AN INTELLIGENT SYSTEM FOR CHURN PREDICTION AND CUSTOMER RETENTION: A CASE OF TELECOMMUNICATIONS COMPANY

by

Parth Sarangi

A thesis progress report submitted in partial fulfillment of the requirements for the degree of Master of Engineering in
Information Management

Examination Committee:	Dr. Vatcharaporn Esichaikul (Chairperson) Dr. Matthew N. Dailey Prof. Sumanta Guha
Nationality:	Indian
Previous Degree:	Bachelor of Technology in Electronics and Communication National Institute of Technology Srinagar, India
Scholarship Donor:	AIT Fellowship

Asian Institute of Technology
School of Engineering and Technology
Thailand
March 2017

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

In the past two decades, many nations are witnessing growth in telephonic services due to availability of affordable cellular devices and increasing fidelity of mobile services. Major policies of deregulation by governments have encouraged private corporations to invest funds and support invention of improved technologies. Telecommunications infrastructure and services are the major contributors to the economic prosperity of any country (Cronin et al., 1993). The telecom industry is largely customer service oriented with goals of loyalty, retention and satisfaction (Gerpott et al., 2001). The major source of revenue is from direct selling of cellular and Internet services. The companies involved in delivering the services have invested in expensive infrastructures and software systems.

Over the period of time most telecom organizations provide almost the same service and similar value proposition to the customer. Companies experience high customer defection when competitors bring in new offers, services and technologies. Incumbent telecom operators face consumer churning on a regular basis.

Profitable telecom companies generally have a large customer base and their databases hold a wealth of information. It has become imperative that company leaders need to look into their own subscriber base and study the trends that can reveal customer behavior. The biggest asset for companies in the services domain is the customer (Poel & Lariviere, 2004). Thus companies are resorting to data mining techniques and tools to predict customer churn prediction (Berson et al., 1999). From previous data mining techniques it is inferred that it is more profitable to retain and service existing users than to bring in new subscribers (Reinartz & Kumar, 2003). A small effort to retain customers results in major contributions.

Reports published by TRAI shows the mobile phone subscriptions for each telecom operator in India (TRAI - Telecom Regulatory Authority of India, n.d.). Data accessed from the reports is tabulated as below:

Operators	Customer Count in Aug 2016	Increase or Decrease in Period				Customer Count in Dec 2016
		Aug - Sep	Sep Oct	Oct Nov	Nov Dec	
Airtel	257	2	2	1	2	265
Vodafone	200	0.5	1	0.8	1.8	204
Tata Indicom	58	- 1	- 1	- 1	- 1.6	52
Reliance Jio	0	15	19	16	20	72

Table 1.1: Approx. subscriber counts(in millions) of select companies in Indian telecom industry.

It can be deduced from this report that "Tata Indicom" is continuously loosing customers and "Airtel" & "Vodafone" are adding new subscribers at relatively the same rate as they did before. Whereas " Reliance Jio", a new entrant is experiencing an extraordinary influx of customers so much so that it almost crossed the numbers held by Tata Indicom in Aug 2016.

This thesis presents an intelligent system which predicts customer churn, helps managers and decision makers to identify the valuable proportion for customer retention strategies. The thesis proposes a system supplemented by a data warehouse on the back-end and a visualizations dashboard as the front-end for decision makers. The predictive model is devised after comparison of prediction performance between Decision tree, Support vector machine and neural networks. The proposal is to build a single system as opposed to using separate softwares for prediction, data manipulation and displaying performance indicators.

1.2 Problem Statement

The telecommunication industrys income is based primarily on the sale of services to customers. A companys income can dwindle severely if the mindset of its customers changes. As of this decade we have witnessed a growth of smart-phones and so the need to consume data has increased. Ever so often rivals advertise customer centric plans. Internet service providers are trying to woo customers with free, limited, high speed, unlimited, day only, night only and various other the Internet data campaigns. In the recent history, in Indian telecommunications market, the incumbent operators like Airtel, BSNL, Vodafone, Idea Cellular lost plenty of customers to a new entrant, Reliance Jio. Jio launched its services September 5th 2016. It has been reported that Jio has signed about 72 million customers for its paid services that were free in the past. (Reuters, 2017). This shows the loyalty factor among the customers staying with Reliance Jio. Thus identification of the correct customer segment and understanding their current and future needs is a proactive decision that needs to be taken by companys management. If leaders are tardy and resist change, they could leave their customers dry and sulky. This would obviously result in customer defection and ultimately loss in revenue.

1.3 Objectives

The overall objective of the thesis is to develop an intelligent system for churn prediction and customer retention (ICPCR).

The specific objectives of the thesis are to:

1. Design models and evaluate their churn prediction performance.
2. Build the system of intelligent churn prediction and customer retention system.
3. Evaluate the system for reliable performance.

1.4 Limitations and Scope

There are many models available for churn prediction. The scope of this thesis is to build the system based on three data mining predictive models viz., Decision tree, Support Vector Machines, Artificial Neural Network tentatively.

1.5 Thesis Outline

The organization of this dissertation is as follows:

- In Chapter 2, the literature review is explored.
- In Chapter 3, the methodology is proposed.

Chapter 2

Literature Review

This thesis chapter introduces concepts, technologies, techniques, consulted papers and articles pertaining to the core concepts of Customer Churn & Retention, OLAP & Datawarehouse, Data mining, Model evaluation metrics, Review of of selected papers and Summary of selected papers.

2.1 Customer Churn & Retention

Customers are the most volatile asset of a services based company. Many frequently churn in search of better services. Customers are frivolous and those with prepaid or prepay plans are most unfaithful. Companies are generally in profit if they are able to retain customers and it pays off to almost six times (Bhattacharya, 1998). Customers spending longer durations with a company are not easily churned and would not be affected by marketing strategies of rival companies. These customers are valuable to the company and generate profit in revenue. Research studies have shown that long standing customers would be engaged in influencing newer customers to buy into a contract with their service provider (Mizerski, 1982).
The ARPU of a stable customer is high compared to that of a churning customer. Thus marketing managers are focusing on advertising competitive products to retain customers from churning. The loss of capital due to a defecting customer is higher than the cost of retention. As per Forbes, Nov 11, 2013, earnings can swing positively by about 10 % if customers are successfully retained.

2.2 OLAP & Datawarehouse

Systems and companies are ever expanding. They are collecting data at unprecedented rates. Managing data becomes easier with the implementation of Data-warehouse. In many a cases the database of a company is segregated into different schema's. Segregation of schema's helps to avoid necessary access privileges and grants confusion. It also helps to maintain the organizational level of segregation in the database, ie., the HR department tables will be unaccessible to an accounts official and vice versa. But company leaders and decision makers should be accessing specific key counts and aggregations from all of their departments. A collection of tables sourcing data from their individual units.

OLAP - Online Analytical Processing is an extension of Data-warehouse technology (Han, 1997). Olap consists of four main processes viz., Drill-down, Roll-up, slice and dice. Multi-demensional data can be fetched by OLAP from the Datawarehouse, and the unit of this is called the OLAP cube.There are two types of OLAP - MOLAP & ROLAP. MOLAP Multidimensional OLAP is a solution used widely.

One very famous open-source OLAP solution is the Kylin™ (Kylin, n.d.). Shown in Figure 2.1 is the architecture of the product.

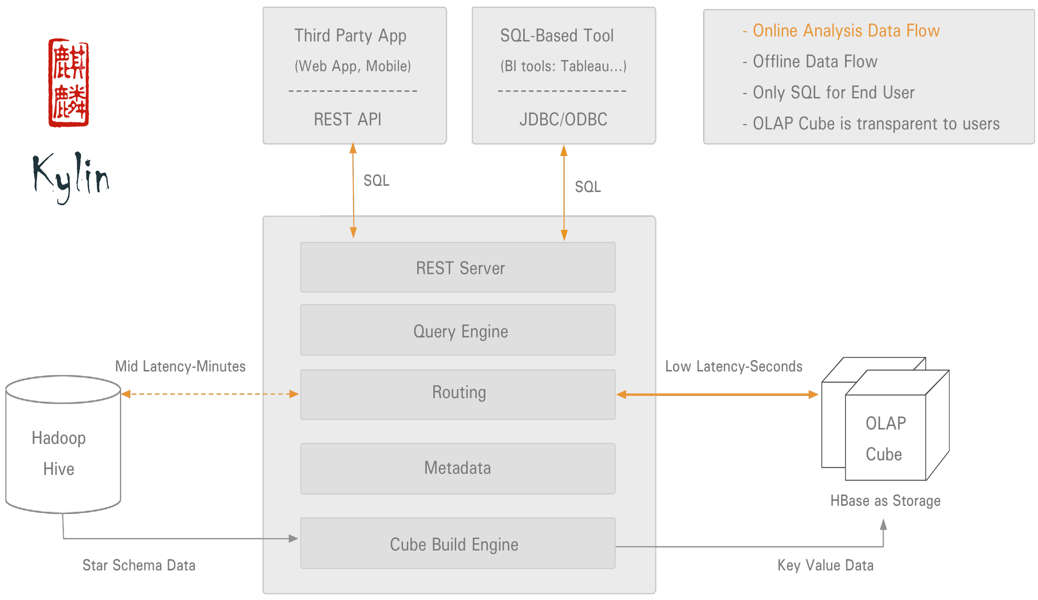


Figure 2.1: OLAP Solution - Apache Kylin

2.3 Data Mining

Data mining is the process of extracting useful trend and patterns from structured and unstructured sources of data. Sometimes many academicians refer to it as KDD (Knowledge Discovery in Databases). John Naisbett (author of famous 'Megatrends') said "We are drowning in information but starved for knowledge." There are various techniques to perform data mining and these can be broadly classified into two categories Supervised Learning, Un-Supervised Learning. A very common terminology used in the data science field is of machine learning and it also used instead of data mining.

2.3.1 Supervised Learning

This part of the data mining consists of classification and regression algorithms. Control and dependent variables of the given data are known entities. The use of these algorithms is to predict the outcome given past data. These algorithms have to be trained with a set of data and then they have to be tested. After reaching certain acceptable level of accuracy, these algorithms are used for prediction.

Below are some of the Supervised learning techniques :

- Linear regression : The prediction of dependent variable is done given the value of known variable. There is only 1 dependent variable. For example, $y = \beta 0 + \beta 1x + \varepsilon$
  y = dependent variable, x = independent variable

- Multiple regression : is an extension of the linear regression but has more number of independent variables.

- Nonlinear regression : there are two variables but they are related in a curvilinear fashion i.e., not governed by the straight line equations.

- Logistic regression : A regression based modeling technique, which is better than linear regression when more variables are considered. Output variable is categorical in nature.

- Decision tree : This is a classification algorithm which when plotted resembles an upside down tree structure. Given that a set of data has many attributes and there is a need to classify them, a decision tree is very suitable method to do so. There are many types of decision trees like the ID3, CART, C4.5 and C5.0. In Figure 2.2 a simple DT for mammal classification model is shown. A decision tree can be designed using **Hunt's Algorithm**.

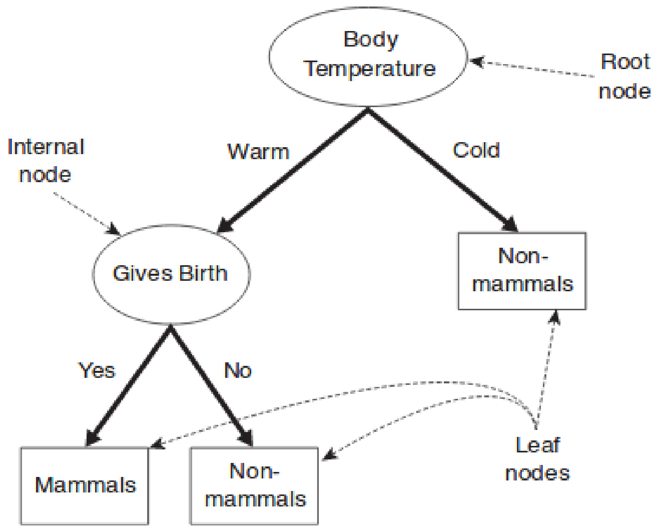


Figure 2.2: Mammal classification problem

- Random Forest : This technique can be used for both classification or regression type problems. A random forest is combination of many decision trees. In some cases random forest is sometimes very accurate.

- Support Vector Machine : This is a classifier technique where the data is segregated by generating hyperplanes. If there are n-features in the data then there have to be n-hyperplanes. The best classification is the hyperplane which clearly separates the data points.

- k-Nearest Neighbors : A learning algorithm that classifies the data into clusters nearest to them. The euclidean distance or manhattan distance could be some of the methods to find the nearest cluster. It is sometimes considered a lazy learning algorithm.

- Naive Bayes : This is an classification rule working on the probabilistic Bayes theorem. $P(H|X) = P(X|H)P(H)/P(X)$.

- Artificial neural networks : Neural networks are classification methods modeled after neurons (karpathy@cs.stanford.edu, n.d.). There are many layers with nodes Figure 2.3. There are many types of neural networks viz., Feed Forward NN, Radial Bias function, Recurrent NN, Backpropagation NN, Perceptron etc. Neural networks are very fast learners.

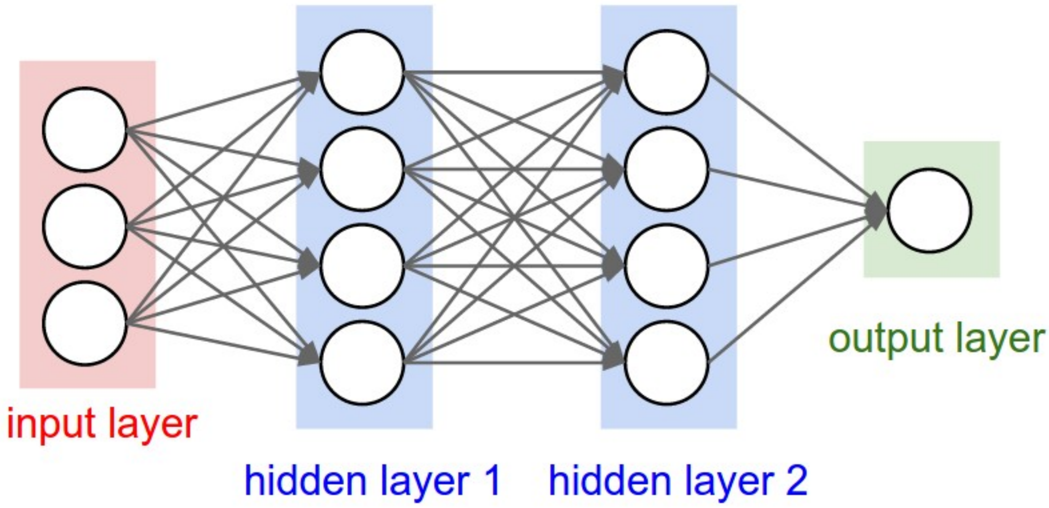


Figure 2.3: A sample neural network

2.3.2 Un-Supervised Learning

The clustering and association techniques in data mining are grouped into Un-Supervised learning. The output variables are not known. Below are some Unsupervised class of algorithms :

- K-means clustering : it is a means of clustering a set of data points with some k centroids. For each data point the distance is calculated and the nearest centroid is chosen and data point is associated with that cluster. After every iteration of cluster formation a new centroid is calculated and the distance of the data points are taken. The clusters are reformed and the iteration is performed till no data point movement happens.

- Apriori clustering : Here in the A priori algorithm is used to create the clusters. A priori is used for frequent item set mining states that sets of items are frequent if the items themselves are frequent.

- Hierarchical clustering : This is a clustering method in which large clusters are further segregated into smaller clusters. This is the Divisive type of HC. In the Agglomerative type of HC, the nearby clusters are joined to form larger clusters. A Dendogram is used to graphically represent the clusters.

- Hidden Markov models : These are used to analyze or predict time series problems in fields of speech, language, medicine, and robotics. Core of the technique is formed on the foundations of Bayes Network. In a markov chain a future state depends only on the current state. It is called Hidden because only certain measurements can be see of the states, not the states itself. Particle filter and Kalman filter are HMM's.
- Self organizing maps (SOM) : This is a type of neural network. Types are of Vector Quantizer or Kohonen SOM. In Figure 2.4 is an illustration of an Kohonen SOM.

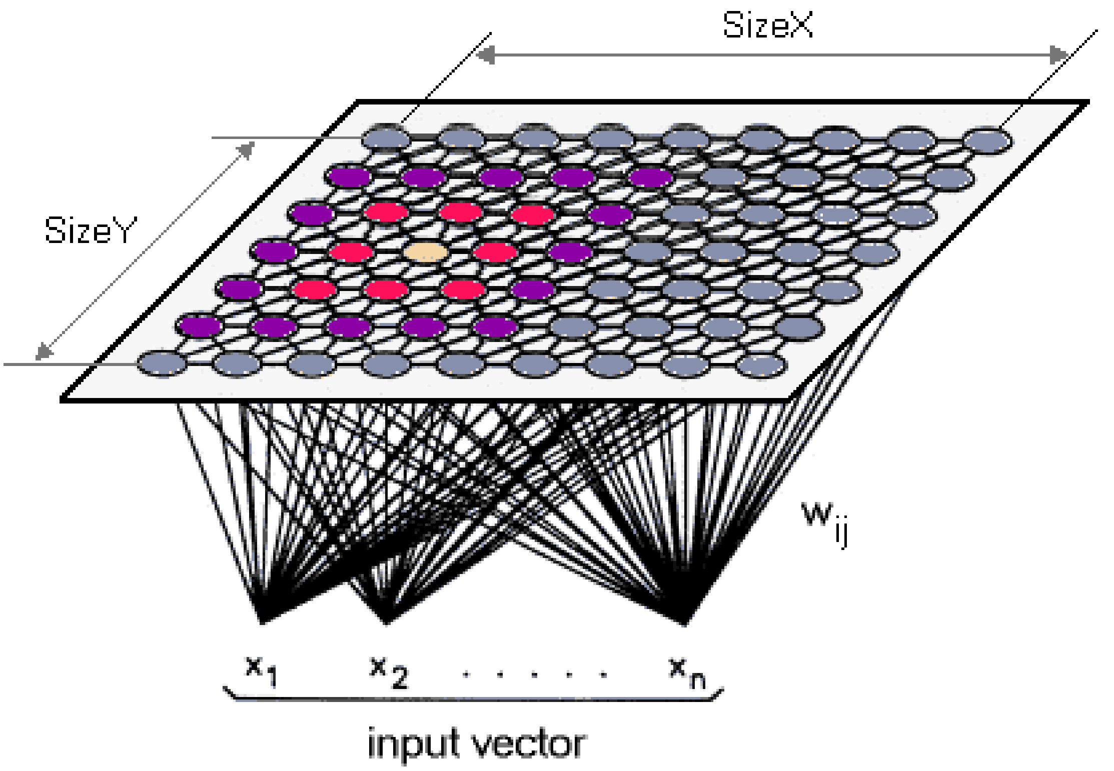


Figure 2.4: Kohonen SOM

2.3.3 Selecting the Right technique

It is of utmost importance that a data scientist select the important mining technique. Of all the process involved in the knowledge discovery process, selection of algorithm is quite difficult. Figure 2.5, from "Choosing the Right Data Mining Technique: Classification of Methods and Intelligent Recommendation" (Gibert et al., 2010) shows the approach which could be taken to select between the various models available for data mining.

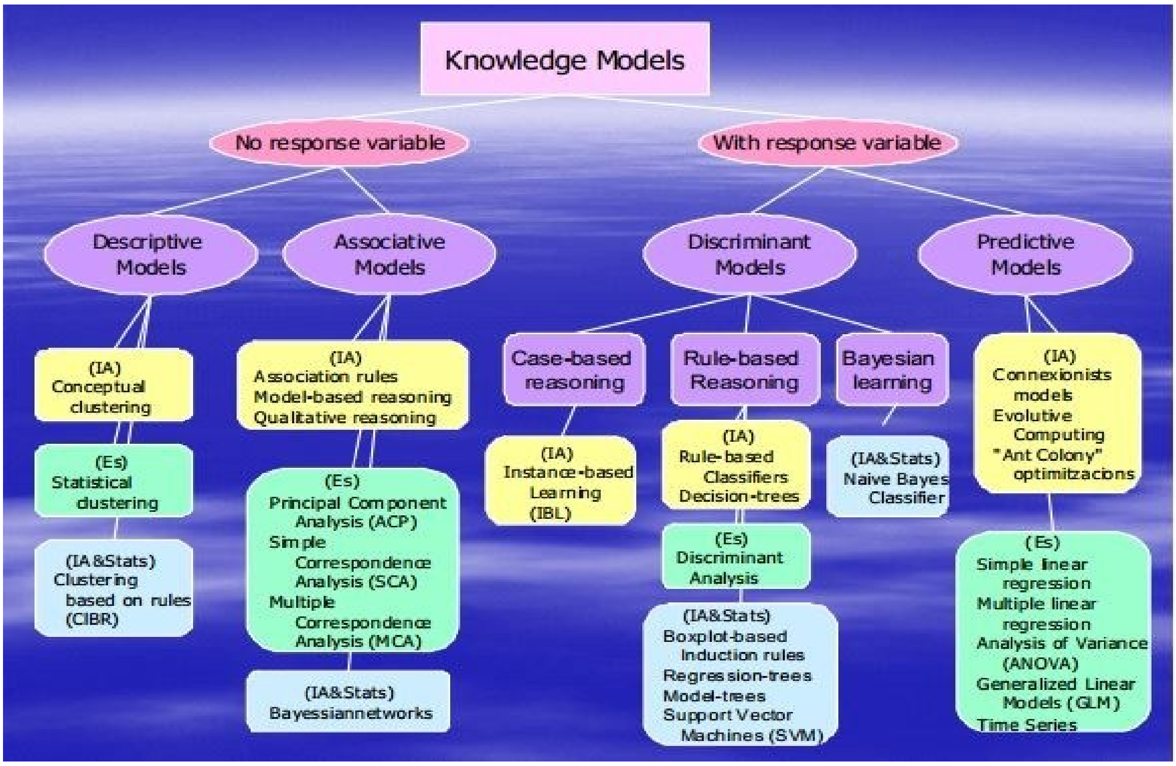


Figure 2.5: Select the Right Mining Technique

In addition to the above there is another approach, shown in Figure 2.6 suggested by the very popular scikit (machine learning library) of python for data mining (Scikit, n.d.).

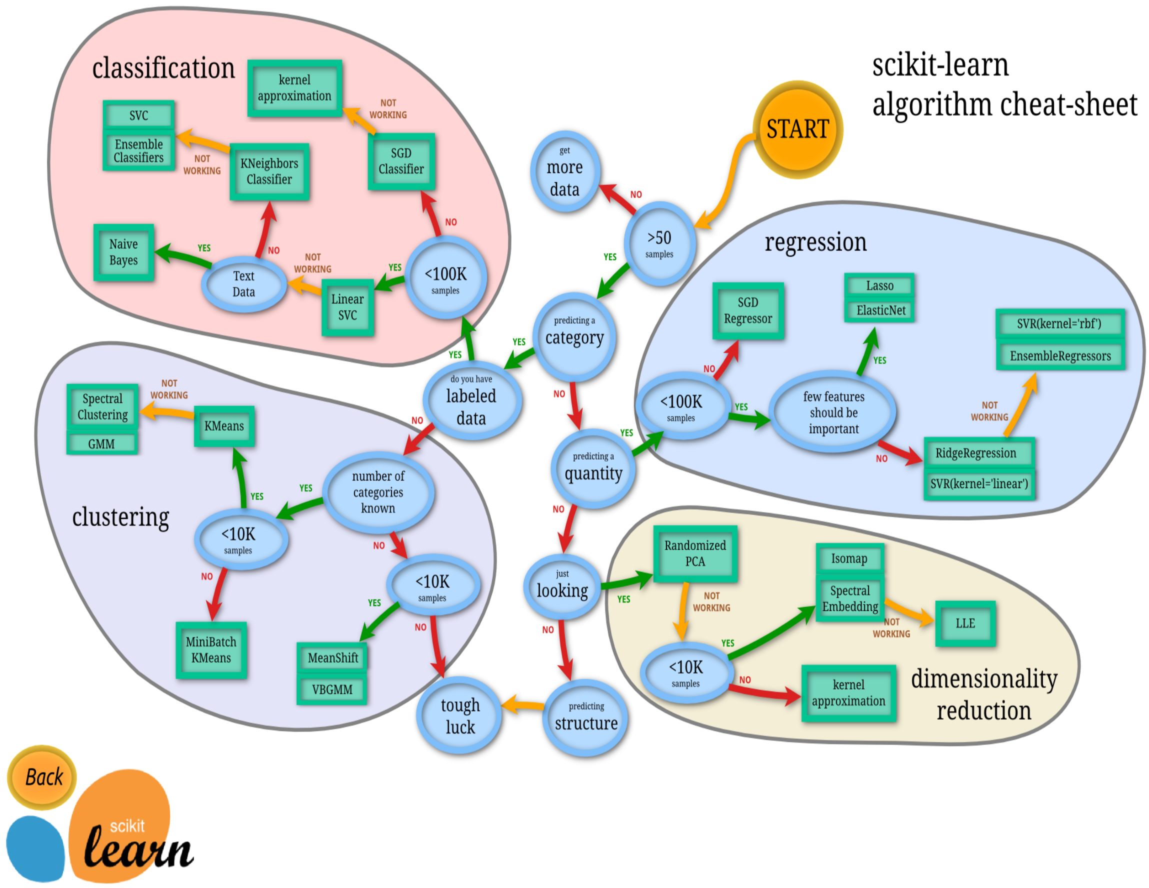


Figure 2.6: Another approach to select the Data Mining. Reprinted from Scikit

2.3.4 Softwares, Libraries and Servers

Data mining techniques have been implemented into modules by a number of generous contributers. There are some very famous solutions that an academician can utilize for implementing predictive analytics. The algorithms and techniques of neural networks, clustering, classification ans associations are available as solutions and API's. Following is a list in no particular order:

- Softwares : The following softwares are open source and available for data analytics by downloading to the desktop.
  - Weka : This is a open source software containing data mining algorithms. This can be used as a software or called by users own Java code.
  - Knime : An open source tool for data mining, comprises of many functions from data cleaning to pattern analysis.
  - Rapidminer : Also another popular tool for machine learning with plenty of algorithms for analysis.

- Libraries : These are available for use as toolbox and academic can program own solution.
  - Tensorflow
  - mlpack
  - H2O
  - Mlib
  - Scikit
- Servers : The following servers have built in modules that can be accessed via web applications and can be modeled to process real time analytics instead of one of processing as with above solutions
  - **DeepDetect** : is an open source deep learning server implemented in C++. It can be supported with back end machine learning applications with TensorFlow XGBoost and Caffe. Model assessment is built in the framework.
  - **Apache Prediction IO** : This a open source stack for academicians to deploy machine learning. The stack has an Event Server that can be used to query from a web application and respond in real time. The Event server co-ordinates with the Engine to respond to API inputs and respond with predicted outcomes Figure 2.7 (PredictionIO, n.d.). PredictionIO provides various templates for varied mining algorithms. Classification templates like Decision trees, Logistic Reressiong, NLP are available for use.

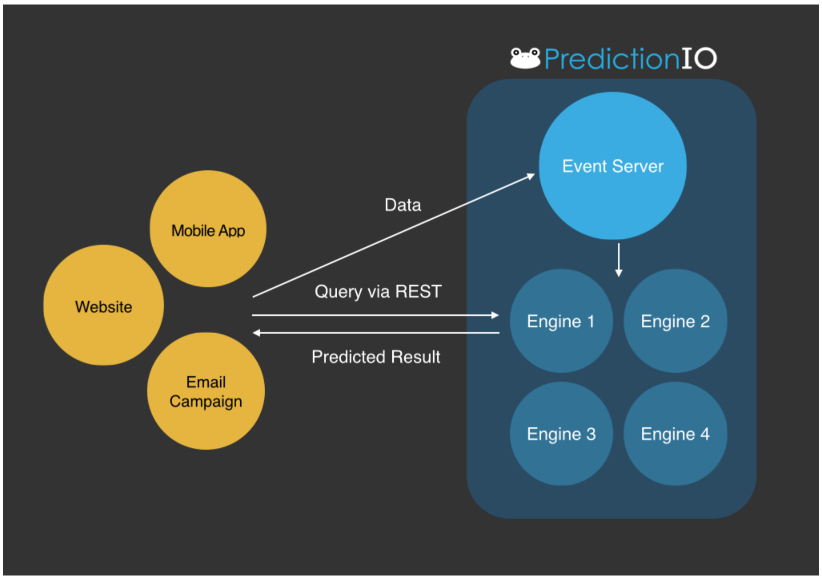


Figure 2.7: PreditionIO Engine interaction with Apps and Prediction Engine

  - **Shiny** : This is an R package and allows for easy to build web applications. It is made of two parts UI script and server script. In Figure it can be seen how Shiny can be implemented to exploit the data mining capabilities of R. Shown in Figure 2.8 how multiple users can access shiny R applications (Rstudio, n.d.).

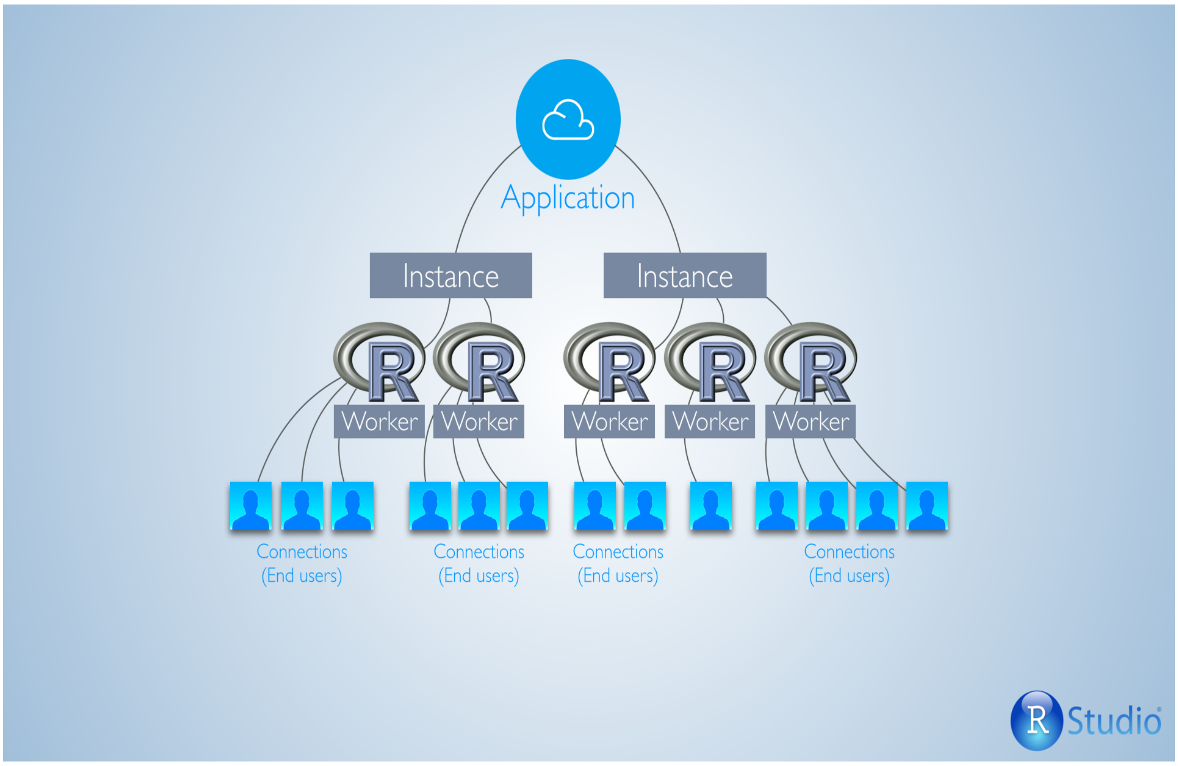


Figure 2.8: R Shiny architecture

2.4 Model Evaluation Metrics

Model development is an important process, but evaluations of the model to ascertain its performance is as much an important procedure. The dataset is partitioned suitably and the testing set is not in the view of the model during training. There are however two methods of evaluations.

- Holdout technique
- K-fold Cross validation technique
- Leave one out CV
- Bootstrap method
- Sensitivity & Specificity

2.4.1 Holdout technique

This method is chosen for evaluation if the dataset is large enough. The data is segregated into three parts viz., Training, Validation and Test sets.

- Training dataset : It is some part of the dataset used for training the models. Predictive models are necessarily trained before actual prediction can be performed eg., Decision Trees, Random forest, Neural network need to be trained.
- Validation dataset : This is a subset of the data used to validate the output after model training. It helps to optimize the models performance. It is not mandatory to have validation sets for certain prediction models.
- Test dataset : Also a part of the whole dataset, it helps to

2.4.2 k-fold cross validation technique

This method of evaluations is chosen if the dataset is small and limited. The data is partitioned into k equal sized sets with an unbiased process. The model is built k times, with every K-1 data sets selected as training set, leaving out 1 set to be used as test set. A round robin process is followed to select the testing set in every iteration.

2.4.3 Sensitivity & Specificity

For calculating the performance of the model, a confusion matrix is plotted. The matrix is a cross table between predicted values and the actual values Figure 2.9. There are generally four types of values that can be calculated from the matrix and those are as follows :

- TP - true positives : The predictor predicts "True" for actual true value of data.
- TN - true negatives : The predictor predicts "False" for actual false values of data.
- FP - false positives : The predictor predicts "False" for actual true value of data.
- FN - false negatives : The predictor predicts "True" for actual false values of data.

Sensitivity : the ratio of the count of the True Positives to the total count of events. This is also called the ***Recall***.

$$Sensitivity(or Recall) = \frac{TP}{TP + FN}$$

Specificity : the ratio of the count of the True Negatives to the total count of non-events.

$$Specificity = \frac{TN}{FP + TN}$$

In addition to the above, True Positive value is called the ***Precision***.

Form the values of *Precision* and *Recall* another statistical measurement called F-score can be derived.

$$F = 2 \times \frac{precision \times recall}{precision + recall}$$

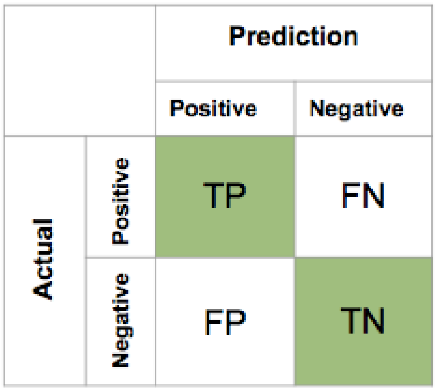


Figure 2.9: Confusion Matrix

2.5 Review of Selected Research Papers

In the paper titled " Modeling & Simulation of a Predictive Customer Churn Model for Telecommunication Industry the authors emulated a neuro fuzzy inference system to study the customer churn in the telecom industry (O et al., 2015). They modeled membership functions for the attributes of the dataset. Then they employed search algorithm for feature selection of the variables that indicate churn. Thereafter they model fuzzy equations to relate the dependent variables to the independent variables. This fuzzy system is trained to tune the Adaptive neuro fuzzy system based on the Sugeno FIS. The call detail records of 5000 subscribers was used to model this FIS. The dataset has 21 attributes but here they selected 9. Then the variables were modeled into three categories. For performance evaluation they calculated the Precision rate and the recall rate. After the testing it was found that accuracy was 95.8% , precision 80.86%, recall 92.7%.

A research study "A Hybrid Churn Prediction Model in Mobile Telecommunication Industry " (Olle & Cai, 2014) presents a combination of LR and VP method. The academics used two algorithms of supervised learning viz., Logistic regression and Voted perceptron. They then combined the two into a Hybrid model for classification in WEKA. The obtained the data from an Asian telcom operator, records of around 2000 customers and 23 attributes.
From the results it was observed that hybrid model performed better than each of them individually.

In the study "A comparison of machine learning techniques for customer churn prediction" by (Vafeiadis et al., 2015) the researchers present a well meted out comparison between the normal model functions and their corresponding boosted models. The performance criteria was based on the F-score. They had used a series of simulations based on the Monte Carlo method. The models selected for analysis were Back-Propagation algorithm , Support Vector Machines, Decision Trees, Naive Bayes and Logistic Regression. The data was obtained from the publicly available churn dataset hosted at UCI Machine learning repository. The 100-fold cross validation technique was used to reduce bias. Ratio of training to testing set is about 2 : 3. A type of the most common boosting algorithm Adaboost, *Adaboost.M1* with DT and BPN as weak classifier was used.
The R programming was used for modeling the simulation experiment. Two steps were followed : Step 1 - tested classifiers run with data and performance of F-score measured. Step 2 - boosting algorithm was applied and performance F-score measured. 100 Monte carlo realizations were generated for cross validation of results. Monte carlo is synthesis of datasets that resemble the actual data. It was derieved from the results that two prediction models performed the best. 2 layer BPN with 15 hidden nodes and Decision tree classifier. An accuracy of 94% and F-measure around 77%. The SVM scored lower followed by Naive Bayes and Logit Regression at last. After application of the Boosting algo, SVM reported the best accuracy of 97% and Fmeasure over 84%.

2.6 Summary of Selected Research Studies

Here some of the past relevant literature in the domain of churn prediction and the results are discussed in Table 2.1.

Table 2.1:: Previous literature review

SNo	Title & Author	Objective	Data & Methodology	Outcome	Further Research
1	Modeling & Simulation of a Predictive Customer Churn Model for Telecommunication Industry (O et al., 2015)	Adaptive neuro fuzzy inference system for prediction emulation of customer churn Neural network + fuzzy logic.	**Data :** 5000 subscribers CDR call detail record with 21 vriables. Partitioned into 5 sets each containing 1000 records. **Method :** Number of predictor variables taken is 9. Target variable is Chrun with value Y or N. Membership function for each variable.	Found that 3 variables are very important. Total no of minute calls, no of customer calls, no of customer service calls, no of repaired calls. Fuzzy churn model Precison 80.86% recall 92.7% and predicted accuracy 95.8%.	None suggested
2	A Hybrid Churn Prediction Model in Mobile Telecommunication Industry (Olle & Cai, 2014)	A model combined with VotedPerceptron and Logisti Regression is performance compared to the models of VP and LR as individual predictors.	**Data :** 2000 customers CDR from an Asian telecom company with 23 attributes. **Method :** A hybrid model of VP and LR was used. WEKA tool was used to model.	The hybrid model performs better than the models prediction accuracy seperately.	None suggested

SNo	Title & Author	Objective	Data & Methodology	Outcome	Further Research
3	A comparison of machine learning techniques for customer churn prediction (Vafeiadis et al., 2015)	The normal model functions were performance compared to their corresponding boosted models.	**Data :** publicly hosted churn dataset at UCI machine learning repository. **Method :** Machine learning techniques of Back-Propagation algorithm , Support Vector Machines, Decision Trees, Naive Bayes and Logistic Regression were used. The boosting algorithm Adaboost.M1 a type of Adaboost was used. R progamming was used for modeling the system.	2 prediction models performed the best : 2-layer BPN with 15 hidden nodes and Decision tree classifier. SVM scored lower followed by Naive Bayes and Logit Regression at last. After application of the Boosting algo, SVM reported the best accuracy of 97% and Fmeasure over 84%.	None suggested
4	Turning telecommunications call details to churn prediction: a data mining approach (Wei & Chiu, 2002)	The company experiences a high monthly churn rate of 1.5 2Neural network requires a long time due to its iterative nature. Highly skewed class distribution between churners and non-churners.	**Data :** Telecom company of Taiwan. Contractual and call details of subscribers Oct 2000 Jan 2001. 9100000 records. **Method :** Multi classifier class combiner, Decision tree C4.5	Churn prediction is relatively high within 1 month duration. Multi classifier performs better than single classifier.	To include more variables from logs and complaints. Evaluation of empirical stats between customers from different geographic locations. Integration with datawarehouse for constantly learning behavior of customer. Research with other industry data from credit card to Internet service providers.

SNo	Title & Author	Objective	Data & Methodology	Outcome	Further Research
5	Applying Fuzzy Data Mining to Telecom Churn Management (Liao & Chueh, 2011).	To determine the most effective marketing strategies of customer retention, by analyzing the responses of customers.	**Data :** Taiwan telecom company, retention activity & response data for customer contract expiry between June and Junly 2008 **Method :** ID3 decision tree for classification.	Using fuzzy set the customer retention shows that marketing via telemarketing is more effective compared with Direct mailing. Also fuzzy marketing technique is better than direct mailing marketing for customers with higher bill amounts.	Fuzzy data mining techniques to analyze the past records of results of various marketing activities to establish a marketing mode.
6	Customer churn prediction using improved balanced random forests (Xie et al., 2009).	a novel learning method, called improved balanced random forests (IBRF), and demonstrate its application to churn prediction	**Data :** Chinese bank data. 1524 [762 train, 762 test]. **Method :** IBRF = Balanced random forest + weighted random forest. Introduce 2 interval variables m middle pt & d length of interval. apply IBRF to a set of churn data in a bank as test the performance of our proposed method, we run several comparative experiments comparison of results from IBRF and other standard methods, namely artificial neural network (ANN), decision tree (DT), and CWC-SVM (Scholkopf, Platt, Shawe, Smola, & Williamson,	Accuracy rate follows this pattern $IBRF > CWC - SVM > ANN > DT$, Top-decile Lift varies as this $IBRF > CWV - SVM > DT > ANN$. IBRF offers great potential compared to traditional approaches due to its scalability, and faster training and running speeds.	Experimenting with some other weak learners in random forests. Improving effectiveness and generalization ability.

SNo	Title & Author	Objective	Data & Methodology	Outcome	Further Research
7	Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques (Coussement & Poel, 2008)	Churn prediction using SVM. Benchmarked to Logit regression and random forest.	**Data :** Belgian newspaper publishing company. Training set 45000, Test set 45000 **Method :** Use of random forest software and SVM-toolbox. SVM compared to Logit regression & random forest. Grid search using 5-fold cross-validation	SVM trained on balanced distribution, outperforms logit regression when parameter selection applied. Random forest surpass SVM. Academincs and practionerx dont need to rely on traditional Logit reg, SVM with parameter selection technique and random forest offer better alternative	No complete working meta-theory to choose kernel function and SVM parameters. Thus deriving a procedure to select proper kernel function and SVM parameter.
8	Customer churn prediction by Hybrid neural networks (Tsai & Lu, 2009)	Very few studies for hybrid data mining apraoch for prediction.	**Data :** CRM dataset from American telephone company, July 2001 to Jan 2002 51,306 subscribers. **Method :** 2 methods developed and compared for performance. M1 SOM + ANN clustering + classification is used. M2 ANN + ANN 2 classifiers are used. 5 fold cross validation, each set of the 5 are tested 5 times. Baseline is 20 ANNs	Baseline ANN models had prediction accuracy of 88% performance : $ANN + ANN > singleANN$ $3 * 3$ SOM is best among $2 * 2$, $3 * 3$, $4 * 4$ and $5 * 5$ clustering Performance of the hybrid models is : $ANN + ANN > SOM + ANN > ANN$	Need to explore dimensionality reduction or Feature selection of data preprocessing. Application of SVM or genetic algorithms. Explore other domains for churn prediction.

SNo	Title & Author	Objective	Data & Methodology	Outcome	Further Research
9	Predicting customer retention and profitability by using random forest and regression forest (Larivière & Poel, 2005)	The paper discusses more than one variable of retention and profit outcome.	**Data :** 100,000 Belgian finance company. Divided into 2 random parts, one for estimation other for evaluation. **Method :** Authors used random forest for regression to predict profitability, next purchase and defection decision. Benchmarked to linear regression model.	Random forest are better than logit and linear regression.	None suggested.
10	Churn prediction using comprehensible support vector machine: An analytical CRM application	The paper discusses more than one variable of retention and profit outcome.	**Data :** 100,000 Belgian finance company. Divided into 2 random parts, one for estimation other for evaluation. **Method :** Authors used random forest for regression to predict profitability, next purchase and defection decision. Benchmarked to linear regression model.	Random forest are better than logit and linear regression.	None suggested.
11	Churn prediction for high-value players in casual social games	Paper presents churn prediction of players of social games and the business impact of retaining high valued players.	**Data :** dataset of high value users of games - Diamond dash and Monster World, for 2 days. **Method :** The researchers trained and predicted neural networks, logistic regression, decision tree and support vector machine. Radial basis function for support vector machine was used with 10-fold cross validation. For business impact of churning the researchers designed A/B test.	Single neural network with tuned learning rate is better that other algorithms. A/B test reveals that sending free coins to high value customers does not affect churn rate.	None suggested.

Chapter 3

Methodology

In this chapter the methodology for implementing the ICPCR system is illustrated. Also the steps that would be followed are outlined.

3.1 Research Methodology

The following steps will be conducted also shown in Figure 3.1 :

Step 1: Data Preprocessing and Datawarehouse Development

- Data Collection
- Meta-data evaluation
- Data cleaning
- Datawarehouse design
- ETL process

Step 2: Development and Evaluation of the Prediction Models

- Select three churn prediction models
- Models to be trained and tested with the data
- Model Evaluation

Step 3: System Development & Evaluation

- Build the ICPCR system as a web application.
- Integration of Web app with OLAP and prediction model.
- Develop the Dashboards to display KPI's.
- Test the system.

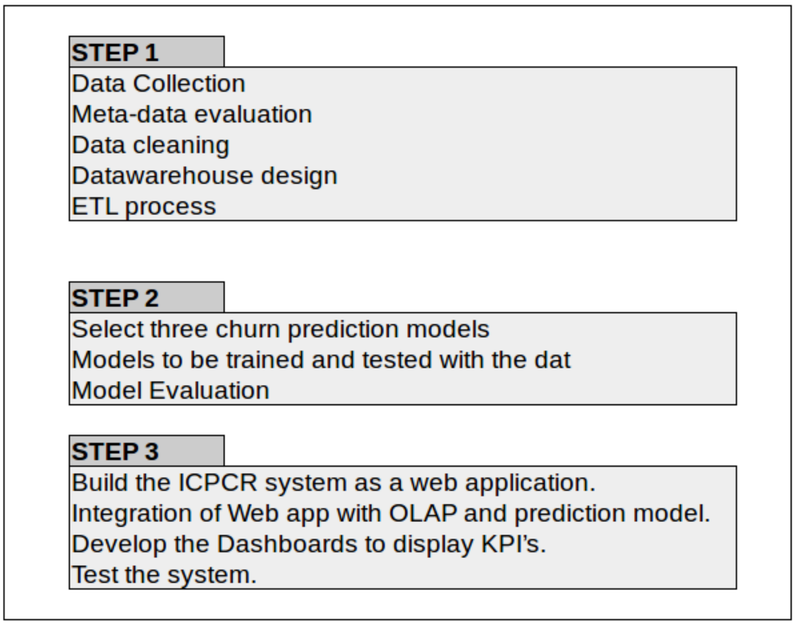


Figure 3.1: Research Methodology

3.2 Data Preprocessing and Datawarehouse Development

3.2.1 Data preprocessing

Data will be collected from available open source sites. In this section a sequence of steps for data preparation are listed. In Figure 3.2 the process flow is shown.

1. Study of meta-data of the dataset. This study reveals the important attributes to be used for prediction.
2. Cleaning of un-usable data, either by replacing with suitable or by entirely removing it. Un-usable data is the one that may be invalid like null or special characters in numeric fields etc.
3. Extract the data and load into the database. This helps in querying the data faster with Structured Query Language.

3.2.2 Datawarehouse development

Following steps will be followed for design of data warehouse:

The attributes generated from above step are summarized. This summary is used to design the OLAP cube. The OLAP will be used in generating reports and KPI's for the dashboard generation. The

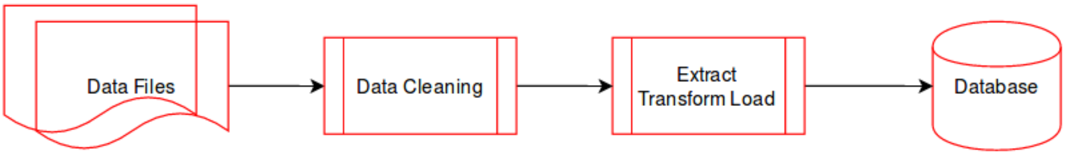


Figure 3.2: Data preprocessing

OLAP will be designed with the star schema. Figure 3.3 shows a typical implementation of the star schema (Tutorials Point, n.d.). A similar structure will be implemented for the study after the dimensions of the data are finalized.

Like for example the count of all the people between the age of 22 to 24 using prepaid service for the year 2013 could be one data whereas the count for 2014 would be another.

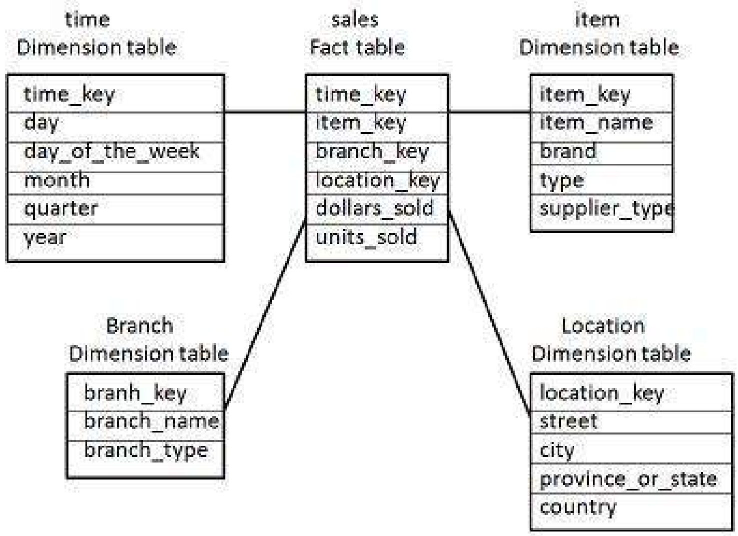


Figure 3.3: OLAP Star Schema

After the Datawarehouse is designed, the tables have to be loaded with data. Thus the next step of ETL is done. Extract Transform and Load processing (ETL) This is a necessary step that would be required to properly extract data from the data file, transform the data types in order that they may be suitable for the database and finally loading to database.

3.3 Development and Evaluation of the Prediction Models

3.3.1 Model Design

In this section, the models are selected for churn prediction. Tentatively it is decided to select Decision tree, Support Sector Machine and ANN. The models will be trained with a training set and then the performance will be evaluated with the testing set. The proposal is to select either the machine learning library of MLib under Apache or Scikit of Python or libraries under R. It will largely depend on the availability of the models in the libraries. In case a model is not available it will be sourced from another library. Also in addition it is proposed that a boosting algorithm like Adaboost would be used to measure change in prediction performance.

3.3.2 Model Evaluations

In order to judge the better performing model or rather the accuracy of predictability by the classification techniques, it is but necessary to perform an evaluation. The evaluations that are commonly performed by academicians are the k-Fold Cross Validation, Sensitivity & Specificity measurements (Larivière & Poel, 2005).

- K-Fold Cross Validation : It is proposed to per form this process to make the classification model more accurate. From previous literature it is learned that k = 100 is highly appropriate.
- Plotting of confusion matrix, as followed by other academicians and then deriving the Sensitivity, Specificity, Precision, Recall and F-score are the proposed evaluation techniques

3.4 System Development & Evaluation

In this section the architecture of the ICPCR system is proposed. The application, shown in Figure 3.4, would be developed in a 3-tire format i.e, Database Layer, Application Layer, and Presentation Layer. The system is designed in two modes. One is the learning phase mode and the other is the Prediction phase mode. In the learning phase the system is fed data and the inference engine learns the trend. Testing and benchmarking along with weighting.

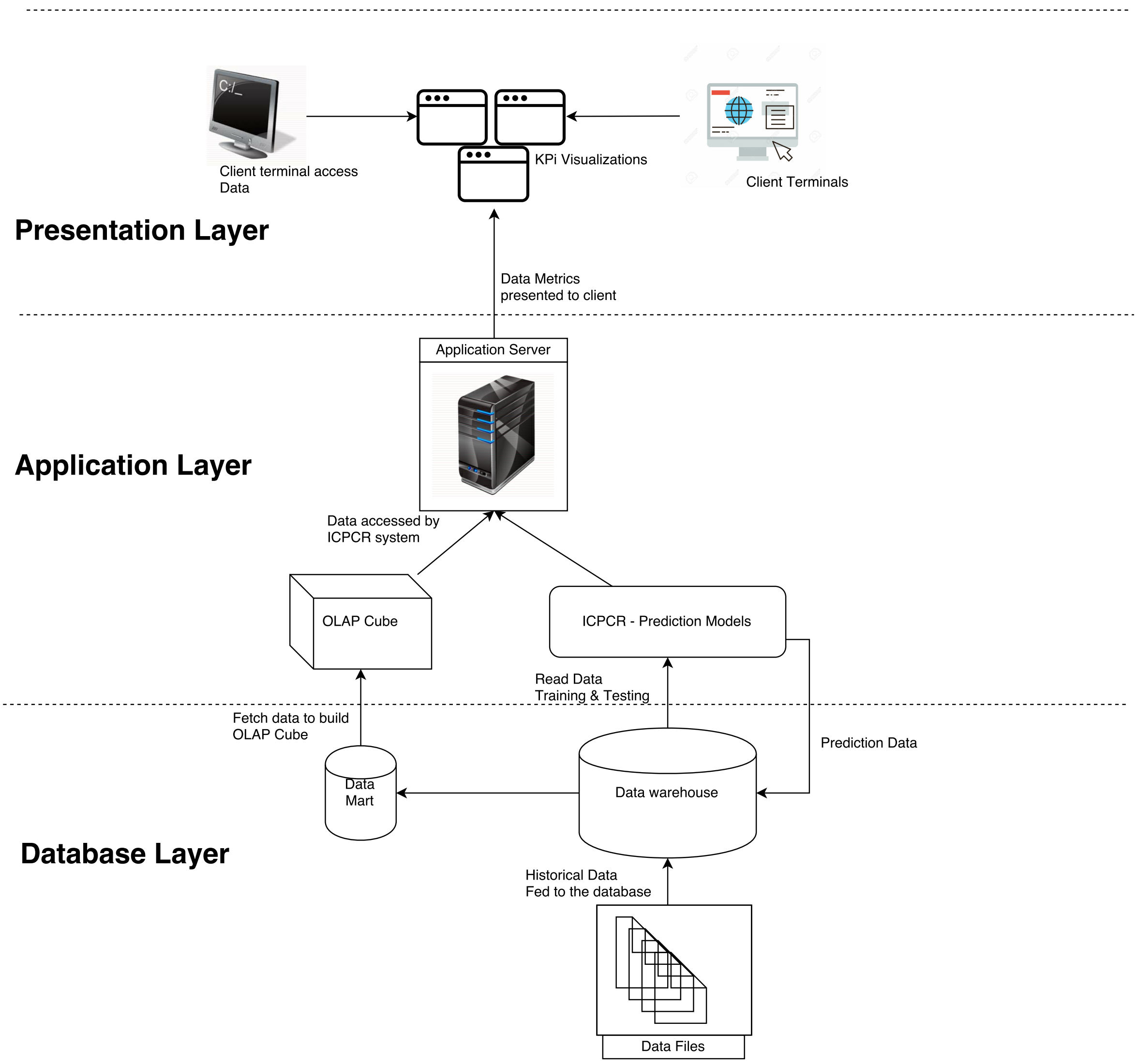


Figure 3.4: The Intelligent Churn Prediction Architecture

3.4.1 Presentation Layer

In this thesis, the presentation layer is the section of the system which is accessible to the user or client. This is used to view the key values obtained from the OLAP and the mining results. There would be a display of metrics of the data.

1. It is proposed to deploy a suitable application to display a dashboard of KPI's.
2. The display of KPI's will be in graphs and charts format. The KPI's are taken from the OLAP cube.

3.4.2 Application Layer

This layer would be comprised of three parts.

1. Application server : This consists of the set of logic codes which will fetch the appropriate data for display in the front end. It may fetch the data directly from the tables or from the OLAP Cube, as is requested from the user.
2. Prediction model : This part is comprised of the predictive model to predict the outcome of data presented to it in the database. The model will go through a phase of training, testing, and prediction of churn value for new data. Also it is proposed that Prediction model be able to identify the variables which could be addressed for retaining the customer.
3. OLAP : This is the MOLAP implementation for building the Key metrics from the data. This part of the system would be responsible for the dashboard metrics display to the user.

3.4.3 Database Layer

This layer will be comprised of the data-warehouse tables. The OLAP calculation and the Model predictions will be updated whensoever a set of ne data is identified. The Olap cube feed tables will also be present here. A Star schema will be implemented for fetching of data for the various dimensions of the OLAP.

3.4.4 System Evaluation

The thesis proposes a system evaluation process to audit the performance. A set of test from latency in display and run will be calculated and improved before the process of deployment. This would ensure that system does not behave erratically under normal situations.

3.5 Timeline

The forecast of the tasks to be carried out in this thesis are shown below in a Gantt chart Figures 3.5.

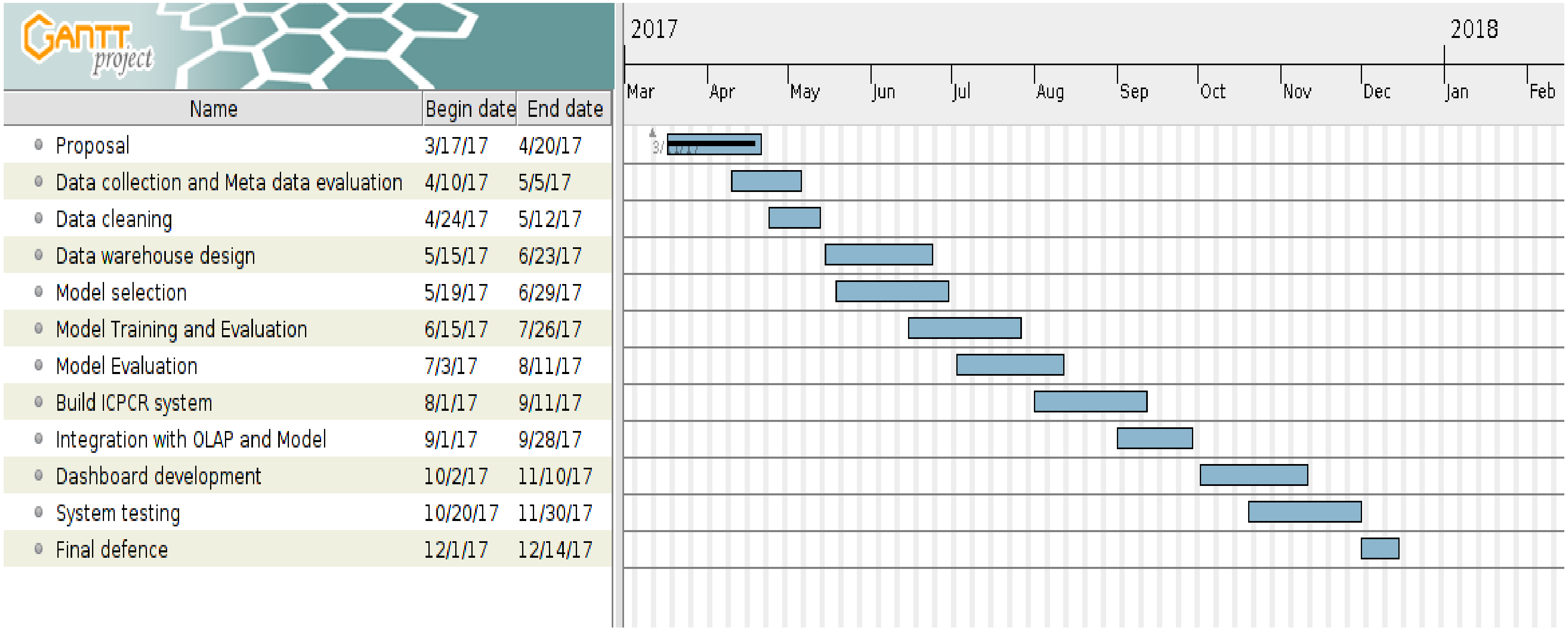


Figure 3.5: Gantt chart tasks

Chapter 4

Data preprocessing & data-warehouse

This chapter presents the progress in data preprocessing and data-warehouse development.

4.1 Data collection

The data was collected from the open source data available at SGI MLC++ website hosted at : `https://www.sgi.com/tech/mlc/db/`. The site has two files suitable for churn prediction. The data was donated to the public domain by Orange telecom.

- Data file *"churn.all"* at location *`https://www.sgi.com/tech/mlc/db/churn.all`*.
- Meta-data file *"churn.names"* for the data at location *`https://www.sgi.com/tech/mlc/db/churn.names`*

The data has 5000 records of call details for customers. Collected over a period of time it lists whether if the customer has churned or retained. The data has 19 dimensions. They are described in tables 4.1 & 4.2.

Table 4.1: Dimension descriptions

Name of Dimension	Description	Type
State	state's of USA	discrete
Account Length	months of active usage	continuous
voice mail plan	Subscribed to voice mail	discrete
number vmail messages	number of voice-mail messages	continuous
international plan	Subscribed to international plan	discrete
total intl minutes	total number of international calls	continuous
total intl calls	total charge of international calls	continuous
total intl charge	total charge of international calls	continuous

Developing the churn prediction system on publicly available data helps to compare :

Table 4.2: Dimension descriptions contd.

Name of Dimension	Description	Type
total day minutes	total minutes of day calls	continuous
total day calls	total number of day calls	continuous
total day charge	total charge of day calls	continuous
total eve minutes	total minutes of evening calls	continuous
total eve calls	total number of evening call	continuous
total eve charge	total charge of evening calls	continuous
total night minutes	total minutes of night call	continuous
total night calls	total number of night calls	continuous
total night charge	total charge of night calls	continuous
number customer service calls	number of calls to customer service	continuous
churn value	if customer churned or not	discrete

- results
- ranking techniques

4.2 Data preprocessing and transformation

The data is in csv format and needs to be processed before loading. The data is loaded to the MySQL database for ease of access and retrieval. The data is loaded into table churn for access by R. A new data set containing the 5 regions of United States are used In MySQL a new is introduced which is the regions table. This is an additional data which is acquired from the google open source data.

In R the data is modified to add two more dimensions and drop two dimensions. The dimension that are irrelevant are :

- Phone number - not relevant since the column has all unique values
- Area code - not relevant since it is state specific and state is already represented

I intend to perform input discretization a process in which the continuous valued dimensions are to be transformed into discrete valued. In addition feature selection is a very important step which needs to be followed and chi-squared test and k-fold cross validations are to be incorporated. This is a necessary step since most of the irrelevant dimensions can be ignored and learning algorithms perform normally. Oversampling also is to be considered because the percentage of churners is quite less compared to the retained customers.

4.3 Data transformation

Dataset contains 5000 records and in order to train the machine learning models, data transformation needs to be done. Data set is split in to two categories Training set and testing set. It is recommended that a split of 75% to 25% be observed. A random function is used to select the indices of churn data set and the split is done.

4.4 Quantitative data analytics

The analysis of input churn data with statistical mathematical and computational techniques is presented below. The analysis of the dimensions of the data is as follows in table 4.3 :

Table 4.3: Dimension analysis

Dimension	min	1st Quart	median	mean	3rd Quart	max
state	na	na	na	na	na	na
account length	1	73	100	100.3	127	243
area code	na	na	na	na	na	na
phone number	na	na	na	na	na	na
internaltional plan	na	na	na	na	na	na
voice mail plan	na	na	na	na	na	na
umber vmail messages	0	0	0	7.75	17	52
total day minutes	0	143.7	180.1	180.3	216.2	351.5
total day calls	0	87	100	100	113	165
total day charge	0.00	24.43	30.62	30.65	36.75	59.76
total eve minutes	0.00	166.4	201.0	200.6	234.1	363.7
total eve calls	0.00	87.0	100.0	100.2	114.0	170.0
total eve charge	0.00	14.14	17.09	17.05	19.90	30.91
total night minutes	0.0	166.9	200.4	200.4	234.7	395
total night calls	0.00	87.0	100.00	99.92	113.0	175.0
total night charge	0.00	7.51	9.02	9.01	10.5	17.7
total intl minutes	0.00	8.50	10.3	10.26	12.0	20.0
total intl calls	0.00	3.00	4.00	4.435	6.00	20.0
total intl charge	0.00	2.30	2.70	2.771	3.24	5.4
number customer service calls	0.00	1.00	1.00	1.57	2.00	9.00
churn	na	na	na	na	na	na

Chapter 5

Development and evaluation of prediction models

This chapter presents the progress in development and evaluation statistics of prediction models.

5.1 Models trained

In the chapter the models trained thus far are Decision tree and Support vector machine.

But before the models are predicted the data set is divided into two

5.1.1 Decision tree

The decision tree is taken from the “rpart“ R library for training a classification tree. Confusion matrix in below table 5.1.

Table 5.1: Decision tree confusion matrix

Prediction	False	True
False	1266	109
True	19	106

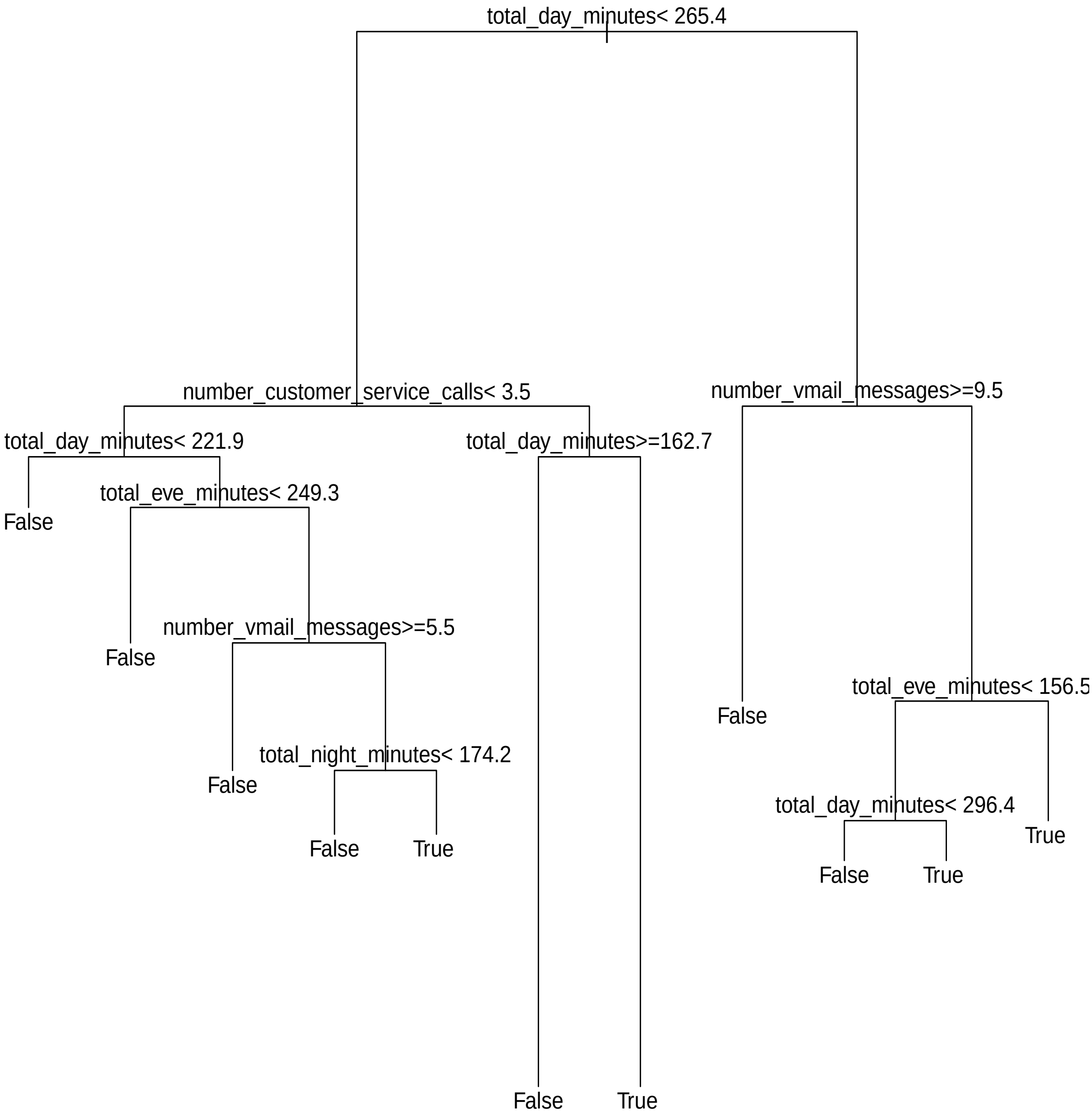
and the statistics are 5.2

Table 5.2: DT-1 Stats

Accuracy	:	0.9146667
95% CI	:	(0.8993698, 0.9283156)
No Information Rate	:	0.8566667
P-Value [Acc >NIR]	:	0.0000000000511805609
Sensitivity	:	0.9852140
Specificity	:	0.4930233

The figure of the decision tree in figure 5.1

Figure 5.1: Decision tree



5.1.2 Support vector machine

I have trained and tested two SVM's linear kernel SVM and Radial kernel. The following are the statistics from the training of the SVM's SVM linear kernel stats :

- Training sameple : 3500
- Testing sample : 1500
- 18 predictors
- 2 classes: 'False', 'True'
- Pre-processing: centered (69), scaled (69)
- Resampling: Cross-Validated (10 fold, repeated 3 times)
- Resampling results
  - Accuracy 0.8595245
  - Kappa 0.003825618

Confusion matrix linear kernel 5.3

Table 5.3: SVM Linear confusion matrix

Prediction	False	True
False	1285	215
True	0	0

Table 5.4: SVM Linear Stats

Accuracy	:	0.8567
95% CI	:	(0.8379, 0.874)
No Information Rate	:	0.8567
P-Value [Acc >NIR]	:	0.5182
Sensitivity	:	1
Specificity	:	0

SVM radial kernel statistics

- Training sameple : 3500
- Testing sample : 1500
- 18 predictors
- 2 classes: ’False’, ’True’
- Pre-processing: centered (69), scaled (69)
- Resampling: Cross-Validated (10 fold, repeated 3 times)
- Resampling results
  - Accuracy 0.8594297
  - Kappa 0

SVM radial kernel confusion matrix 5.5

Table 5.5: SVM Radial confusion matrix

Prediction	False	True
False	1266	68
True	19	147

Table 5.6: SVM Radial Stats

Accuracy	:	0.8566667
95% CI	:	(0.8379028, 0.8740215)
No Information Rate	:	0.8566667
P-Value [Acc >NIR]	:	0.5181819
Sensitivity	:	1
Specificity	:	0

5.1.3 Neural networks

To train the neural networks the churn data set needs to be transformed so that dimensions of data type factor are converted to numeric data type. This is my current

5.2 Dashboard of ICPCR

The dashboard is prepared in shiny and below are a few screen-shots.

Figure 5.2 enables the viewer to quickly scan the data set being used and understand the values of dimensions. A find search and sort functionality is also included.

Figure 5.2: Dashboard 1

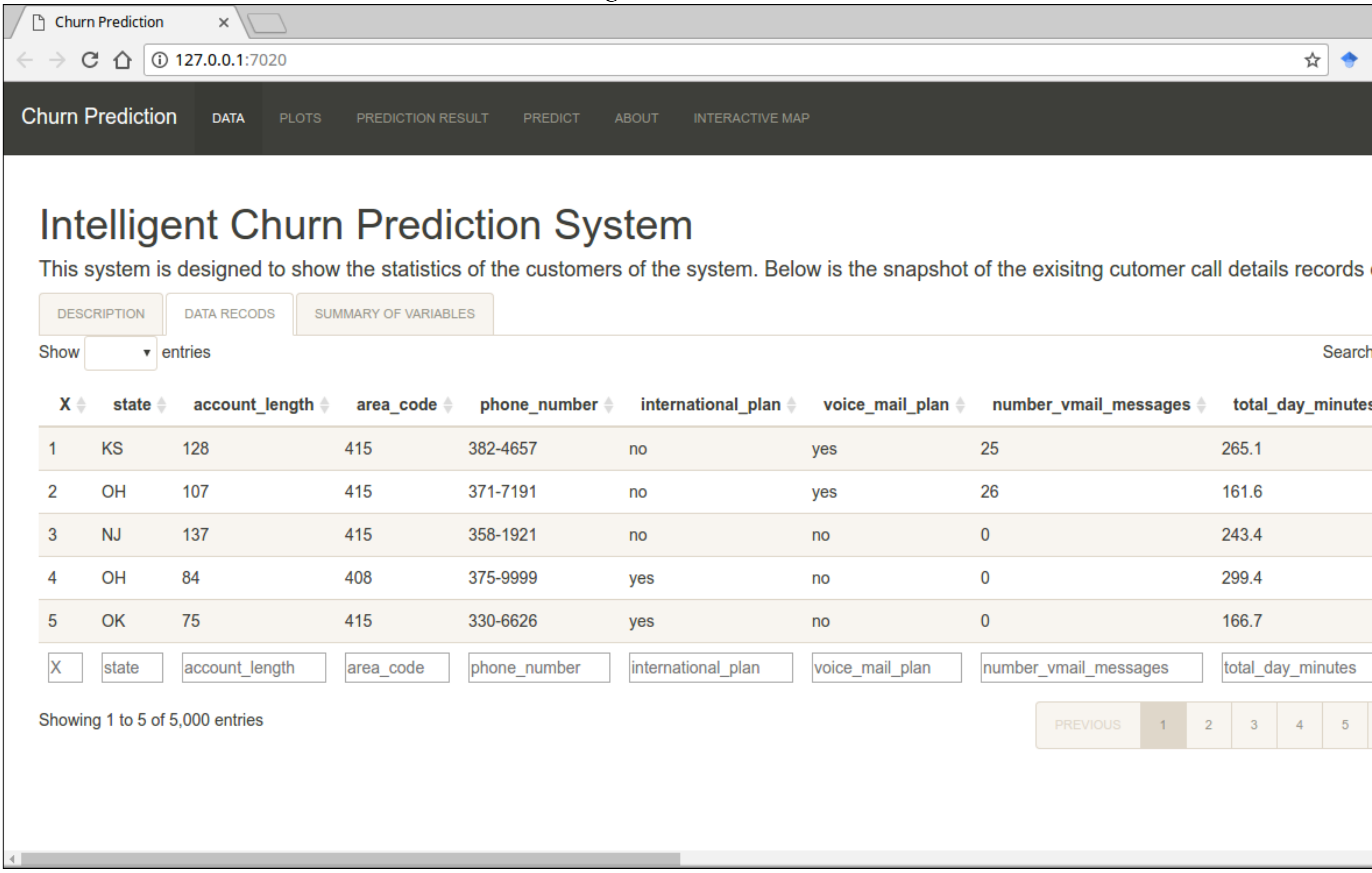


Figure 5.3 This describes the quantitative analysis of the dimensions.

Figure 5.3: Dashboard 2

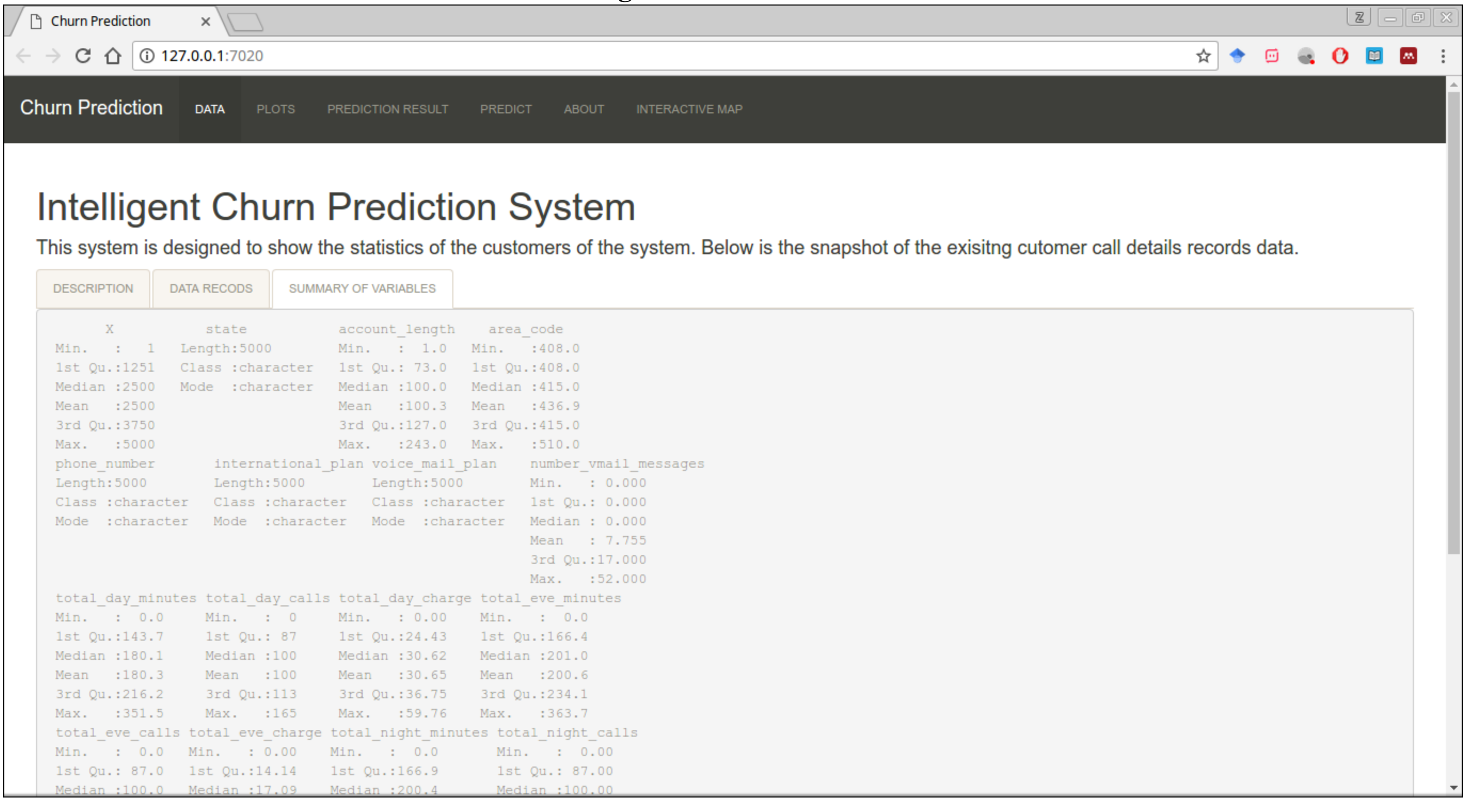


Figure 5.4 This displays the bar plot region wise for the churners to non-churners

Figure 5.4: Dashboard 3

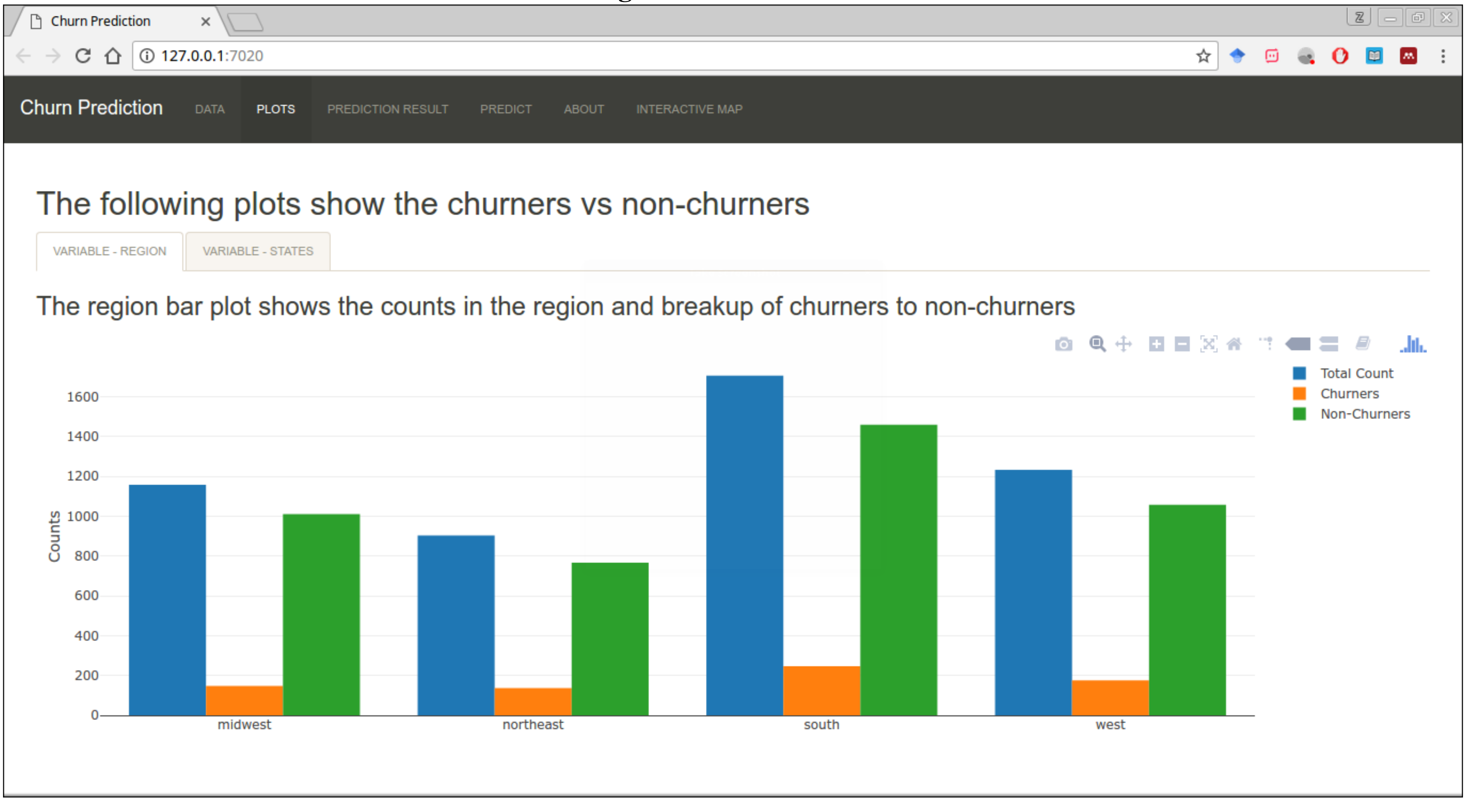


Figure 5.5 This displays the bar plot of the state wise churners to non-churners

Figure 5.5: Dashboard 4

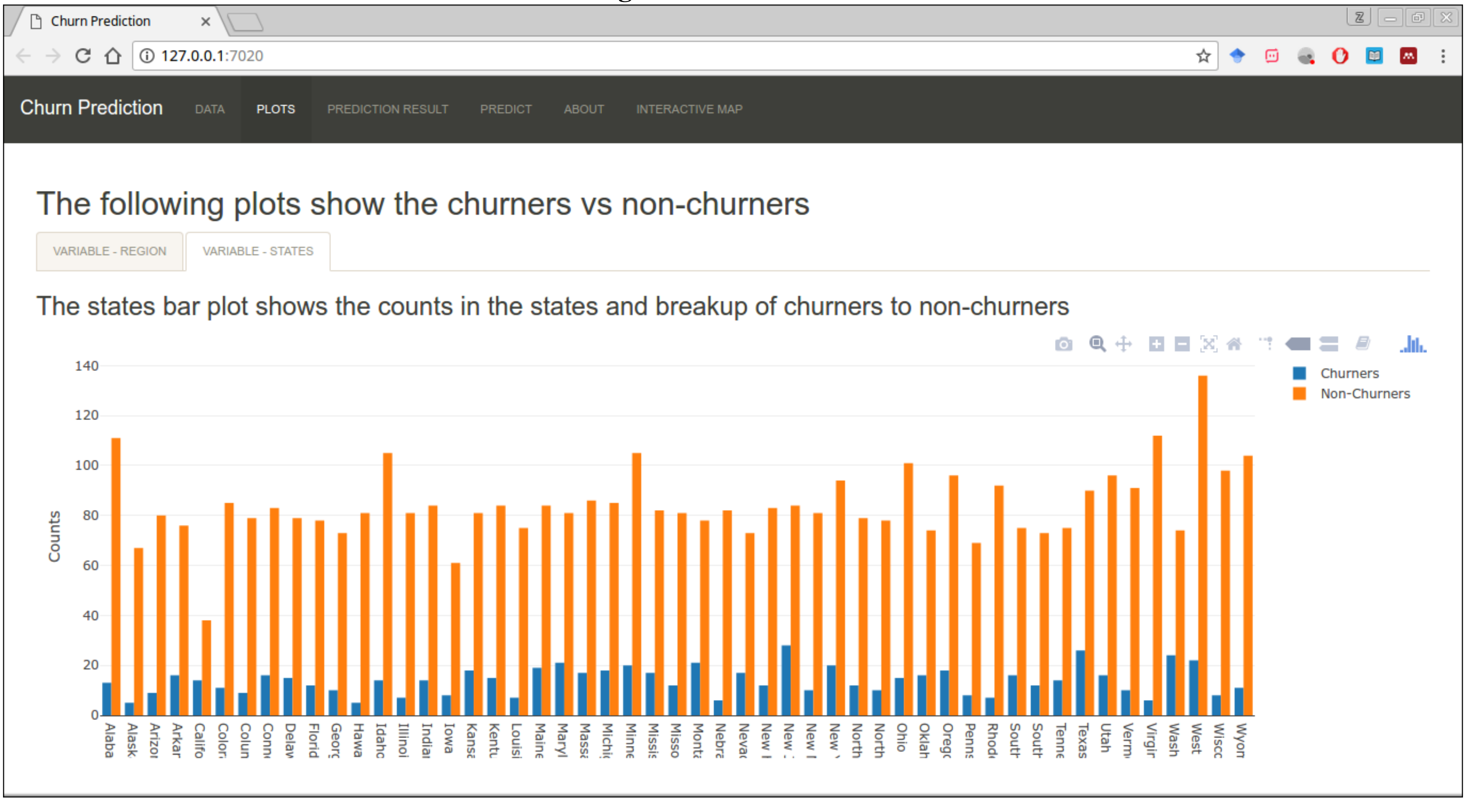


Figure 5.6 The screen shot is about a functionality to be implemented to enable the user to upload the data in CSV format to predict the churning of customers

Figure 5.6: Dashboard 5

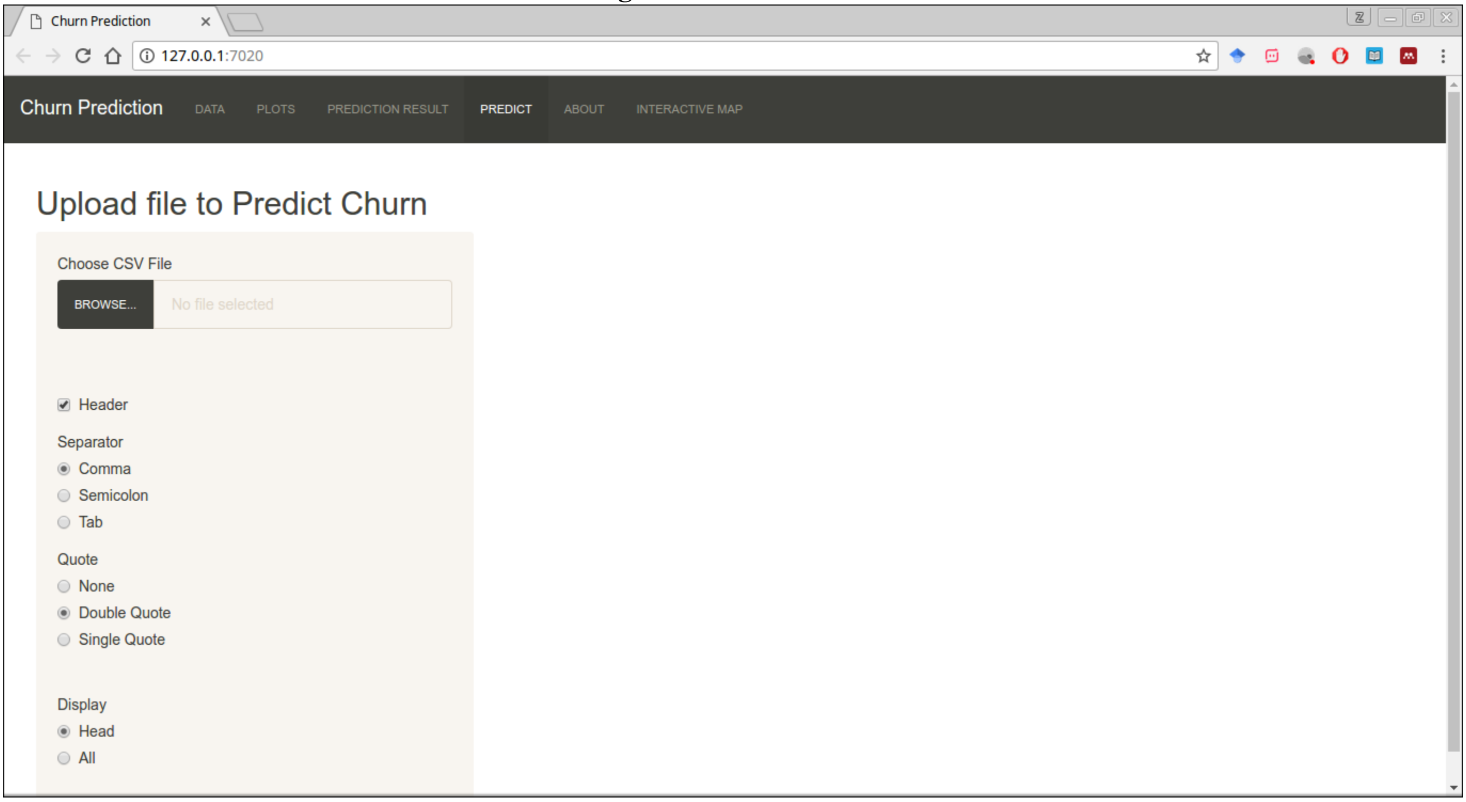
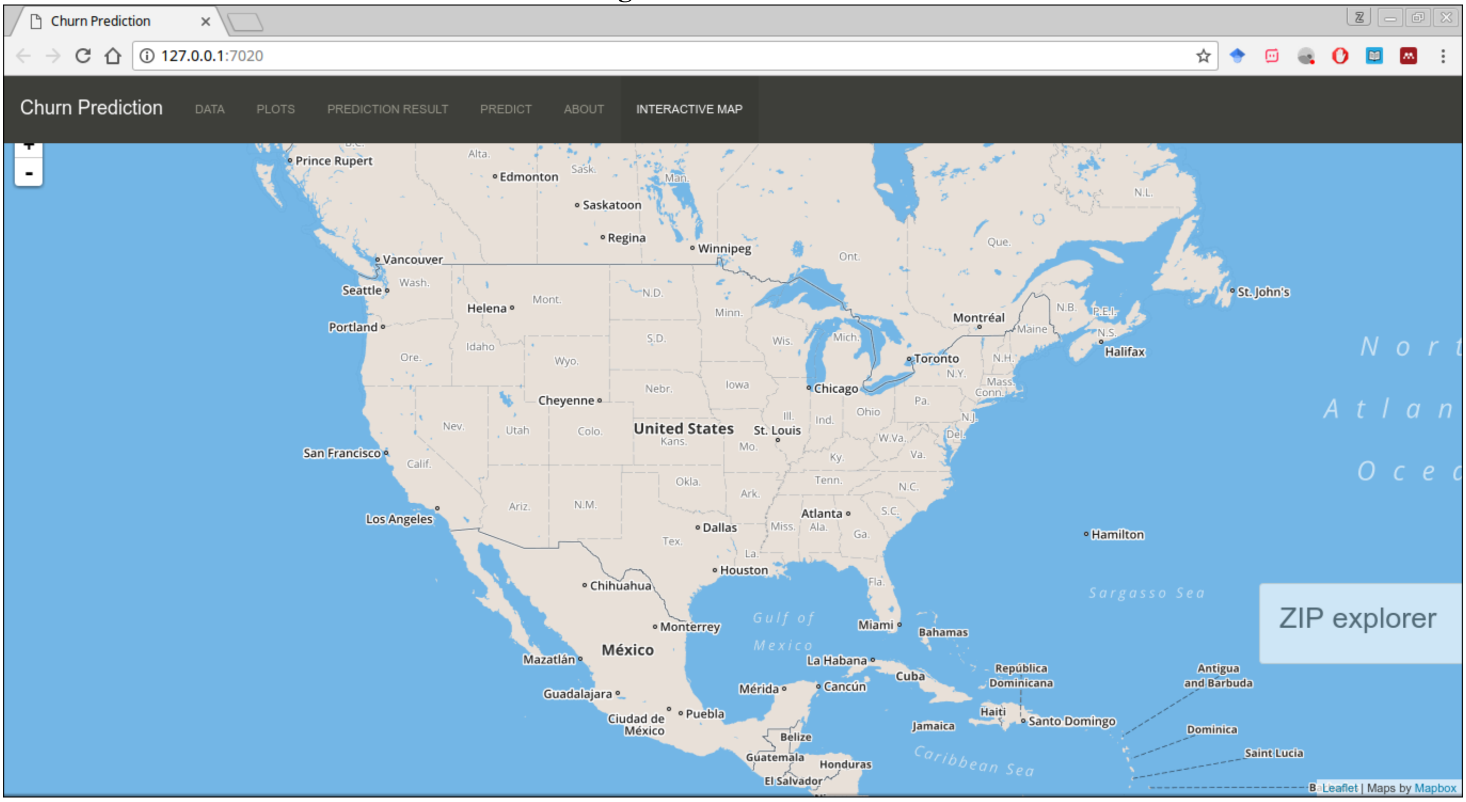


Figure 5.7 This functionality will be showing state wise distribution of telecom population and user will be able to customize to select churning or retaining populations.

Figure 5.7: Dashboard 6



References

Berson, A., Smith, S., & Thearling, K. (1999). *Building data mining applications for crm* (1st ed.). McGraw-Hill Professional.

Bhattacharya, C. B. (1998). When customers are members: Customer retention in paid membership contexts. *Journal of the Academy of Marketing Science*, *26*(1), 31-44.

Coussement, K., & Poel, D. Van den. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert systems with applications*, *34*(1), 313–327.

Cronin, F. J., Colleran, E. K., Herbert, P. L., & Lewitzky, S. (1993). Telecommunications and growth: The contribution of telecommunications infrastructure investment to aggregate and sectoral productivity. *Telecommunications policy*, *17*(9), 677–690.

Gerpott, T. J., Rams, W., & Schindler, A. (2001, may). Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market. *Telecommunications Policy*, *25*(4), 249–269.

Gibert, K., Sànchez-Marrè, M., & Codina, V. (2010). Choosing the right data mining technique: classification of methods and intelligent recommendation.

Han, J. (1997). Olap mining: An integration of olap with data mining. *2*, 1–9.

karpathy@cs.stanford.edu. (n.d.). *Cs231n convolutional neural networks for visual recognition.* `http://cs231n.github.io/neural-networks-1/`. (Accessed: 2017-03-29)

Kylin, A. (n.d.). *Apache kylin — home.* `http://kylin.apache.org/`. (Accessed: 2017-03-27)

Larivière, B., & Poel, D. Van den. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*, *29*(2), 472–484.

Liao, K.-H., & Chueh, H.-E. (2011). Applying fuzzy data mining to telecom churn management. 259–264.

Mizerski, R. W. (1982). Journal of consumer research. *An Attribution Explanation of the Disproportionate Influence of Unfavourable Information*, 301-310.

O, A. A., O, A. B., O, A. I., & R, A. E. (2015). Journal of Emerging Trends in Computing and Information Sciences Modeling & Simulation of a Predictive Customer Churn Model for Telecommunication Industry. *6*(11).

Olle, G. D. O., & Cai, S. (2014). A hybrid churn prediction model in mobile telecommunication industry. *International Journal of e-Education, e-Business, e-Management and e-Learning*, *4*(1), 55.

Poel, D. Van den, & Lariviere, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European journal of operational research*, *157*(1), 196–217.

PredictionIO. (n.d.). *Predictionio — a quick intro.* `http://predictionio.incubator.apache.org/start/`. (Accessed: 2017-03-27)

Reinartz, W. J., & Kumar, V. (2003). The impact of customer relationship characteristics on profitable lifetime duration. *Journal of marketing*, *67*(1), 77–99.

Reuters. (2017). *India's reliance jio signs up 72 million paying customers — reuters.* http://www.reuters.com/article/us-reliance-jio-prime-idUSKBN1722BJ. (Accessed: 2017-04-01)

Rstudio, S. by. (n.d.). *Shiny - scaling and performance tuning with shiny apps.* https://shiny.rstudio.com/articles/scaling-and-tuning.html. (Accessed: 2017-03-27)

Scikit. (n.d.). *Scikit-learn choose the right estimator.* http://scikit-learn.org/stable/tutorial/machine_learning_map/. (Accessed: 2017-03-27)

TRAI - Telecom Regulatory Authority of India. (n.d.). *Telecom Subscriptions Reports — Telecom Regulatory Authority of India.* http://www.trai.gov.in/release-publication/reports/telecom-subscriptions-reports. (Accessed: 2017-04-01)

Tsai, C.-F., & Lu, Y.-H. (2009). Customer churn prediction by hybrid neural networks. *Expert Systems with Applications*, *36*(10), 12547–12553.

Tutorials Point. (n.d.). *Data warehousing schemas.* https://www.tutorialspoint.com/dwh/dwh_schemas.htm. (Accessed: 2017-03-31)

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, *55*, 1–9.

Wei, C.-P., & Chiu, I.-T. (2002). Turning telecommunications call details to churn prediction: a data mining approach. *Expert systems with applications*, *23*(2), 103–112.

Xie, Y., Li, X., Ngai, E., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, *36*(3), 5445–5449.