proposed Method

In our work, we have used the churn dataset made publicly available by IBM. The first and the foremost step towards our work is to discard useless features and treat missing values using appropriate methods. The next step towards our implementation is that of sampling the data. In telecommunication datasets, there is a problem of skewed data distribution [48]. Due to this, classifiers perform unsatisfactorily in predicting churn. Therefore, in this work we handle skewed class distribution and then apply suitable ensemble [49] approach to build a classifier that exhibits better performance than any of the individual models. The block diagram in fig 2 illustrates the various steps undertaken in this work. The dataset is initially preprocessed to handle missing values. SMOTE based sampling [50] is applied to investigate its effect on the prediction results. Feature extraction algorithms such as Information Gain, Gain Ratio, Correlation and OneR are employed to study the effect each of these methodologies imposes individually on the data under study. In our work, we study the effect of CART [51], bagged CART [52] and PART [53] algorithms on our dataset to measure how effective these algorithms are on this particular dataset. Adaboost.M1 [54] algorithm is applied on the newly blended dataset that has been achieved by incorporating the predictions of base classifiers to act as additional features on the initial dataset. Sampling is employed in combination with feature selection techniques and classifiers are used to study the improvement in predictions, if any using AUC [55], specificity and sensitivity measures [56]. Figure 4.1 shows the basic block diagram of the proposed approach.

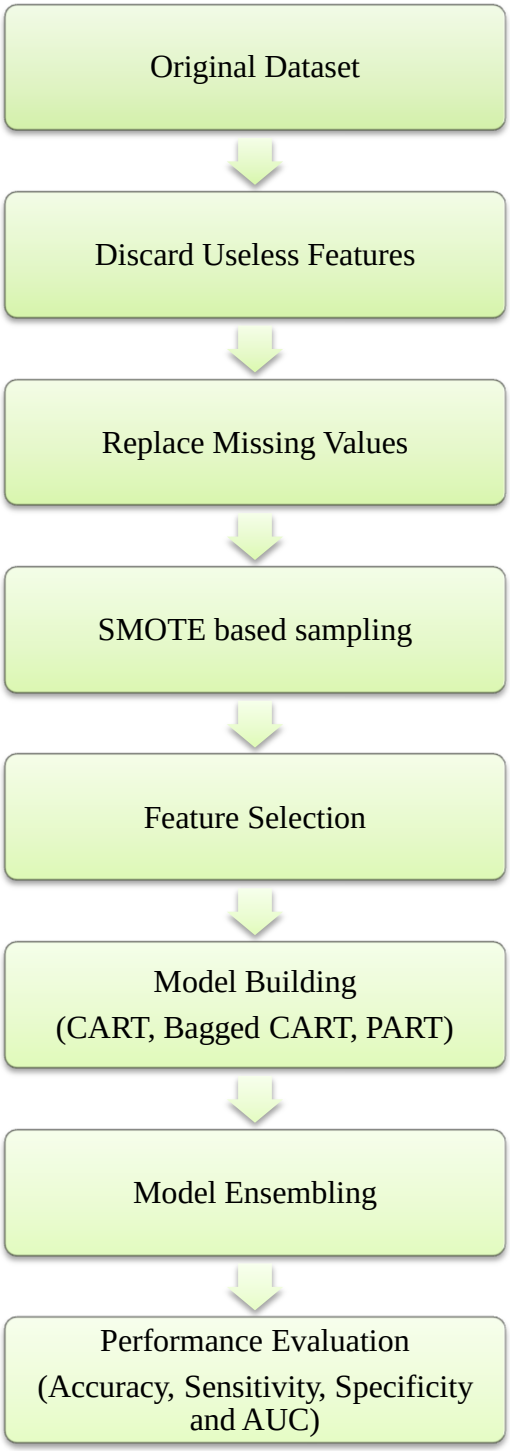

Figure 4.1: Basic block diagram of proposed approach

4.1.1 Dataset

The dataset that has been used in our study has been taken from IBM Watson Analytics and is publicly available. Our dataset has 7,043 instances and 21 features. The dataset has 4 numerical and 17 nominal attributes. It has 1,869 records of minority and 5,174 records of majority class that amounts to 26.5% contribution of the minority class in the whole dataset. We can see that the dataset is slightly biased towards the non-churners.

4.1.2 Data Preprocessing

We preprocessed the data to discard useless features. In data preprocessing, data cleansing is performed in our data set.

Fill in missing values

- Fill in the missing values by ignoring the instance in cases where the class tag is missing.
- Use the attribute mean to fill in the missing value
- Use of mean attributes for all samples belonging to the same class.
- Predict lost value using a learning algorithm: consider the missing attribute as a dependent variable (class) and execute a learning algorithm to predict the missing value.

4.1.3 Data Balancing

In our work, we are examining telecommunication data for customer churn prediction. Churn is not a very frequent activity in any sector. Due to this characteristic of telecom churn data, it is very natural that any given data about customer churn is largely biased towards one class i.e., the non-churners which means that the non-churner class outweighs the churner class. Therefore, it is not unusual to have majority of instances in non-churner class and minority in the churner class.

In our data, the minority class (churner) accounts to 26.5% of the whole dataset. Here, majority instances are of non-churner class which has 5,174 instances and minority class has 1,869 instances. Figure 4.2 pictorially represents the share of the two classes in the dataset.

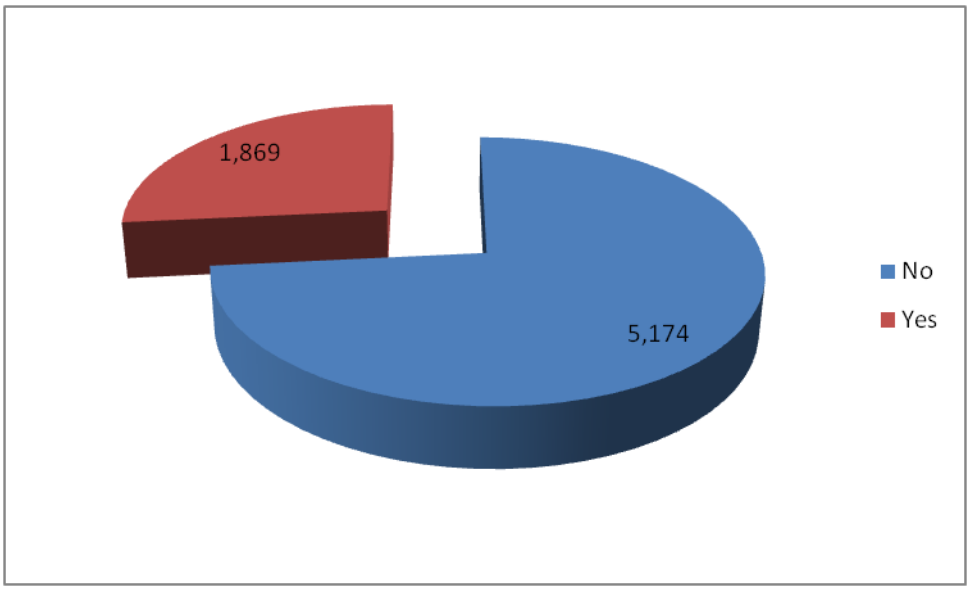

Figure 4.2: Share of churners and non-churners

Classification on class imbalanced data tends to be in favor of majority class as minority class samples are far less to adequately represent a class. So, RUS (random under-sampling) and ROS (random over-sampling) techniques were introduced to balance the classes. Random under-sampling and random over-sampling are conventional techniques that may not always prove to be effective in balancing the data along with introducing new appropriate minority instances into the data or removing some existing instances from majority class whichever might be the case. This is because random under-sampling and random over-sampling techniques are very basic and rudimentary. The three prominent sampling techniques are explained below.

a) Random Under-sampling technique

Random Under-sampling is the method of randomly deleting instances of majority class from the dataset to achieve a class balanced dataset. This method has its own limitations as it is a random act of selecting instances randomly and discarding them. It has no major idea on which instances should be discarded and which should be retained. This idea is just a hit-and-trial technique which may/may not work for a particular problem. Since, this technique is too naïve, it has the possibility that some of the useful instances may be discarded.

b) Random Over-sampling

Random Over-sampling technique is also based on random selection. In this method also, instances are selected randomly from the dataset and replicated. Therefore, the dataset now consists of replicated instances. This method is also too basic and does not exhibit good performance. Due to simple replication, learning on this type of data causes overfitting. In ROS, repeated instances cause decision boundary to tighten.

c) SMOTE based Sampling

As noted above, ROS and RUS techniques may cause overfitting [55] and deletion of useful instances respectively. Therefore, to overcome this technique, another sampling technique SMOTE (Synthetic Minority Over-sampling Technique) was developed. In this technique, minority class is over-sampled not by random selection of instances with replacement but by creating "synthetic" examples.

This approach was inspired from a technique used in handwritten character recognition. In this technique, additional training examples were created by performing certain operations on the data. Operations such as rotation and drift were considered to be natural ways of changing training data. SMOTE creates synthetic samples by operating in "feature space" instead of "data space". The minority class is over-sampled by taking each sample of the minority class and introducing synthetic samples along the line segments that join to any of the k nearest neighbors of the minority class. Depending on the amount of over-sampling required, the neighbors of the nearest neighbors are chosen at random. For example, if we take k = 5 and the required sampling quantity is 200%, then only two neighbors from five nearest neighbors are chosen and a sample is generated in the direction of each. The synthetic samples are generated as follows:

Take the difference between the feature vector (sample) under consideration and your nearest neighbor. Multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration. This makes the selection of a random point along the line segment between two specific characteristics. This approach effectively forces the minority class decision region to become more general.

SMOTE technique being simple and effective is one of the most adopted approaches for sampling. The SMOTE technique has been developed due to the fact that over-sampling causes over-equipping [57] due to repeated cases that cause the decision limit to be tightened. On the other hand, SMOTE creates similar examples. To the machine learning algorithm, these newly constructed examples are not duplicates of existing ones and thus soften the decision boundary. As a result, the classifier is more general and does not over-adjust. The dataset obtained through the process of SMOTE based balancing gives a better performance, as now the data is evenly distributed among the classes i.e., the churners and the non-churners which is bound to improve the classification score. This therefore extends enhanced learning to the base classifiers.

4.1.4 Feature Selection

Feature selection is a technique where appropriate and the most relevant features are selected and the least relevant that hold no or minimal meaning in predicting the dependent variable are discarded. Such variables, if kept in the feature space tend to bring down the accuracy and performance of the classifiers. The purpose of function selection is to eliminate features that are irrelevant to our problem or redundant. The elimination of these characteristics does not lead to the loss of information. There is a difference between redundant and irrelevant features and these are two different ideas as a relevant feature might as well be redundant with another relevant feature which is highly correlated with it.

Besides feature selection and feature extraction are two different concepts. Feature extraction is the process of generation of new features from the functions of original features. On the other hand, feature selection is drawing a subset of the original feature set. Feature selection techniques are generally implemented in applications where the datasets have large number of features and relatively fewer samples. Examples of archetypal cases of feature selection are domains such as DNA microarray data and written texts analysis. In these areas, there are thousands of features and few tens to hundreds of samples. Feature selection techniques are generally classified into three categories, as explained below:

1. Filter Methods

Filter type methods select features independent of the machine learning algorithm used. They are based on general methods such as the correlation of a feature with the class variable. Filter methods ignore the least important features. The other variables which are of higher importance will be a part of the classification or regression carried out to predict data. These methods are advantageous because of less computation time and robustness to overfitting. However, filter methods have a tendency to select redundant variables because they do not consider inter-variable dependencies. Therefore, they are mainly used as a preprocess method. Figure 4.3 illustrates the functioning of filter methods of feature selection.

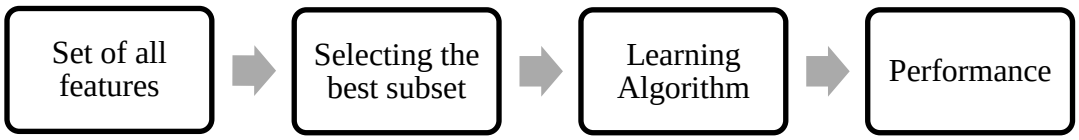

Figure 4.3: Filter Method

a) Pearson's correlation

The Pearson correlation measure is used to quantify the relationship between two linear variables that are linearly dependent X and Y. Its value ranges from -1 to +1. Pearson's correlation is given as:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{4.1}$$

b) LDA

Linear discriminant analysis is used to find a linear combination of characteristics that characterizes or separates two or more classes (or levels) of a categorical variable.

c) ANOVA

ANOVA stands for Analysis of variance. It works similarly to the LDA. It is measured between one or more independent categorical variables and a continuous dependent variable. It provides a statistical test of whether the media of several groups are equal or not.

d) Chi-square

It is a statistical test applied to groups of categorical variables to evaluate the probability of correlation or association between them using their frequency distribution.

Table 4.1 shows the techniques used for calculating feature dependence based upon the types (continuous and categorical) of features and response variables.

Table 4.1: Techniques used for calculating feature dependence based on feature and response types

Feature/Response	Continuous	Categorical
Continuous	Pearson's Correlation	LDA
Categorical	ANOVA	Chi-Square

1. Wrapper Methods

The wrapper methods use subsets of features and train a model using them. Based on the inferences extracted from the previous model, they add or eliminate characteristics of the subset. The problem is essentially a search problem. These methods are usually computationally very expensive. Some common examples of wrapping methods are the selection of advanced functions, the elimination of backward functions, the elimination of recursive traits, etc.

a) Forward Selection

Forward selection methods are iterative methods of feature selection. They start with empty sets and add a feature that best improves the performance of a learner. They keep

on adding new features until the addition of more features does not improve the model performance anymore.

b) Backward Elimination

Backward elimination starts with all the features in the feature set. They also work iteratively like forward selection. With each iteration, they remove the least significant feature to improve the performance of the models. This step continues until the performance of models does not improve anymore with the removal of features.

c) Recursive Feature elimination

It is a greedy optimization algorithm that aims to find the subset of features with better performance. It repeatedly creates models and keeps aside the best or worst performance function at each iteration. Build the next model with the functions on the left until all the features are exhausted. It then classifies features based on the order of their elimination. Figure 4.4 depicts the operation of wrapping methods for feature selection.

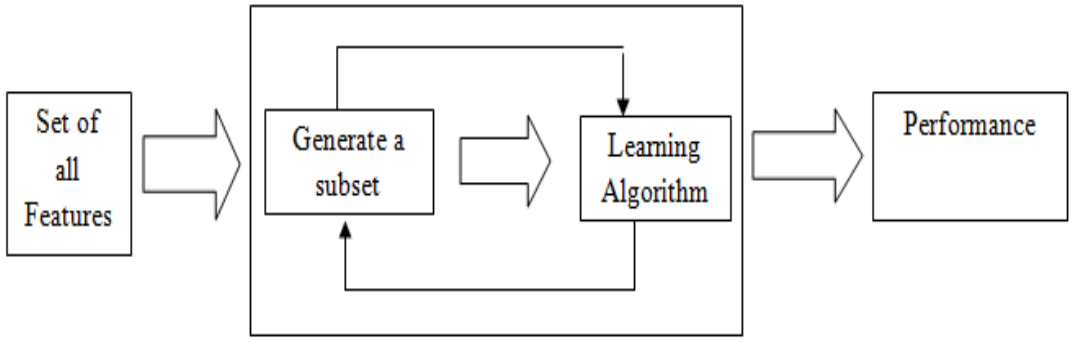

Figure 4.4: Wrapper method

2. Embedded Methods

Embedded methods are a combination of filter methods and wrapper methods. They have their own feature selection algorithms and are therefore called embedded methods. RIDGE and LASSO regression techniques are classic examples of embedded methods of feature selection which have inbuilt penalization functions to eliminate overfitting. Figure 4.5 shows the work flow of embedded methods for feature selection.

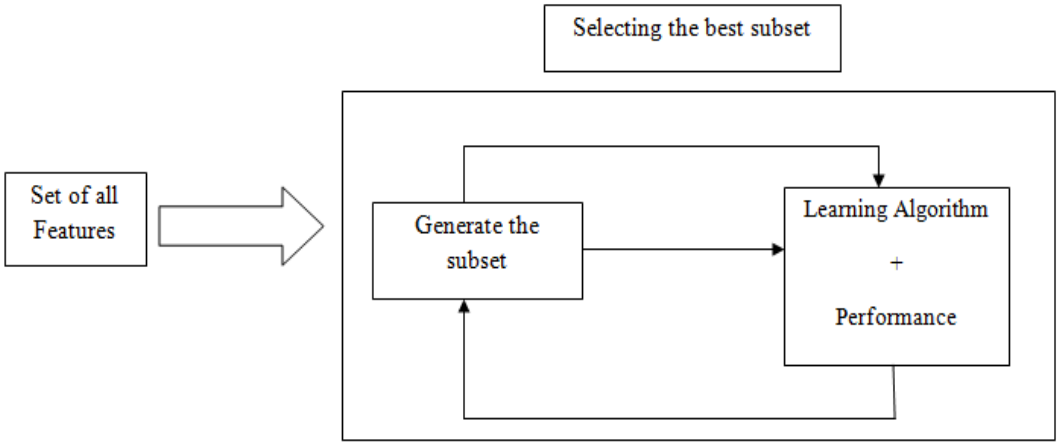

Figure 4.5: Embedded method

In our work, we have used 4 kinds of feature selection techniques. They are

1. Correlation Feature Selection

The co-relation based feature selection [58] method is a heuristic for measuring the worth of individual features for predicting the class label. There are mainly two techniques to find co-relation of an attribute with the class label. They are Pearson's Co-relation Coefficient and Spearman's Rank-Order Correlation. The attributes which show higher co-relation with the class label are selected and those with lower co-relation are discarded. In co-relation feature selection, features are rated according to the importance of features. In our work, Pearson's Correlation Coefficient has been used.

Pearson's Correlation

The Pearson correlation coefficient developed by Karl Pearson is also known as the Pearson product-moment correlation coefficient (PPMCC), Pearson's bivariate correlation coefficient is a statistical measure to define the linear correlation between two variables X and Y. Values between the range + 1 and -1. Here, +1 is total positive linear correlation, -1 is total negative linear correlation and 0 is no linear correlation. It is widely used in science.

The Pearson correlation coefficient is mathematically defined as the covariance of two variables divided by the product of their standard deviations. This definition implies the average (the first on the origin) of the product of the random variables adjusted to the mean, also denominated "moment of the product"; hence the modifier "product moment" in the name.

For a population

Pearson's correlation coefficient when applied to a population is commonly represented by the Greek letter ρ (rho) and can be referred to as the *population correlation coefficient* or the *population Pearson correlation coefficient*. The formula for ρ is:

$$\rho_{X,Y} = \frac{cov\,(X,Y)}{\sigma_X \sigma_Y} \tag{4.2}$$

where:

σ_X is the standard deviation of X

σ_Y is the standard deviation of Y

and,

$$\text{cov}(X,Y) = E[(X - \mu_X)\,(Y - \mu_Y) \tag{4.3}$$

where:

μ_X is the mean of X

μ_Y is the mean of Y

E is the expectation

For a sample

The Pearson correlation coefficient when applied to a sample is commonly represented by the letter r and can be called the sample correlation coefficient or the Pearson correlation of the sample. A formula for r can be obtained by substituting covariance estimates and variations based on a sample in the above formula.

So if there is one dataset $\{x_1,...,x_n\}$ containing *n* values and another dataset $\{y_1,...,y_n\}$ containing *n* values then the formula for *r* is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (4.4)$$

where,

n is the number of samples

x_i is i^{th} value of variable X

y_i is i^{th} value of variable Y

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and similarly for $\bar{y}$

Figure 4.6 exhibits scatter diagrams between x-axis and y-axis labels with different values of correlation coefficient (ρ). $\rho = -1$ signifies perfect inverse correlation between the variables. It shows a perfect downward slope meaning that the variables are indirectly correlated. Here, the data points fall completely on the line.

$-1 < \rho < 0$ signifies that the variables are somewhat indirectly correlated and not completely. The slope of the line is lesser as compared to the previous case. Also, the data points are slightly dispersed away from the line.

$\rho = +1$ signifies that the variables are completely correlated to each other. In such cases, one of the two variables can be discarded to avoid redundant features. The slope of the line is rising upwards and all the data points fall exactly on the line.

$0 < \rho < +1$ signifies that the variables are somewhat correlated and not completely as in the case above. The points are relatable to each other to a limit depending upon how close the value of ρ falls to +1. The data points here are slightly dispersed.

$\rho = 0$ is also an extreme case which signifies that the two variables are absolutely uncorrelated with one another. It shows that there is no similarity between these variables and cannot be compared whatsoever. Therefore, the data points here are also completely dispersed signifying no relation between the variables.

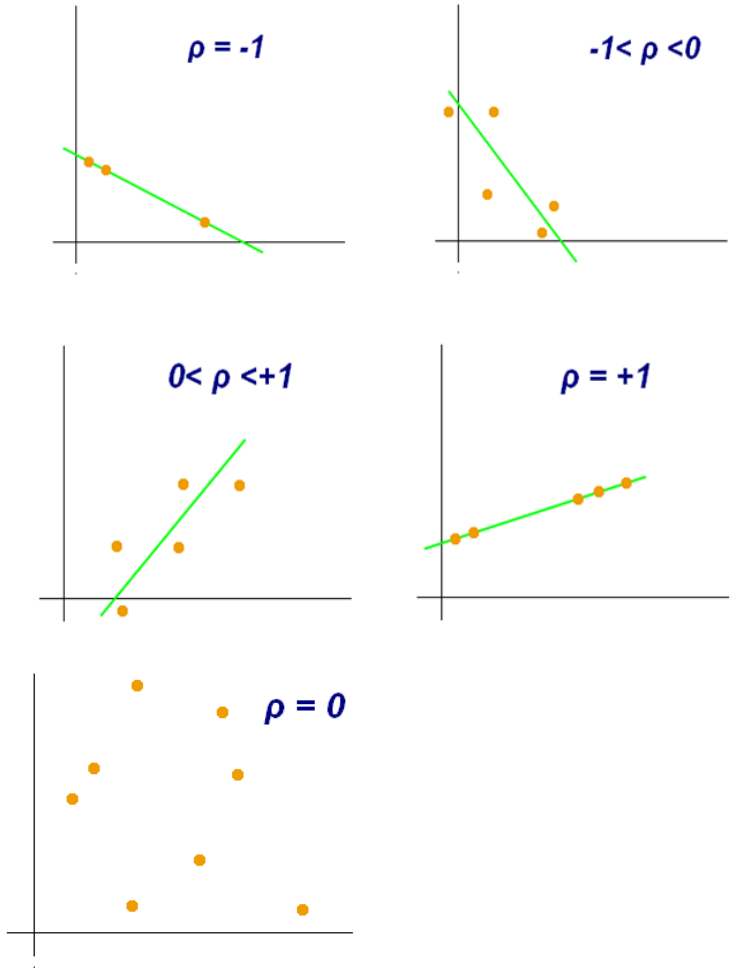

Figure 4.6: Examples of scatter diagrams with different values of correlation coefficient (ρ)

2. Information Gain Attribute Selection

IG (Information Gain) measures the entropy of a system that signifies the degree of disorder in a system. It is a criterion of goodness in the field of automatic learning. Therefore, the entropy of a subset is a fundamental calculation to calculate GI.

Let $\boldsymbol{Attr}$ be the set of all attributes and $\boldsymbol{Ex}$ the set of all training examples, $\boldsymbol{values}(\boldsymbol{x}, \boldsymbol{a})$ with $\boldsymbol{x} \in \boldsymbol{Ex}$ defines the value of a specific example $\boldsymbol{x}$ for attribute $\boldsymbol{a} \in \boldsymbol{Attr}$, $\boldsymbol{H}$ specifies the entropy. The $\boldsymbol{values}(\boldsymbol{a})$ function denotes set of all possible values of attribute $\boldsymbol{a} \in \boldsymbol{Attr}$. The information gain for an attribute $\boldsymbol{a} \in \boldsymbol{Attr}$ is defined as follows:

$$IG(Ex, a) = H(Ex) - \sum_{v \in values\ (a)} \left(\frac{|\{x \in Ex | value\ (x,a)=v\}|}{|Ex|} \cdot H(\{x \in Ex | value(x, a) = v\}) \right) \quad (4.5)$$

The information gain is equal to the total entropy of an attribute if for each of the attribute values a single classification can be made for the result attribute.

3. Gain Ratio Attribute Selection

A decision tree is a simple structure of nodes and edges wherein nodes represent test cases on attribute, the edges represent the various answers possible on the test nodes while the terminal nodes represent decision outcomes to these test cases. The IG measure [59] is used to select attributes for the terminal nodes of the decision tree. The IG measure has a limitation that it selects the attributes with large number of values. This limitation of the IG method is overcome by the Gain Ratio method. ID3 is a decision tree based on IG. C4.5 is an enhancement of ID3 that uses gain ratio which is an extension of IG measure aimed at improving ID3. The intrinsic value for a test is defined as follows:

$$IV(Ex, a) = -\sum_{v \in values\ (a)} \left(\frac{|\{x \in Ex | value\ (x,a)=v\}|}{|Ex|} \cdot \log_2 \left(\frac{|\{x \in Ex | value\ (x,a)=v\}|}{|Ex|} \right) \right) \quad (4.6)$$

The information gain ratio is given by:

$$IGR(Ex, a) = IG/IV \quad (4.7)$$

The measure of information gain causes a bias in the decision tree, since it considers attributes with a large number of different values. Therefore, the measure of the gain ratio

solves the drawback of the information gain, ie the gain of information applied to the attributes that can take a large number of different values. This can result in the decision tree learning the established training too well. For example, if you are going to build a decision tree for some data that contain the clients of a company. The information gain is used to find the most important attributes to test near the root of the tree. One of the attributes could be the credit card number of the customers. This attribute would have a high information gain because each value of this attribute is unique but cannot be included in a decision tree because deciding how to treat a customer based on their credit card number is unlikely to generalize unseen customers before.

4. OneR Attribute Selection

OneR, abbreviation for "One Rule", is a simple but accurate classification algorithm. It generates a rule for each of the predictor / input variables in the data. In the second step, select the rule with the lowest total error called "a rule". It constructs a frequency table for each feature against the target to create a rule for a feature. It has been seen that ”one rule” generates rules that are easy to interpret and also these rules do not exhibit very low classification accuracy as compared to other complex classification algorithms.

4.1.5 Model Building

In our work, we have used three different classifiers – CART (Classification and Regression Trees), Bagged CART (an ensemble of trees) and PART (Projective Adaptive Resonance Theory).

a) CART

CART programs build binary trees for classification and regression problems. These trees are referred to as models. These programs follow a two-stage procedure. The tree is constructed by the following process:

Step 1: First, you will find the only variable that best divides the data into two groups.

Step 2: The data is separated using this variable.

Step 3: Steps 1 and 2 are applied separately to each subgroup, and so on until the subgroups reach a minimum size or until they can not be improved.

Step 4: The second step of the procedure is to use cross-validation [60] to trim the entire tree. This step is necessary since the resulting model after the first stage is, with a certainty, too complex and the question regarding the criterion of unemployment arises as it does with all the step-wise procedures.

b) Bagged CART

Bagged CART or Bagged tree classification is a type of ensemble machine learning algorithm called Bootstrap Aggregation or Bagging [61]. Bagging is an ensemble technique which is both simple and powerful. An ensemble technique is an approach to combine predictions of two or more models to create a final prediction model that is more effective and has higher prediction accuracy than any of the individual models. The basic idea behind ensemble approach is to combine models that have very different predictions from one another to ensure that the each base model strives to overcome the weakness of another base model. The bagging technique has been invented to reduce the variance of those classifiers which have a high variance. Algorithms such as Classification and Regression Trees (CART) have been known to have high variance and low bias [62]. Decision trees are specific to the data on which it is trained which means that if they are trained on some other data then they may result in a completely different decision tree and in turn different predictions. Bagging is applied to machine learning algorithms having high-variance such as decision-trees. It is the application of the bootstrap procedure to such algorithms. Assuming we have to apply CART algorithm on a sample dataset of say, 1000 instances (x), bagging of the CART algorithm would work as follows:

1. Create many (e.g. 100) random sub-samples of the dataset with replacement.
2. Train a CART model on each sample.
3. Given a new dataset, calculate the average prediction from each model.

Bagging can be used for classification and regression problems just like decision trees.

c) PART

PART is a rule-based classifier. Rules learners are supervised learning algorithms. This algorithm induces a set of rules of the training instances. These rules then apply to test

data for classification. C4.5 and RIPPER classifiers are two well-known classifiers based on beginners. Both approaches take two steps to induce rules. The first step determines the initial set of rules and in the second step, these rules are adjusted or discarded according to a global optimization strategy. C4.5 generates an unprimed decision tree and generates rule sets from this decision tree. Each rule is simplified using a rule classification strategy. Finally, it rules out rules until the error rate of the set of rules in the instances does not increase. RIPPER implements a divide and conquer strategy for rule induction. Generates rules iteratively for a set of instances, and then removes those instances. In the second iteration, it generates rules for another set of instances. In an iterative way, new rules are derived for the remaining instances.

PART uses the strategy of dividing and conquering the RIPPER algorithm and the decision tree approach of C4.5. More exactly, PART generates a set of rules according to the strategy of divide and conquer, it eliminates all the instances of the formation collection that are covered by this rule and proceeds recursively until there is no case left. To generate a single rule, PART creates a partial decision tree for the current set of instances and selects the sheet with the largest coverage as the new rule. Subsequently, the partial decision tree is discarded which avoids early generalization.

4.1.6 Model Ensemble

After performing the above steps, we have used ensemble approach to investigate the performance of the ensemble technique. Ensemble approach in machine learning is a learning approach where we try to combine two or more learners together to see if the combination improves results. There are three types of ensemble methodologies. They are:

1. Bagging

Given a standard training set D of size n, bagging generates m new training sets D_i, each of size n' by sampling from D uniformly and with replacement. By sampling with replacement, some observations may be repeated in each D_i. If $n'=n$, then for large n the set D_i is expected to have the fraction (1 - 1/e) (≈63.2%) of the unique examples of D, the

rest being duplicates. This kind of sample is known as a bootstrap sample. The *m* models are fitted using the above *m* bootstrap samples and combined by averaging the output (for regression) or voting (for classification).

Bagging can improve the performance of "unstable procedures" to a large extent including artificial neural networks, classification and regression trees, and selection of subsets in linear regression. Figure 4.7 illustrates the operation and focus of the basic bagging algorithm for set.

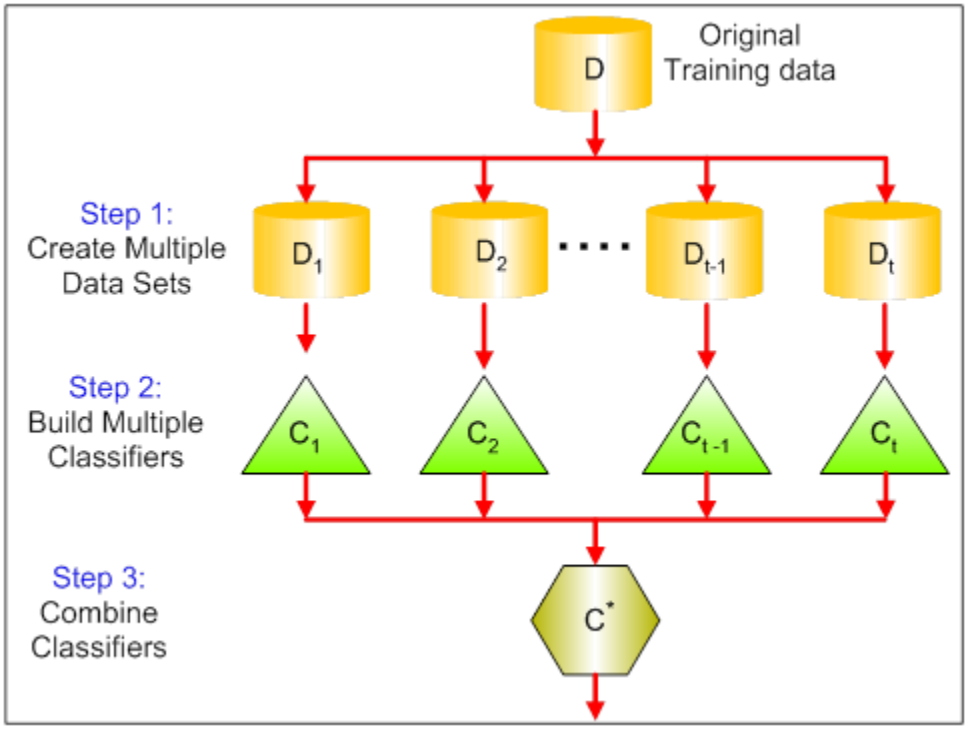

Figure 4.7: Illustration of basic bagging technique

2. Boosting

Boosting is a machine learning ensemble meta-algorithm for reducing bias and variance. It is a form of supervised learning which converts a group of weak learners to produce a strong learner. The boosting algorithms work in three steps.

Step 1: In the first step, base learners consider all the data points and assign equal weight to all the data points.

Step 2: If the base learner classifies some instances incorrectly then higher weight is assigned to those misclassified instances and another base classifier is applied.

Step 3: Repeat Step 2 till the limit of the base learner is attained or higher accuracy is achieved.

Therefore, boosting creates an ensemble of models by incrementally increasing the weights of the appropriate instances according to the error rate of the instances in the previous prediction. At each iteration, the algorithm identifies the points which have been misclassified and updates their weights in the next iteration. It trains each model with the same dataset. The main idea is to force the models to focus on tough instances. Boosting is a sequential operation so unlike bagging parallel operations cannot be used here. Figure 4.8 illustrates the approach of the boosting technique of ensemble method by increasing the weights of misclassified samples in previous iteration.

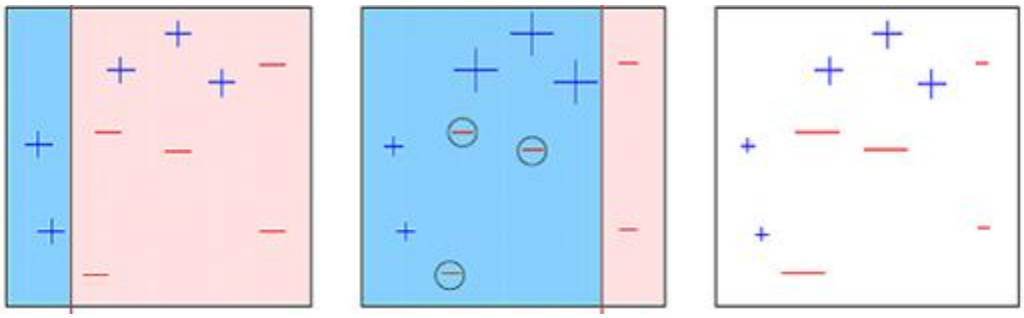

Figure 4.8: Illustration of boosting technique

3. Stacking

Stacking, or stacked generalization [63] is an ensemble approach where classifiers are combined together. Initially, base classifiers are used to train models and then another model is used to train a new model called the combiner algorithm. The basic idea is to train machine learning algorithms with training dataset and then the predictions of these base models are used to create a new dataset. Then this new dataset is used to train the combiner learning algorithm. Figure 4.9 shows the approach of the stacked generalization technique.

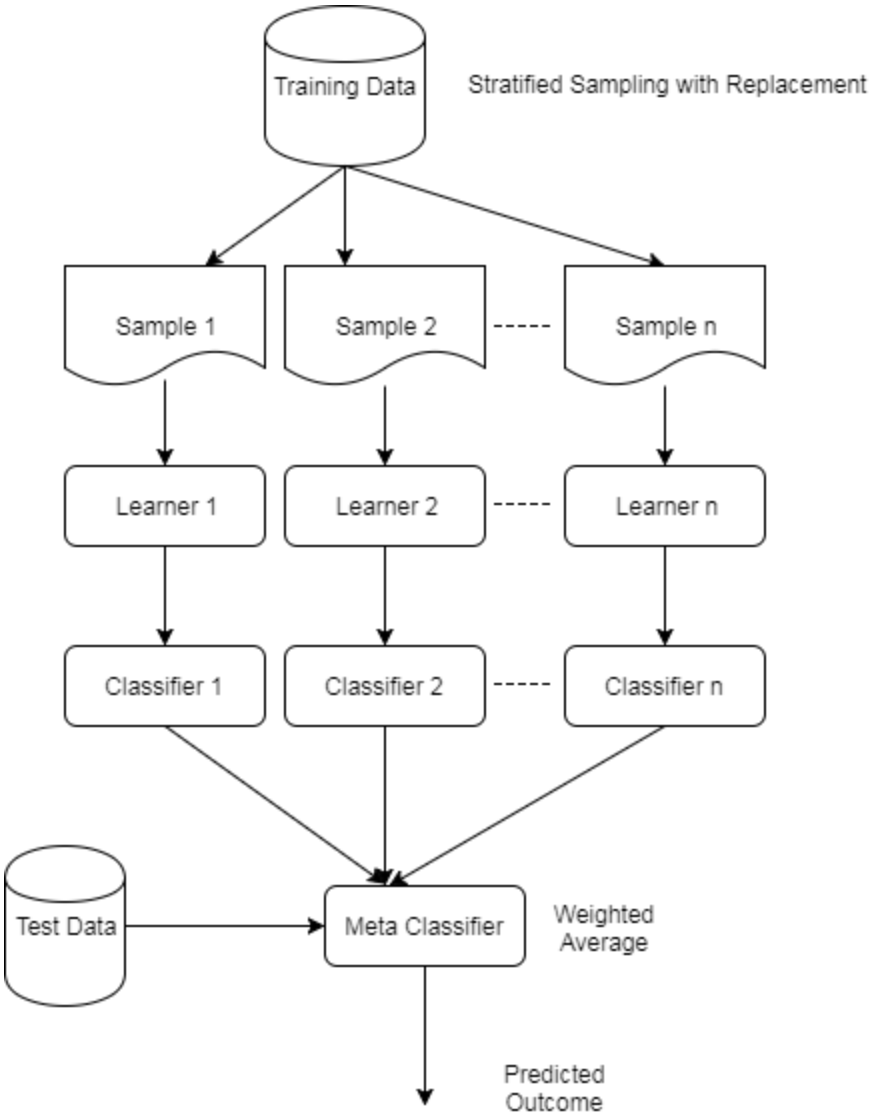

Figure 4.9: Stacked Generalization method

In our work, we have used the stacking ensemble approach to study its effect on the accuracy and other performance measures. To ensure that training and testing is performed on different sets of instances we have partitioned the dataset into three sets. All the three partitions consist of roughly the same number of instances. The partitions are called ensemble data, blender data and testing dataset. We have separately trained the ensemble data with each of the base classifiers – decision tree, bagged tree and the rule-based classifier. Afterwards, the learned classifiers are made to make predictions upon new set of instances called the blender data. The predictions of these models are stored in

the blender data as well as the testing data and three additional features are added to the original dataset. This way the dataset on which new learner will be applied contains not only the original attributes but the predictions made by each of the three base classifiers are also taken into consideration. For the rest of the process, the new predictors are a combination of the old predictors and the predictions of the base learners.

In the next step, blender data is trained again with a new classifier. This leads to another learned model that is further used to predict the outcome on testing data. This prediction made by the stacked model is the final prediction. The results of these predictions are further used to assess the prediction accuracy achieved by our stacked model if any. If these predictions show in some manner that some of the performance measures have improved results, then we can say that the stacked ensemble approach is useful in predicting telecommunication churn, in general with satisfactory growth in performance than either of the base classifiers could perform even with appropriate preprocessing, sampling and feature selection.

We have used the AdaBoost.M1 classifier as the classifier that learns and builds a novel model upon the stacked data.

AdaBoost Algorithm

AdaBoost stands for Adaptive Boosting. It works on the lines of boosting algorithms. In this a number of weak classifiers are learned on training data which is differently weighted. In the beginning, it predicts on original data and all observations are assigned equal weight. If there are observations that have been predicted incorrectly, then it updates weights of those observations and assigns higher weight to such observations. As this process is iterative, these steps continue to build new models until the limit of number of learners is exceeded or desired accuracy is achieved.

The most used models on AdaBoost algorithm are decision stumps but it accepts any learner as base learner if the learner allows to assign weight on training data. AdaBoost algorithm can be used for classification as well as regression problems. In Figure 4.10, we clearly explain the methodology on how AdaBoost algorithm works. The working is as follows:

Box 1: In Box 1 in figure 4.11, each data point has equal weight and we have applied a decision boundary to distinguish the plus(+) and the minus(-) points. The decision boundary is D1 which has been created on the left side of the box as a vertical line. But D1 wrongly classifies three plus(+) data points as minus(-) data points. Due to this misclassification, higher weight is assigned to these three plus(+) points and another decision stump is applied.

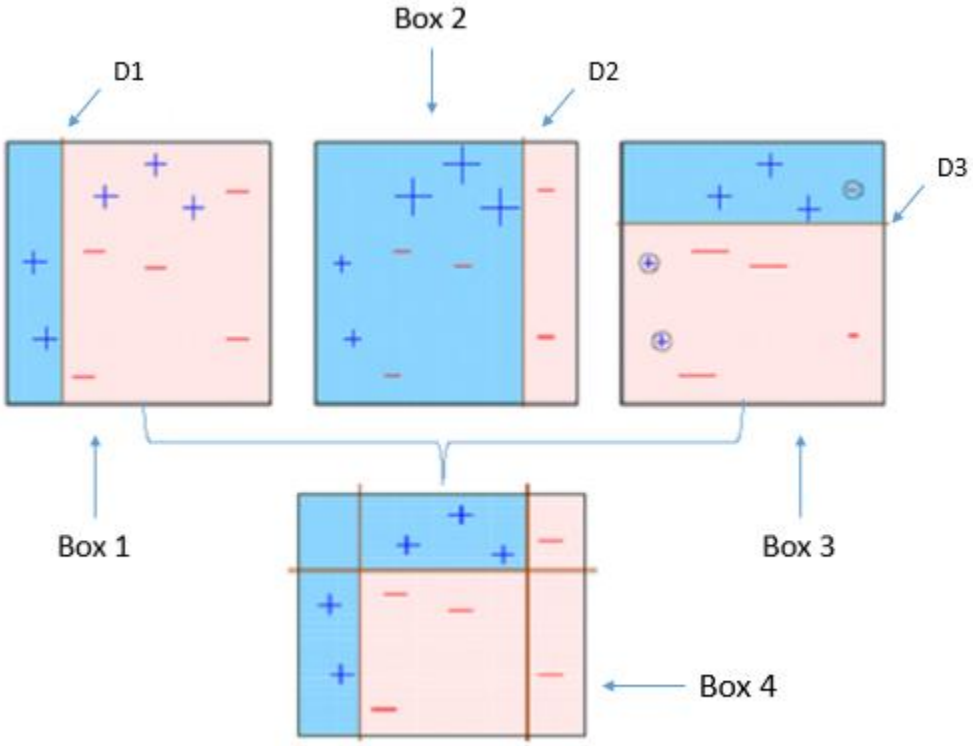

Figure 4.10: Illustration of working of AdaBoost Algorithm

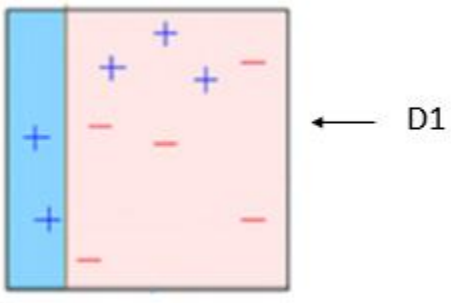

Figure 4.11: AdaBoost Algorithm - Decision Stump D1

Box 2: In Box 2 in figure 4.12, three incorrectly predicted plus (+) points from Box 1 are now bigger in size because we have increased their weights. Now we apply another decision boundary called D2 which will try to classify these points correctly. This time the decision boundary is placed vertically on the right side of the box. This leads to misclassification of the three minus (-) points as plus (+) points on the left side of the box. Therefore, we assign higher weight to these misclassified minus (-) points and then apply another learner.

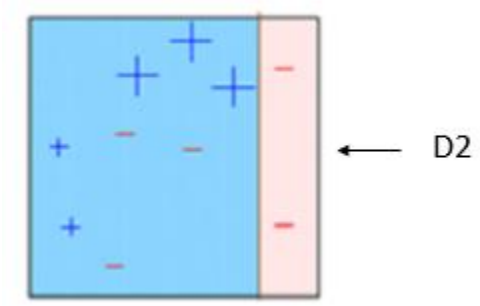

Figure 4.12: AdaBoost Algorithm - Decision Stump D2

Box 3: In Box 3 in figure 4.13, we can see that the three misclassified points from Box 2 are assigned higher weight than the rest of the instances. A new decision boundary D3 is applied to classify again those points. This leads to a new horizontal boundary called D3 on the upper side of the box.

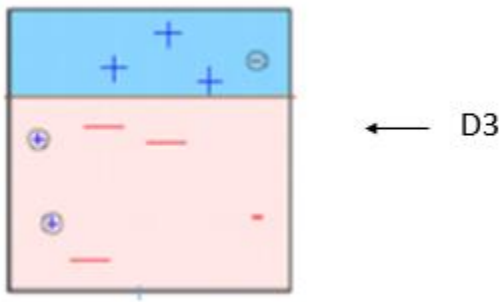

Figure 4.13: AdaBoost Algorithm - Decision Stump D3

Box 4: In Box 4 in figure 4.14, D1, D2 and D3 have been combined to form a strong prediction having a complex rule as compared to individual weak learners. It can be seen

that this algorithm has classified these observations satisfactorily as compared to any of the individual weak learners.

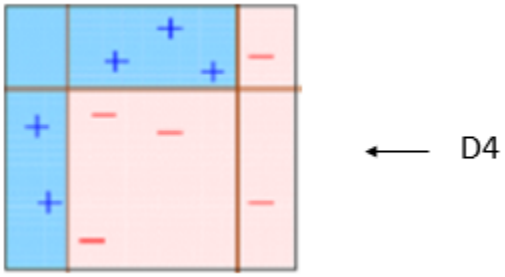

Figure 4.14: AdaBoost Algorithm - Decision Stump D4

4.1.7 Evaluation Parameters

The evaluation parameters that we have used in our work are AUC(Area Under the Curve), Sensitivity, Specificity, and Accuracy.

Concept of Confusion Matrix

1. A classifier predicts 0 and the class label is actually 0: this is called a **True Negative**, i.e. classifier correctly predicts the class as negative (0). For example, a non-churner is detected as non-churner.
2. A classifier predicts 0 while the class label is actually 1: this is called a **False Negative**, i.e. classifier incorrectly predicts the class as negative (0). For example, a churner is detected as a non-churner.
3. A classifier predicts 1 while the class label is actually 0: this is called a **False Positive**, i.e. classifier incorrectly predicts the class as positive (1). For example, a non-churner is detected as churner.
4. A classifier predicts 1 and the class label is actually 1: this is called a **True Positive**, i.e. classifier correctly predicts the class as positive (1). For example, a churner is detected as churner.

In general, Positive = identified and negative = rejected. Therefore:

- True positive = correctly identified
- False positive = incorrectly identified
- True negative = correctly rejected
- False negative = incorrectly rejected

Figure 4.15 shows a confusion matrix and types of errors.

		predicted condition	
	total population	prediction positive	prediction negative
true condition	condition positive	**True Positive (TP)**	**False Negative (FN)** (type II error)
	condition negative	**False Positive (FP)** (Type I error)	**True Negative (TN)**

Figure 4.15: Confusion Matrix

a) Sensitivity

Sensitivity, or recall or true positive rate measures the proportion of positives that are correctly classified or the percentage of churners which are correctly classified as churners. It can be expressed as:

$$sensitivity = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + number\ of\ false\ negatives} \tag{4.8}$$

$$= \frac{number\ \ of\ true\ \ positives}{total\ \ number\ \ of\ churners\ \ in\ the\ data} \tag{4.9}$$

$$= probability\ of\ a\ positive\ prediction\ given\ that\ a\ customer\ is\ a\ churner \tag{4.10}$$

b) Specificity

Specificity, or true negative rate measures the proportion of negatives that are correctly classified such which means the percentage of non-churners which are classified as non-churners. It can be mathematically defined as:

$$specificity = \frac{number\ of\ true\ negatives}{number\ of\ true\ negatives + number\ of\ false\ positives} \tag{4.11}$$

$$= \frac{number\ \ of\ true\ \ negatives}{total\ \ number\ \ of\ non-churners\ \ in\ the\ data} \tag{4.12}$$

$$= probability\ of\ a\ negative\ prediction\ given\ that\ a\ customer\ is\ a\ non\ churner \tag{4.13}$$

Figure 4.16 and figure 4.17 illustrate the sensitivity and specificity parameters.

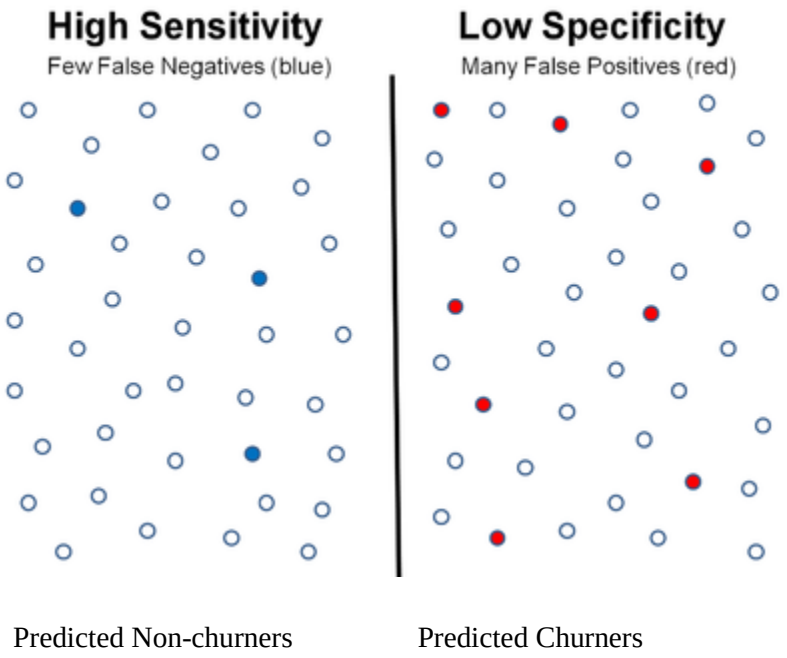

Figure 4.16: High sensitivity and low specificity

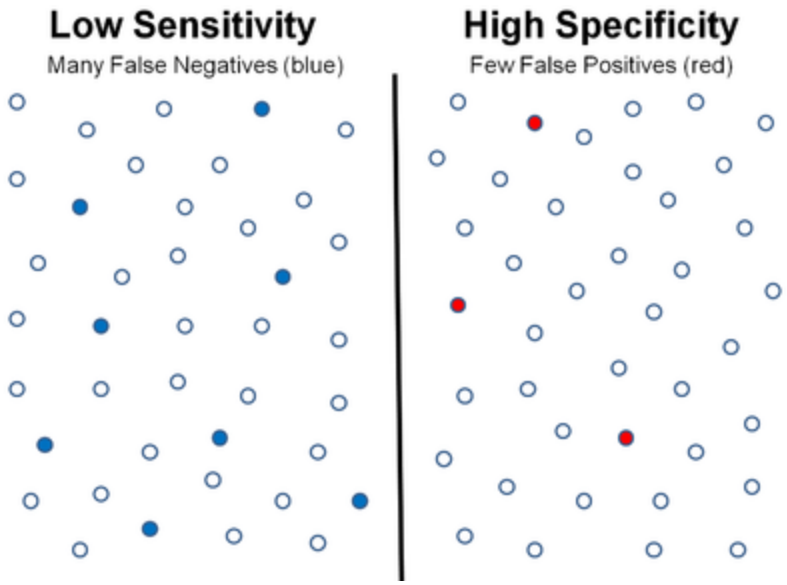

Figure 4.17: Low sensitivity and high specificity

c) AUC

AUC or area under the curve is a measure used in binary classification problems. It is used to determine the goodness of a binary classifier. It determines which of the classifiers learns the best.

ROC curves are an example of application of AUC. It is a plot of false positive rate against true positive rate. The AUC is measured on a scale of 0.5 to 1.0 with 0.5 being the lowest possible area under the curve a model can achieve while an AUC score of 1.0 signifies an excellent model and it is the best score achievable. Generally, an AUC of 1.0 is not possible. An example for an AUC curve is shown in figure 4.18. As can be noted, AUC is a plot between false positive rate (100- Specificity) and true positive rate (Sensitivity). Therefore, it can be said that AUC increases with increase in true positive rate and decrease in false positive rate. So, in worst scenario, true positive rate will be equal to the false positive rate which would give an AUC of 0.5. So, models with higher AUCs are preferred over those with lower AUCs.

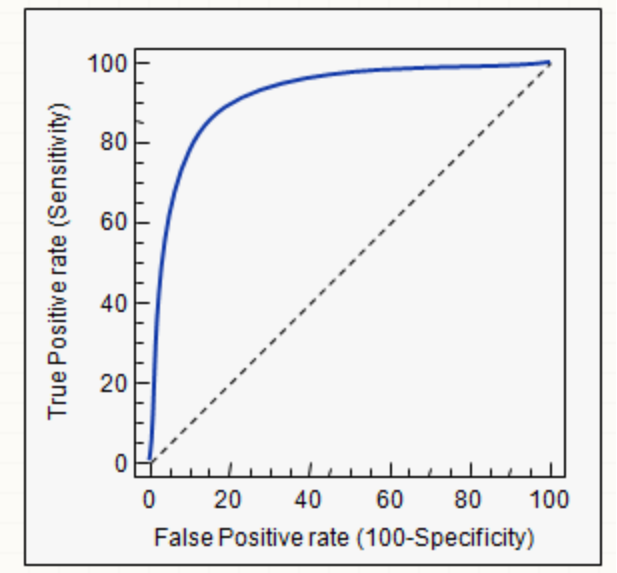

Figure 4.18: An AUC Curve

An AUC score of 0.5 is no better than guessing at random. An AUC score of 0.9 suggests a very good model, but a score of 0.9999 would be too good to be true and will indicate overfitting.

d) Accuracy

Accuracy is a measure of classification goodness which signifies the percentage of correct classifications made i.e., the total number of true positives and true negatives out of the total population. The best accuracy measure is 1.0 and worst is 0.0. Accuracy can be expressed as:

$$Accuracy = \frac{number\ \ of\ true\ \ positives + number\ \ of\ true\ \ negatives}{total\ \ population} \tag{4.14}$$

$$= \frac{number\ \ of\ correct\ \ predictions}{total\ \ population} \tag{4.15}$$

4.2 Technologies Used

4.2.1 R Programming Language

Implementation of proposed language is done in R programming language. R is a language and environment for statistical computing and graphics. It is an open source programming language. R has a wide variety of statistical and graphical techniques and is very extensible. The available techniques in R are classification and grouping, time series analysis, linear and nonlinear modeling and classic statistical tests. R is an environment commonly used for statistical computing, data analysis and scientific research. It is widely used by statisticians, researchers, data analysts and marketers to retrieve, clean, analyze, visualize and present data. It has been used for our work because of its easy-to-use interface and expressive syntax.

R is an integration of facilities such as data manipulation, calculation and graphical visualization. It includes:

- An effective data storage and handling capacity.
- A large, integrated and coherent collection of data analysis tools
- Graphical installations for the analysis of data
- A simple, effective and well-developed programming language that includes conditionals, loops, user-defined recursive functions, and input and output facilities.

Some business intelligence and big data tools that integrate with R are:

- Mongo DB
- Microstrategy
- SAP
- Azure
- Spark
- Hortonworks
- Hadoop
- Hp-vertica, etc

Figure 4.19 illustrates the key features and advantages of R.

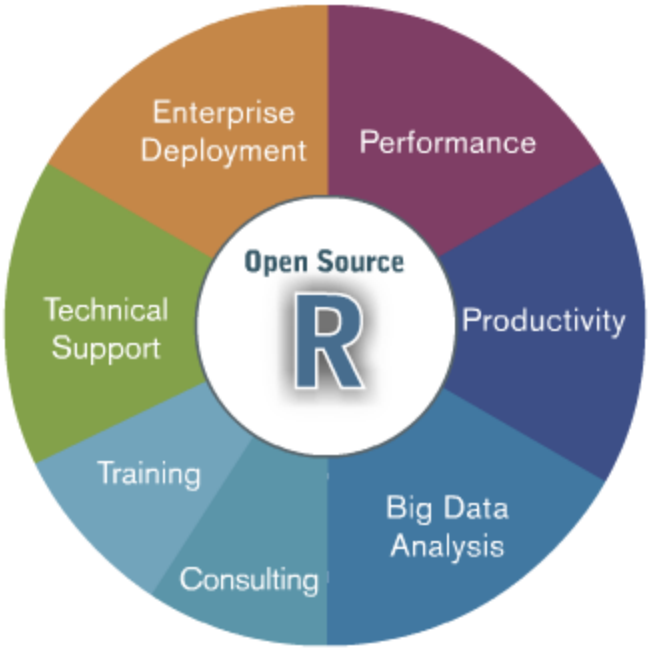

Figure 4.19: R Language

4.2.2 RStudio

RStudio is a free and open source integrated development environment (IDE) for R, a programming and statistical computing environment. RStudio was found by JJ Allaire. RStudio is written in C ++ programming language and uses the Qt framework for its graphical user interface. It can be run locally as a regular desktop application as well as RStudio server wherein we are allowed to access RStudio using a web browser.

Some of the features of RStudio are:

- **It is an IDE that is built just for R**
  - It helps with its useful features such as code completion, syntax highlighting and smart indentation
  - It executes R code directly from the source editor
  - It can quickly jump to function definitions
- **It brings the workflow together**
  - it integrates R help and documentation
  - it easily manages multiple workspace directories using projects

- it has easy-to-use feature such as workspace browser and data viewer

- **It has excellent features such as powerful authoring and debugging**
  - It consists of an interactive debugger that can diagnose and fix errors quickly
  - It boasts of extensive package development tools
  - Authoring with Sweave and R Markdown

Chapter 5

Experimental Results

Initially, the unprocessed dataset is used to deploy the various combinations of feature extraction and classification methods. Then, similar experimentation is conducted upon the processed dataset. The performances of various techniques applied into the study are compared as to performance they display on processed as well as unprocessed dataset. The testing is done on 10-fold cross validation and performance measures used are AUC, sensitivity and specificity.

5.1 Performance Analysis after Basic Preprocessing

Initially, CART, Bagged CART and PART are implemented on original dataset without applying any sampling or feature reduction strategy. Table 1 shows, all three classifiers show poor performance in terms of specificity, sensitivity and AUC.

The original unprocessed dataset contains 7,074 records and 21 attributes. In basic preprocessing the only optimization done is to discard useless features and instances with missing values have been treated. Table 5.1 shows the performance of the base classifiers (CART, Bagged CART and PART) based on four parameters (Accuracy, sensitivity, specificity and AUC). Figure 5.1 shows the graph projecting the AUC scores of the base classifiers upon the preprocessed dataset. The performance of PART classifier is greatest with an AUC of 0.6851. Bagged CART follows close with an AUC of 0.6603 while the CART shows least AUC of 0.6063.

Table 5.1: Performance of base classifiers on preprocessed dataset

Classifier/Evaluation Parameters	CART	Bagged CART	PART
Accuracy	0.7643	0.7619	0.7686
Sensitivity	0.275	0.441	0.507
Specificity	0.937	0.88	0.863
AUC	0.6063	0.6603	0.6851

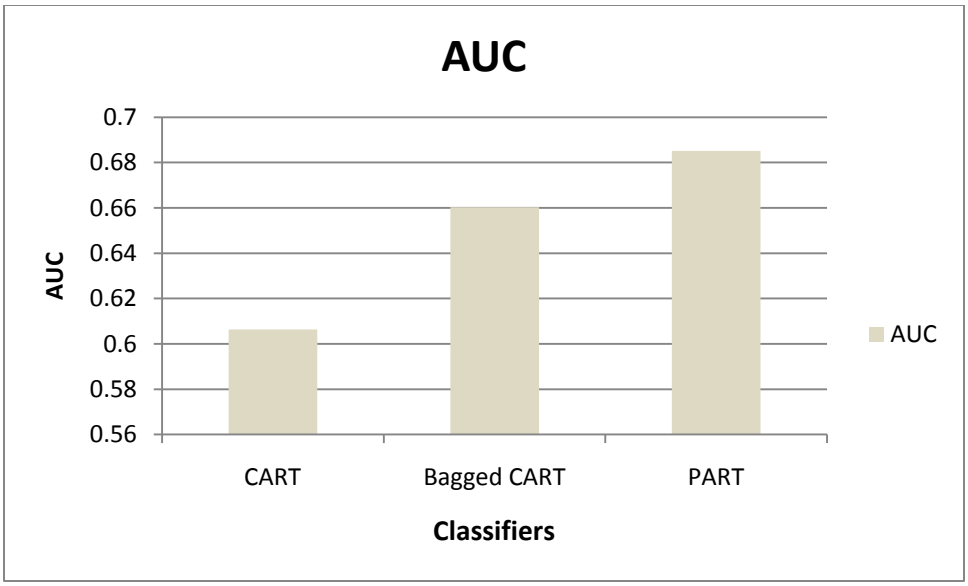

Figure 5.1: scores of base classifiers on preprocessed dataset

5.2 Performance based on SMOTE Sampling Technique

SMOTE (Synthetic Minority Over-Sampling Technique) is a widely adopted technique due to its simplicity and effectiveness. The SMOTE method for data sampling is a combination of under-sampling and over-sampling. The over-sampling here is not random over-sampling but by generating new examples from the existing ones using an algorithm. KNN algorithm is used to generate new samples. It studies the existing samples and re-frames them to discover new possibilities that the data may hold.

In traditional oversampling, the minority class reproduces exactly. In SMOTE, the new instances of minorities are constructed as follows:

1. For each instance of minority class c and number of neighbors k
2. Neighbors = Get KNN (k)
3. At random choose one of the neighbors
4. Create a new minority class instance r using the feature vector of c and the feature vector's difference of n and c multiplied by a random number

$$\text{i.e.,r.feats} = \text{c.feats} + (\text{c.feats} - \text{n.feats}) * \text{rand}(0,1) \quad (5.1)$$

All three classifiers show improvements in AUC. CART, Bagged CART and PART achieve 0.6445, 0.6998 and 0.7139 AUC values respectively, which show that there

definitely has been an improvement in performance due to the sampling technique used. The dataset obtained after the SMOTE based sampling evenly justifies the presence of both the classes. Now the minority class instances have not been suppressed by the majority class instances. Due to this an improved training level was possible by the used classifiers. This was possible because of KNN classifier involved in the instance generation of SMOTE based sampling. Therefore, the SMOTE based sampling serves the idea of achieving balanced data for better performance.

Table 5.2 shows the performance of base classifiers on the dataset after performing data sampling. Figure 5.2 shows the graph of AUC scores obtained by the base classifiers on the dataset.

The AUC of CART has increased by 0.0382 from 0.6063 to 0.6445. Bagged CART has an increase in the accuracy by 0.0395 from 0.6603 to 0.6998. The AUC of PART has increased by 0.0288 from 0.6851 to 0.7139. Therefore, the AUC of classifiers on sampled data has considerably increased from that on original dataset after preprocessing.

Table 5.2: Performance of base classifiers on sampled data

Classifiers/ Evaluation Parameters	CART	Bagged CART	PART
Accuracy	0.7661	0.7690	0.7701
Sensitivity	0.259	0.410	0.436
Specificity	0.940	0.891	0.923
AUC	0.6445	0.6998	0.7139

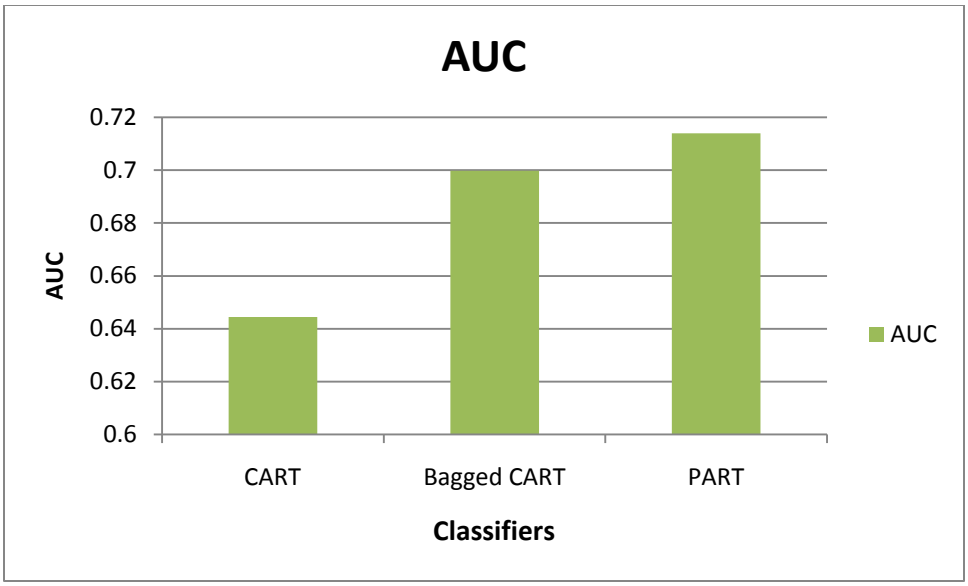

Figure 5.2: AUC scores of base classifiers on sampled data

5.3 Performance Analysis after Feature Selection

It was observed that after feature selection, the performance of the individual models has improved substantially. The AUC of CART on sampled dataset is 0.6445 and that after applying co-relation is 0.6581. The AUC shows an improvement of 0.0136 on sampled data after applying Co-relation feature selection. The AUC increased by 0.0204 from 0.6998 to 0.7202 after applying co-relation on Bagged CART classifier. AUC improved from 0.7139 to 0.7243 after applying co-relation on sampled data for PART classifier. Overall, we observe that co-relation feature selection performs much better than Gain Ratio, Information Gain and OneR feature selection methods.

Table 5.3 shows the performance of classifiers based on four parameters after correlation feature selection on sampled data. Table 5.4 shows the performance of classifiers after gain ratio based feature selection on sampled data. Table 5.5 shows performance of classifiers after OneR based feature selection. Likewise, table 5.6 shows the performance of classifiers after IG based feature selection.

Table 5.3: Performance of classifiers after correlation feature selection

Classifier/ Parameter	Accuracy	Sensitivity	Specificity	AUC
CART	0.7922	0.401	0.933	0.6581
Bagged CART	0.7709	0.465	0.882	0.7202
PART	0.7638	0.516	0.921	0.7243

Table 5.4: Performance of classifiers after gain ratio based feature selection

Classifier/ Parameter	Accuracy	Sensitivity	Specificity	AUC
CART	0.7492	0.205	0.950	0.6317
Bagged CART	0.7473	0.243	0.929	0.7024
PART	0.7738	0.519	0.866	0.6973

Table 5.5: Performance of classifiers after OneR based feature selection

Classifier/ Parameter	Accuracy	Sensitivity	Specificity	AUC
CART	0.7553	0.266	0.932	0.6373
Bagged CART	0.7515	0.185	0.954	0.6993
PART	0.7686	0.503	0.868	0.7172

Table 5.6: Performance of classifiers after IG based feature selection

Classifier/ Parameter	Accuracy	Sensitivity	Specificity	AUC
CART	0.7440	0.574	0.803	0.645
Bagged CART	0.7530	0.519	0.850	0.7181
PART	0.7406	0.527	0.838	0.7035

Table 5.7 shows the performance comparison of AUC scores obtained by base classifiers after preprocessing, sampling and feature selection.

Table 5.7: The performance comparison of AUC scores obtained by CART, Bagged CART and PART

Classifiers/ Techniques	CART	Bagged CART	PART
Basic Preprocessing	0.6063	0.6603	0.6851
Sampled Data	0.6445	0.6998	0.7139
Correlation	0.6581	0.7202	0.7243
Gain Ratio	0.6317	0.7024	0.6973
OneR	0.6373	0.6993	0.7172
Information Gain	0.645	0.7181	0.7035

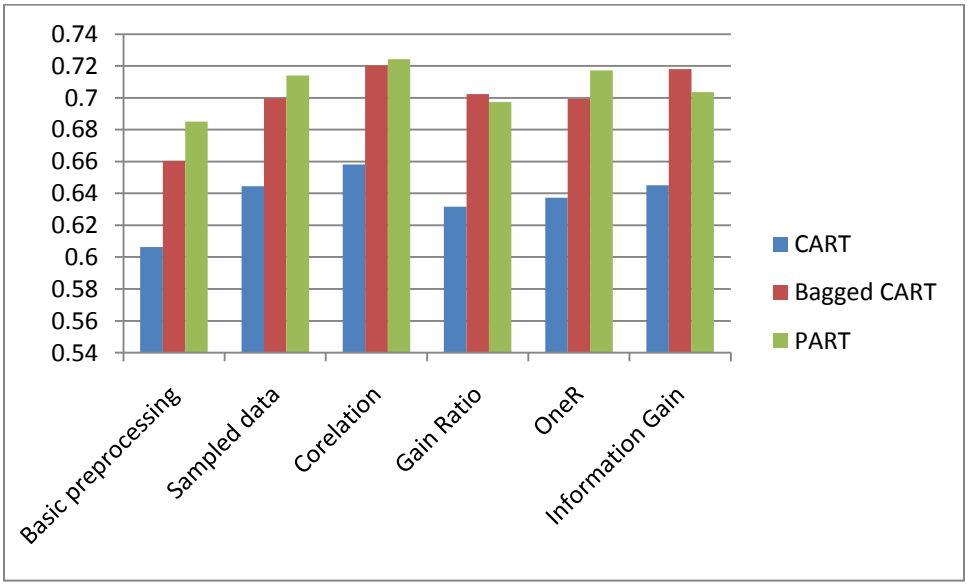

Figure 5.3: Performance comparison of base classifiers based on AUC scores

5.4 Performance Analysis of Ensemble Method

In the final step, we have generated model ensemble after applying SMOTE based sampling, correlation feature selection, and using the adaboost algorithm for training the final data that we are considering for ensembling after the application of sampling and feature selection technique. For ensemble model, we have used the stacking approach.

Table 5.8 shows the performance of ensemble model after preprocessing, sampling and feature selection phases. From the results, we observe that the AUC of final model is 0.7788 which is 0.0607 more than the highest AUC 0.7181 attained by any of the base classifiers and techniques used on the dataset in previous experiments. We have not seen much improvement in the accuracy parameter. The highest accuracy attained in the previous experiments is 0.7922 and that of the final model is 0.7940. There is a shift of 0.0018 in the accuracy which is negligible. Table 5.9 shows the AUC scores of base classifiers and ensemble model after preprocessing, sampling and feature selection.

Table 5.8: Performance of Ensemble model after preprocessing, sampling and feature selection

Accuracy	Sensitivity	Specificity	AUC
0.7940	0.482	0.869	0.7788

Table 5.9: AUC scores of classifiers after preprocessing, sampling and correlation feature selection

CART	Bagged CART	PART	Ensemble
0.6581	0.7202	0.7243	0.7788

The graph shown in figure 5.4 is a comparison of AUC scores obtained by the base classifiers as well as the ensemble model. It can be noted that that the AUC score obtained by the ensemble model is far better than those obtained by any of the classifiers. The AUC score of ensemble model is 0.7788 which is more than the AUC score obtained by the best classifier i.e., PART. The AUC of PART classifier being 0.7243, there is a sharp increase of 0.0545 in the AUC after performing ensemble.

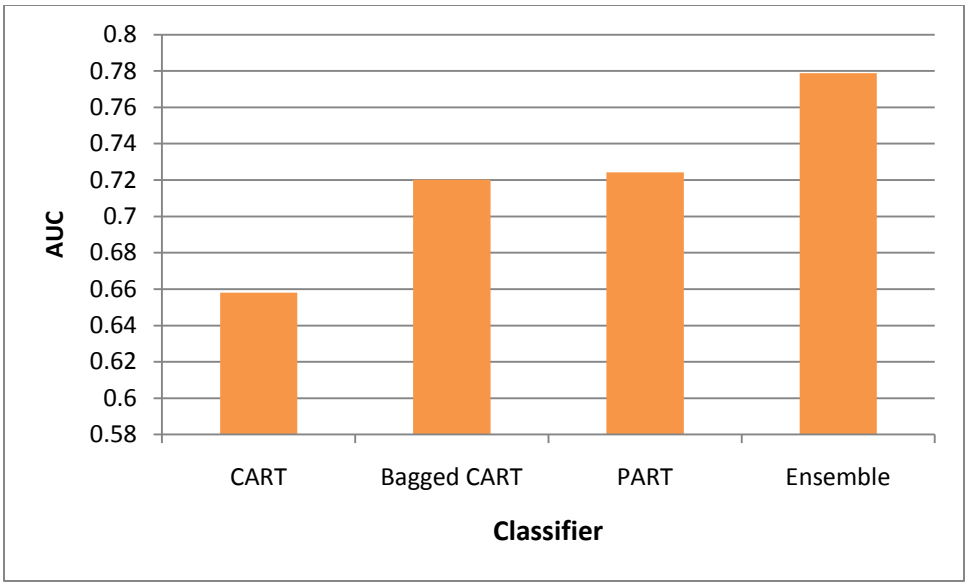

Figure 5.4: AUC scores of classifiers after sampling and correlation feature selection

Table 5.10: Comparison of best scores obtained before ensemble and scores obtained after ensemble

	Accuracy	Sensitivity	Specificity	AUC
Best scores before ensemble	0.7922	0.185	0.954	0.7243
Ensemble scores	0.7940	0.482	0.869	0.7788

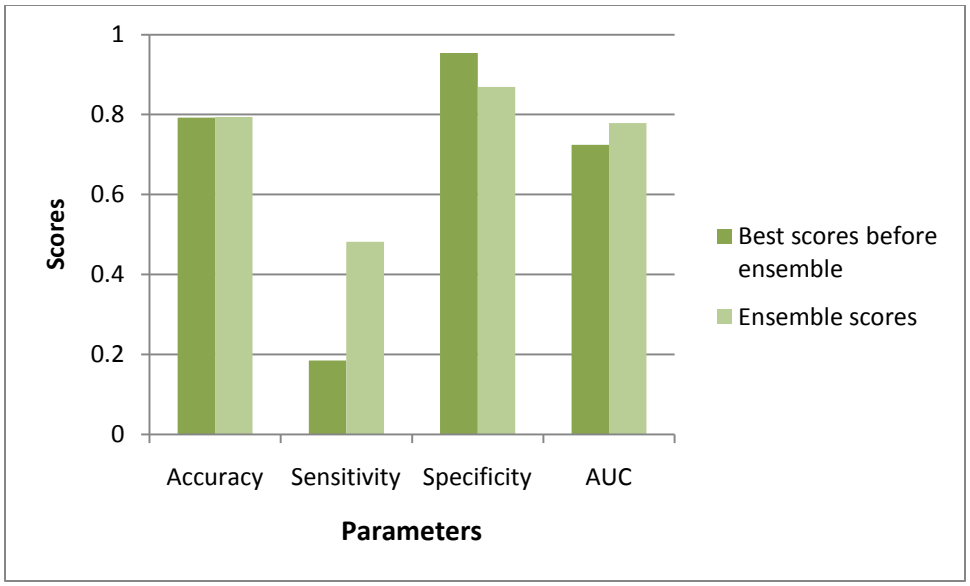

Figure 5.5: Comparison of best scores before and after stacking

Table 5.10 gives the comparison of best scores obtained by any of the base classifiers in terms of accuracy, sensitivity, specificity and AUC to the scores obtained by the ensemble model. The AUC is the most improved parameter and it gives a clear indication that AUC scores have drastically improved using our technique. Sensitivity has increased from the best score of 0.185 before ensemble to 0.482 after ensemble approach. The sensitivity or probability of detection has clearly increased which is a desirable quality. specificity, or true negative rate has reduced from 0.954 to 0.869 which means that the probability of identifying non-churners as such has fallen by 0.085 which does not make much of a difference to our system as misclassifying some non-churners as churners is not an issue in our problem but vice-versa is. Accuracy of our system before and after ensemble is more or less the same. Previously, the accuracy was 0.7922 and that after ensemble is 0.7940. So, there is an increase of 0.0018 which does not prove to be much helpful.

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

This work validates the argument that adequate pre-processing and data balancing in case of imbalanced datasets are bound to improve the classification performances of the used classifiers. The SMOTE based classifier extends desired performance level to the classifiers measured in terms of AUC. Further, appropriate feature extraction strategies are employed to explore the power of discriminating features in the training of the classifiers and hinder their performance and lower the learning of the models. Finally, to further enhance the productivity of the learners, ensemble approach has been used, which works by combining the results of the base classifiers. The proposed approach is a promising contribution of SMOTE analysis, correlation based feature extraction and ensemble of classifiers with CART, Bagged CART and PART as the base classifiers. The AdaBoost classifier gives the best results in terms of prediction performance measures (AUC, sensitivity, specificity). Thus, the proposed approach can be understood as a viable solution for accurately predicting customer churn in telecommunication industry.

6.2 Future Scope

This work successfully validates how and why the proposed approach could become a valuable contributor in predicting customer churn in the telecommunication sector but, it would be even more useful if the system could determine techniques and propose schemes to retain customers. This would be a challenging task as it would require the system to be able to understand each type of customer and segment them based upon their status and needs. Their needs and requirements need to be grasped and then propose viable solutions to hold on to such customers. This process shall also be paying attention towards reasons why such customers have a tendency to switch to a competitor and then recommend appropriate steps to avoid such a possibility.

References

[1] J. Peppard, "Customer relationship management (CRM) in financial services.", European Management Journal, vol.18, no.3, pp. 312-327, 2000.

[2] D.E. Goldberg and J.H. Holland, "Genetic algorithms and machine learning.", Machine learning, vol. 3, no. 2, pp. 95-99, 1988.

[3] Z. Xuegong, "Introduction to statistical learning theory and support vector machines.", Acta Automatica Sinica, vol. 26, no.1, pp. 32-42, 2000.

[4] R.A. Wood, T.H. McInish and J.K. Ord, "An investigation of transactions data for NYSE stocks." The Journal of Finance, vol. 40, no. 3, pp.723-739, 1985

[5] S.R. Ahmed, "Applications of data mining in retail business." Information Technology: Coding and Computing, vol. 2, pp. 455-459, 2004.

[6] G.Weiss, "Data mining in telecommunications.", Data Mining and Knowledge Discovery Handbook, pp.1189-1201, 2005

[7] W. Raghupathi, "Data mining in healthcare." Healthcare Informatics: Improving Efficiency through Technology, Analytics, and Management, pp. 353-372, 2016.

[8] T. Dasu, and J. Theodore, "Exploratory data mining and data cleaning." Vol. 479, 2003.

[9] W.H. Inmon, "The data warehouse and data mining", Communications of the ACM, vol. 39, no.11, pp. 49-51, 1996.

[10] S Harizopoulos, DJ Abadi and S Madden, "OLTP through the looking glass, and what we found there", Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pp. 981-992, 2008.

[11] E. Rahm, and H. D. Hong, "Data cleaning: Problems and current approaches." *IEEE* Data Eng. Bull., vol. 23, no. 4, pp. 3-13, 2000.

[12] M.L. Brown and F.K. John. "Data mining and the impact of missing data.", Industrial Management & Data Systems, vol. 103, no.8, pp.611-621, 2003.

[13] B. Shneiderman, "Inventing discovery tools: combining information visualization with data mining.", Information visualization, vol.1, no.1, pp.5-12, 2002.

[14] G. Caprio and K. Daniela, Bank insolvencies cross-country experience, 1996.

[15] J. N. Weinstein, "Predictive statistics and artificial intelligence in the US National Cancer Institute's Drug Discovery Program for Cancer and AIDS." Stem Cells, vol.12, no.1, pp. 13-22, 1994.

[16] J. M. Bernardo and A.F.M Smith. "Bayesian theory.", pp. 221, 1994.

[17] J. H. Friedman, "Stochastic gradient boosting." Computational Statistics & Data Analysis, vol.38, no.4, pp. 367-378, 2004.

[18] J.M. Keller, R.G. Michael, and A.G. James. "A fuzzy k-nearest neighbor algorithm." IEEE transactions on systems, man, and cybernetics, vol. 4, pp. 580-585, 1985.

[19] J.A. Hartigan and A.W. Manchek, "Algorithm AS 136: A k-means clustering algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 28, no.1, pp. 100-108, 1979.

[20] T. Kohonen, "The self-organizing map." Neurocomputing, vol. 21, pp. 1, pp.1-6, 1998.

[21] L.L. De, "Singular value decomposition." Proc. EUSIPCO-94, Edinburgh, Scotland, UK, vol. 1, 1994.

[22] S. Daskalaki, "Data mining for decision support on customer insolvency in telecommunications business." European Journal of Operational Research, vol. 145, no. 2, pp. 239-255, 2003.

[23] X.X. Han, "Application of Data Mining in CRM." Applied Mechanics and Materials, Vol. 444, 2014.

[24] A. Idris, R. Muhammad, 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.

[25] A. Idris, A. Khan, and Y.S. Lee, "Genetic programming and adaboosting based churn prediction for telecom.", Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on. IEEE, 2012.

[26] C.P. Wei and C. I-Tang. "Turning telecommunications call details to churn prediction: a data mining approach.", Expert systems with applications, vol. 23, no. 2, pp. 103-112, 2003.

[27] K. Coussement and D. Poel. "Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques." Expert systems with applications, vol. 34, no. 1. pp. 313-327, 2008.

[28] Y. Xie, "Customer churn prediction using improved balanced random forests." Expert Systems with Applications, vol. 36, no. 3, pp. 5445-5449, 2009.

[29] L. Bin, P. Shao and J. Liu. "Customer churn prediction based on the decision tree in personal handyphone system service." Service Systems and Service Management, 2007 International Conference on. IEEE, 2007.

[30] E. Shaaban "A proposed churn prediction model.", International Journal of Engineering Research and Applications, vol. 2, pp. 4, pp. 693-697, 2012.

[31] V. Umayaparvathi and K. Iyakutti, "Applications of data mining techniques in telecom churn prediction.", International Journal of Computer Applications vol. 42, no. 20, pp. 5-9, 2012.

[32] A.A. Khan, S. Jamwal and M. M. Sepehri, "Applying data mining to customer churn prediction in an internet service provider.", International Journal of Computer Applications vol. 9, no. 7, pp. 8-14, 2010.

[33] M. Owczarczuk, "Churn models for prepaid customers in the cellular telecommunication industry using large data marts.", Expert Systems with Applications, vol. 37, no. 6, pp. 4710-4712, 2010.

[34] N. Lu, "A customer churn prediction model in telecom industry using boosting.", IEEE Transactions on Industrial Informatics, vol. 10, no. 2, pp. 1659-1665, 2014.

[35] P.D. Kusuma, "Combining customer attribute and social network mining for prepaid mobile churn prediction." Proc. the 23rd Annual Belgian Dutch Conference on Machine Learning (BENELEARN), 2013.

[36] C.F. Tsai and L. Yu-Hsin, "Customer churn prediction by hybrid neural networks.", Expert Systems with Applications, vol. 36, no. 10, pp. 12547-12553, 2009.

[37] W. Verbeke, "Building comprehensible customer churn prediction models with advanced rule induction techniques.", Expert Systems with Applications, vol. 38, no.3, pp. 2354-2364, 2011.

[38] H. Chen, H.L.C. Roger and V.C. Storey, "Business intelligence and analytics: From big data to big impact.", MIS quarterly, vol. 36, no. 4, 2012.

[39] R. Kohavi, J.R. Neal and S. Evangelos, "Emerging trends in business analytics." , Communications of the ACM, vol. 45, no. 8, pp. 45-48, 2002.

[40] L. Kaufman and P.J. Rousseeuw, Finding groups in data: an introduction to cluster analysis. vol. 344, 2009.

[41] C. Phua, A. Damminda and Vincent Lee, "Minority report in fraud detection: classification of skewed data.", Acm sigkdd explorations newsletter vol. 6, no.1, pp. 50-59, 2004.

[42] T.G. Dietterich, "Ensemble learning." The handbook of brain theory and neural networks, vol. 2, pp. 110-125, 2002.

[43] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* 16 (2002): 321-357.

[44] Denison, David GT, Bani K. Mallick, and Adrian FM Smith. "A bayesian CART algorithm." *Biometrika* 85.2 (1998): 363-377.

[45] Breiman, Leo. "Bagging predictors." *Machine learning* 24.2 (1996): 123-140.

[46] E. Frank and H.W. Ian, "Generating accurate rule sets without global optimization.", 1998.

[47] Y. Freund, S. Robert and A. Naoki, "A short introduction to boosting.", Journal-Japanese Society For Artificial Intelligence, vol. 14, pp. 771-780, 1999.

[48] A.P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms.", Pattern recognition, vol. 30, no. 7, pp. 1145-1159, 1997.

[49] A.P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms.", Pattern recognition, vol. 30, no. 7, pp. 1145-1159, 1997.

[50] P. Domingos, "Bayesian averaging of classifiers and the overfitting problem.", ICML, vol. 2000, 2000.

[51] M.A. Hall, "Correlation-based feature selection for machine learning.", 1999.

[52] A.G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Comparative study of attribute selection using gain ratio and correlation based feature

selection.", International Journal of Information Technology and Knowledge Management vol. 2, no. 2, pp. 271-277, 2010.

[53] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection.", Ijcai, vol. 14, no. 2, 1995.

[54] L. Breiman, "Bagging predictors.", Machine learning, vol. 24, no. 2, pp. 123-140, 1996.

[55] T.G. Dietterich and E.B. Kong, Machine learning bias, statistical bias, and statistical variance of decision tree algorithms, Technical report, Department of Computer Science, Oregon State University, 1995.

[56] D.H. Wolpert, "Stacked generalization." Neural networks, vol. 5, no. 2, pp. 241-259, 1992.

List of Publications

International Conference

Kriti Mishra and Dr. Rinkle Rani, "Churn Prediction in Telecommunication using Machine Learning", International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017), Conference on August 1, 2017 [Accepted].