

Churn Prediction Model Using Linear Discriminant Analysis (LDA)

Naveen Kumar Rai¹, Vikas Srivastava², Rahul Kumar³

¹(Department of Information Technology, Birla Institute of Technology, Mesra, India)

²(Department of Information Technology, MCKVIE Vidya Sagar University, India)

³(Department of Electronics, Priyadarshini College of Engineering and Architecture, India)

Abstract: Customer churn refers to customers terminating the service contract with the company or turning to services provided by the other company. Churn analysis is the calculation of the rate of attrition in the customer base of any company. It involves identifying those consumers who are most likely to discontinue using your service or product. Churn rate is a measure of customer or employee attrition, and is defined as the number of customers who discontinue a service or employees who leave a company during a specified time period divided by the average total number of customers or employees over that same time period and the customer who discontinue a service or leave are simply called Churners.

Keywords: Churn, LDA, PDA, Churn Prediction, Modeling, Prediction Algorithms

I. Introduction

Prediction of the churn and non-churn customers according to the given inputs among given set of customers and identify the reasons of churn Predictive modeling is the process by which a model is created or chosen to try to best predict the probability of an outcome. In many cases the model is chosen on the basis of detection theory to try to guess the probability of an outcome given a set amount of input data.

Customers become “churners” when they discontinue their subscription and move their business to a competitor. Different kinds of churn:

1. Voluntary churn: This includes customers, of their own free will, decide to take their business elsewhere.
2. Involuntary Churn: It is also known as forced attrition, occurs when the company rather than the customers, terminates the relationship-most commonly due to unpaid bills.
3. Expected churn: It occurs when the customer is no longer in the target market for a product.

II. Churn analysis applications

Churn analysis is necessary to handle the crucial problem of commercial companies in general and telecommunication companies in particular suffer from is a loss of valuable customers to competitors, that is, it increases churners.

One needs to understand the behavior of customers, and classify the churn and non-churn customers, so that the necessary decisions will be taken before the churn customers switch to a competitor.

Churn Analysis is important because lost customers must be replaced by new customers, and new customers are expensive to acquire and generally generate less revenue in the near term than established customers.

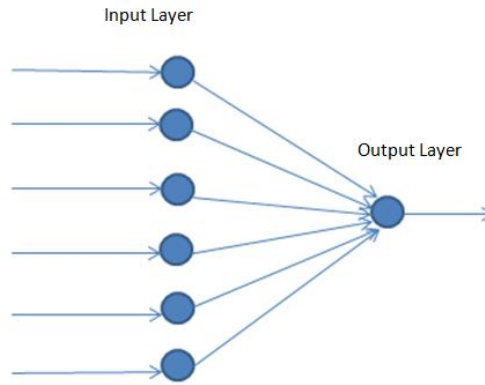
To know why customers are changing into ‘churners’.

III. Aim and proposed model

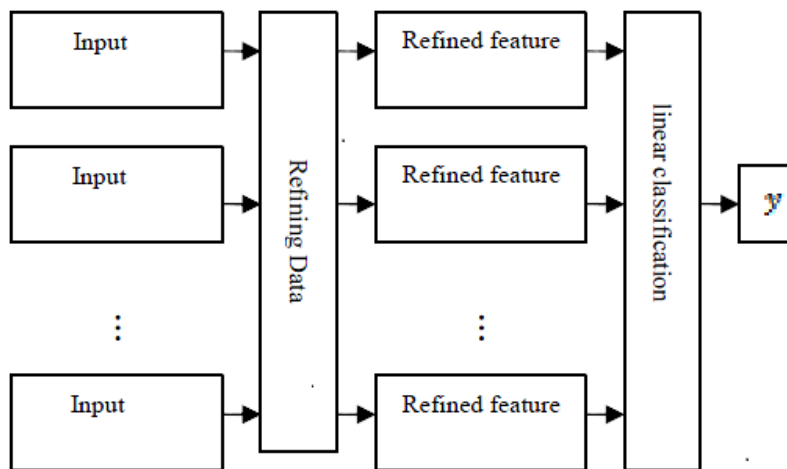
Developing a Churn Analysis system based upon data mining technology to analyze the customer database and predict churn for mobile and telecommunication. Building a new corporate Customer Data Warehouse aimed to support Marketing and Customer Care areas in their initiatives. Classify the churn and non-churn customers, so that the necessary decisions can be taken before the churn customers switch to a competitor. Our main aim is to design a model which can predict whether a customer is churner or not.

Propose an algorithm which can classify the customers into Churners and non-Churners.

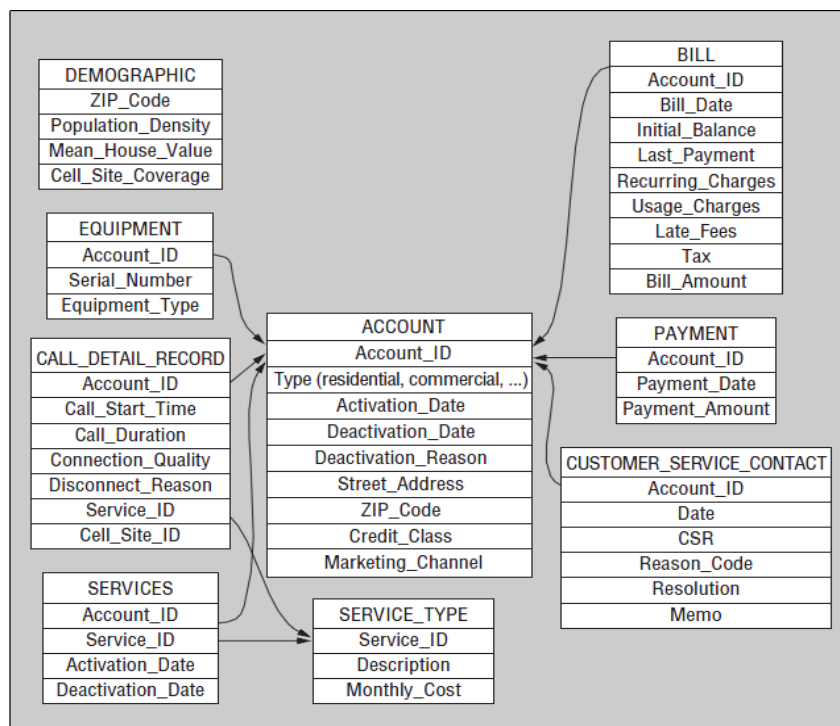
- Our proposed model uses Linear Discriminant Analysis (LDA) to predict the churn.
- It is a two layer network consisting one input layer and one output layer as shown below:



Using data mining we obtained a number of attributes related to customers but not all of them are selected in our model as input, only some features were selected which could affect the prediction of churners.



Customer data obtained from data warehouse using data mining are classified as in the following tables



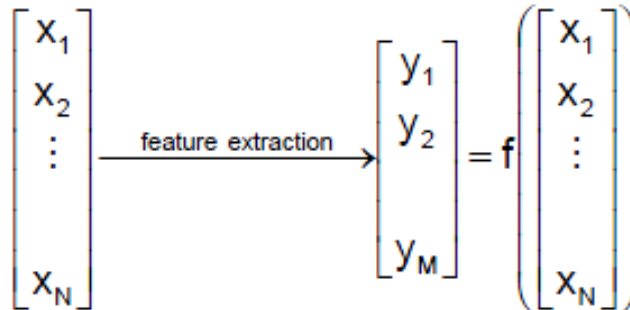
IV. Churn prediction algorithm

Feature extraction: creating a subset of new features by combinations of the existing features.

- **The problem of feature extraction can be stated as:**

Given a feature space $X_i \in \mathbb{R}_N$ find a mapping $y=f(x): \mathbb{R}_N \rightarrow \mathbb{R}_M$ with $M < N$ such that the transformed feature vector $Y_i \in \mathbb{R}_M$ preserves (most of) the information or structure in \mathbb{R}_N .

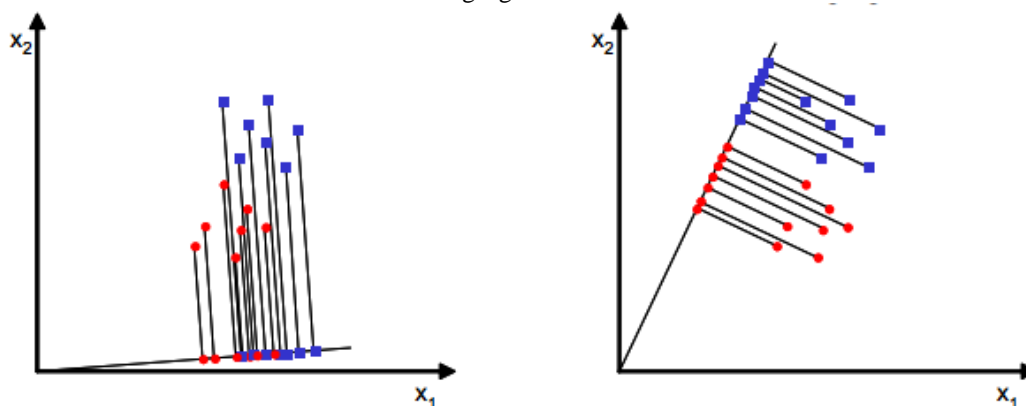
We manually chose the attributes which could affect the prediction.



The types of information that are useful for a churn analysis include:

- Call statistics: length of calls at different times of the day, number of long distance and local calls like day minutes, initialization minutes, night call, day call.
- Billing information for each customer – what the customer is paying for local and long distance like day charges, evening charges, and night charges.
- Extra service information, that is, what extra plan the customer is registered on, e.g. special long distances rates like voice mail, voice-mail messages, credit history .
- Voice and data product and services purchased by the customer, e.g., broad-band services, private virtual networks, dedicated data transport links, service calls etc.
- The objective of this algorithm is to perform dimensionally reduction while preserving as much of the class discriminatory information as possible
- We have a set of M -dimensional samples $\{ x^{(1)}, x^{(2)}, \dots, x^{(N)} \}$, N_1 of which belong to class ω_1 (Non-Churners), and N_2 to Class ω_2 (churners).
- Here M is the no. of attributes that we have obtained after refining the whole information and N is the number of samples taken for training.
- **In general, the optimal mapping $y=f(x)$ will be a non-linear function**
- However, there is no systematic way to generate non-linear transforms. The selection of a particular subset of transforms is problem dependent
- For this reason, feature extraction is commonly limited to linear transforms: $y=Wx$.
- That is, y is a linear projection of x .
- We seek to obtain a scalar y by projecting the samples x onto a line

$$y=W^T X$$
- Of all the possible lines we would like to select the one that maximizes the separability of the scalars
 This is illustrated for the 2-D case in following figure



- In order to find a good projection vector , we need to define a measure of separation between the projections

The mean vector of each class in x and y feature space is:-

$$\mu_i = \frac{1}{N_i} \sum_{x \in \omega_i} x \quad \text{and} \quad \tilde{\mu}_i = \frac{1}{N_i} \sum_{y \in \omega_i} y = \frac{1}{N_i} \sum_{x \in \omega_i} w^T x = w^T \mu_i$$

- we could then choose the distance between the projected means as our objective function

$$J(w) = |w^T(\mu_1 - \mu_2)|$$

- However, the distance between the projected means is not a very good measure since it does not take into account the standard deviation within class
- The solution proposed is to maximize a function that represents the difference between the means, normalized by a measure of the within-class scatter

For each class we define the scatter, an equivalent of the variance, as

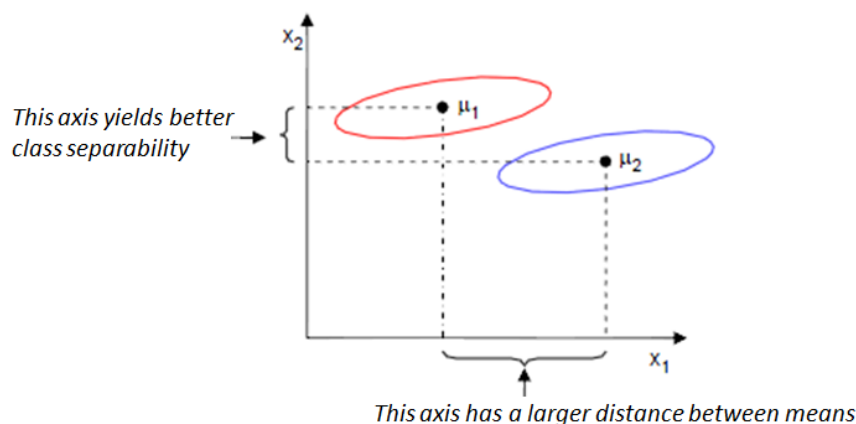
$$\tilde{S}_i^2 = \sum_{k \in \omega} (y_k - \tilde{\mu}_i)^2$$

where the quantity \tilde{S}_i^2 is called the within class scatter of the projected examples.

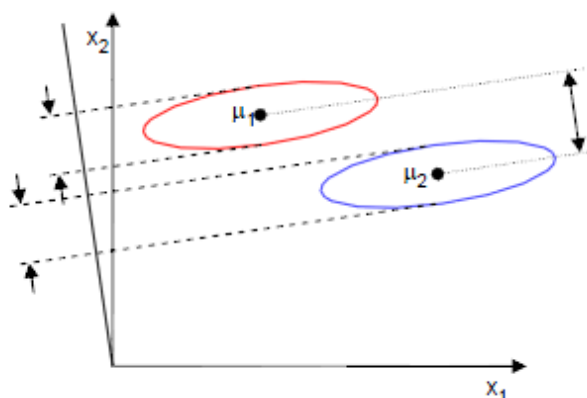
$$J(w) = |\tilde{\mu}_1 - \tilde{\mu}_2|^2 / (\tilde{S}_1^2 + \tilde{S}_2^2)$$

therefore, we will be looking for a projection where examples from the same class are projected very close to each other and, at the same time, the projected means are as far apart as possible

Analysis without taking into account the standard deviation



Analysis of projection considering both means and standard deviation



- In order to find the optimum projection w^* , we need to express $J(w)$ as an explicit function of w , We define a measure of the scatter in multivariate feature space x , which are scatter matrices

$$S_i = \sum_{k \in \omega} (x - \mu_i)(x - \mu_i)^T$$

$$S_1 + S_2 = S_w$$

Where S_w is called the within class scatter matrix.

- The scatter of the projection y can then be expressed as a function of the scatter matrix in feature x

$$\tilde{S}_i^2 = \sum_{k \in \omega} (y - \tilde{\mu}_i)^2 = \sum_{k \in \omega} (w^T x - w^T \mu_i)^2 = \sum_{k \in \omega} w^T (x - \mu_i)(x - \mu_i)^T w = w^T S_i w$$

$$(\tilde{S}_1^2 + \tilde{S}_2^2) = w^T S_i w$$

- Similarly, the difference between the projected means can be expressed in terms of the means in the original feature space

$$(\tilde{\mu}_1 - \tilde{\mu}_2)^2 = (w^T \mu_1 - w^T \mu_2)^2 = w^T (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w = w^T S_B w$$

The matrix S_B is called the between class scatter. Note that, since S_B is the outer product of two vectors, its rank is at most one

- We can finally express the Fisher criterion in terms of S_w and S_B as

$$J(w) = w^T S_B w / w^T S_w w$$

- To find the maximum of $J(w)$ we derive and equate to zero

$$\frac{d}{dw}[J(w)] = \frac{d}{dw}[w^T S_B w / w^T S_w w] = 0$$

$$\Rightarrow [w^T S_w w] \frac{d}{dw}[w^T S_B w] - w^T S_B w \frac{d}{dw}[w^T S_w w] = 0$$

$$\Rightarrow [w^T S_w w] 2S_B w - [w^T S_B w] S_w w = 0$$

- Dividing by $w^T S_w w$

$$\Rightarrow [w^T S_B w / w^T S_w w] S_B w - [w^T S_B w / w^T S_w w] S_w w = 0$$

$$\Rightarrow S_B w - J S_w w = 0$$

$$\Rightarrow S_w^{-1} S_B w - J w = 0$$

- Solving the generalized eigenvalues problem ($S_w^{-1} S_B w = J w$) yields

$$W^* = \operatorname{argmax}\{w^T S_B w / w^T S_w w\} = S_w^{-1} (\mu_1 - \mu_2)$$

Use of weight Matrix

- W^* is the weight matrix for the attributes taken as input. Now this would be used to calculate the y_1 and y_2 values for the two classes by using $y = W^T X$
- Then mean of attributes $Y(\text{mean})$ of y_1 and y_2 will be calculated and then the mean of y_1 and y_2 will be selected as the threshold for the customers to be classified as churners and non-churners.
- The customers whose Y value are less than $Y(\text{mean})$ will be classified as churners and those whose Y value are greater than $Y(\text{mean})$ will be considered as non-churners.

Process of classification

- We have a customer sample dataset of a telecommunication company.
- Now this data set consists of various information about the customer. Here ,our aim is to develop a model that will analyze the nature of the behavior on the basis of previous data and predict the behavior of the customer that is basically related to churning behavior of the existing customer.
- Now, we develop a particular model for prediction of such behavior by training our model.
- During this training process, we adjust the weight of several attributes of the customer and the value of weight keeps on improving as our training process progresses .In our case, we train our model with the help of two-third of the data set and adjust the values for weight of all attributes and normalize it.
- We select the attributes from the available one on which our model depends a lot on the basis of values of weight for several attributes.
- This selected attributes were used to plot three-dimensional graph to represent the nature of customer under classifications that is churners and non-churners.
- Then, with our developed model, we try to predict the nature of customers and compare it with the actual one. This particular process is considered as the testing phase. In our case, we test our model with one-third of the available data-set .
- Then we find percentage accuracy of our model by calculating and plotting true positive, true negative ,false positive ,false negative.
- If this percentage is high enough, our model is acceptable else we need to improve our model either by incorporating new attributes or by training the model more suitably.
- Finally, we come out with a model that predicts churning behavior of the customer.

V. Experiment Result

Linear Discriminant Vs. Principle Component Analysis

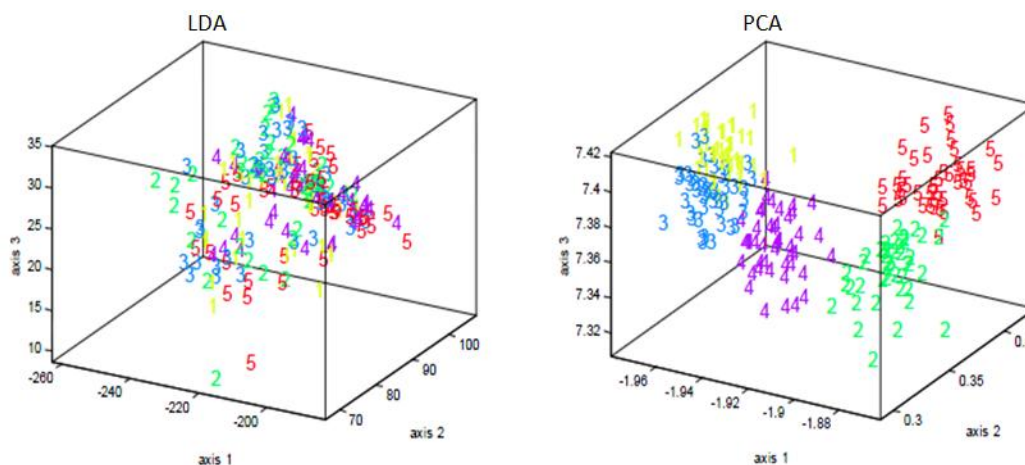
Coffee discrimination with a gas sensor array

Five types of coffee beans were presented to an array of chemical gas sensors

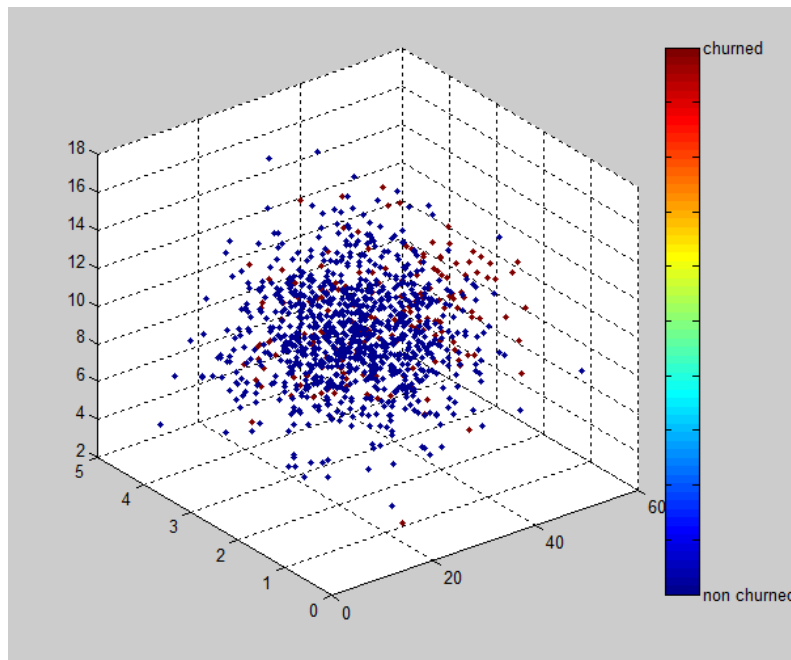
- For each coffee type, test were performed and the response of the gas sensor array was processed in order to obtain a 60-dimensional feature vector

Results

- From the 3-D scatter plots it is clear that LDA outperforms PCA in terms of class discrimination
- This is one example where the discriminatory information is not aligned with the direction of maximum variance
- These figures show the performance of PCA and LDA on an odour recognition problem

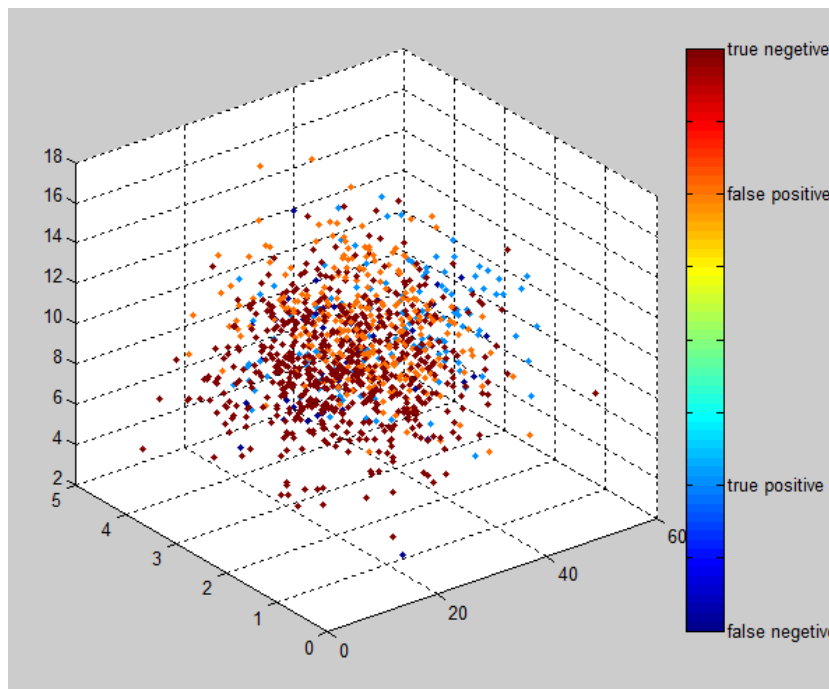


Output -1



This is the scatter plot of the customers classified by the model developed as churners and non-churners.

Output-2



- True-positive denote the fraction of churned customers who were perfectly identified by the model
- False-positive denote the fraction of churned customers who were identified by the model as non-churn customers
- True-negative denote the fraction of non-churned customers who were perfectly identified by the model
- False-negative denote the fraction of non-churned customers who were identified by the model as churners.

```

w =
  1.0e+004 *
  Columns 1 through 7
    0.0529   -0.0011   -0.1793    0.0001    1.0531   -0.0474   -0.0001
  Columns 8 through 14
    0.5554    0.0254   -0.0000   -0.5652    0.1830    0.0019   -0.6900
  Column 15
   -0.0120

tp =
    0.7416

fp =
    0.2584

tn =
    0.7031

fn =
    0.2969
  
```

This shows that:

- 74 % of the customers were positively identified by the model as churners.
- 26 % of the customers were negatively identified by the model as churners, who actually did not churn.
- 70 % of the customers were positively identified by the model as non-churners.
- 30% of the customers were positively identified by the model as non-churners, who actually churned.

VI. conclusion

- In our model, we have incorporated the following valuable information about the customers that serves as input in our developed model : voice-mail, voice-mail messages, day-minutes, day calls ,day charges, evening minutes, evening charges, night calls, night charges, initialization minutes, initialization calls, initialization charges, service calls.
- Firstly our model is trained on the certain part of the customer data and the required weight is calculated for various attribute in our model, which gets on improved as training progresses. At the end of the training we have developed model that takes input and imparts the probability for the customer whether they will churn in near future or not.
- We matched our result with the actual result about the churning behaviour of the customers and we successfully came out with an accuracy of 74% . Thus, finally we came out with a model that studies and predict the behaviour of customer churning and classify them as churners and non-churners.

Limitations of the model

- It did not include market value, sales trend and other factors which affect the churn prediction.
- It has accuracy up to 74%.

References

Books:

- [1] A New Prediction Model of Customer Churn Based on PCA Analysis By Zhao Xin, Wang Yi and chahongwang
- [2] Introduction to pattern analysis , Lecture Texas A&M university
- [3] Data Mining Techniques , By Michael J.A. Berry and Gordon S. Linoff
- [4] Classification of Churn and non-Churn Customers for Telecommunication Companies, by Tarik Rashid (*Computing Faculty/Research and Development Department College of Computer Training (CCT), Ireland.*)