REPORT

Comparison of 3 Algorithms on Bank Churn Data Using CAP curve

Submitted in partial fulfilment of the requirement for the award of the degree
Of
Bachelor of Technology
Computer Science Engineering and Information Technology
By
Shefali Singhal(13ITU035)
Shikhar Agarwal(13ITU036)

Under supervision of
Dr. Geetika Munjal
Assistant professor

Department of CSE & IT
THE NORTHCAP UNIVERSITY
Gurgaon

CERTIFICATE

This is to certify that the Project Report entitled, "Comparison of 3 algorithms on Bank Churn Data using CAP curve" submitted by **Shefali Singhal, Shikhar Agarwal** to **THE NORTHCAP UNIVERSITY, Gurgaon, India**, is a record of bona fide project work carried out by him/her under my/our supervision and guidance and is worthy of consideration for the award of the degree of **Bachelor of Technology** in **Computer Science Engineering / Information Technology** of the Institute.

Dr. Geetika Munjal
Assistant Professor

Date: 20/03/17

Acknowledgment

The satisfaction that accompanies successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success.

This report cannot be realized without help from numerous sources and people in the organization. We take this opportunity to express our profound sense of gratitude and respect to all those who helped us throughout the report. With reverence and veneration honor, we acknowledge all those who's guidance and encouragement has made this report successful.

This project report is the result of the dedication and encouragement of many individuals. Our sincere and heartfelt appreciation goes to all of them. We would like to thank Ms Geetika Munjal for giving us useful tips for the project. We have endeavored to present this in most clear and interesting way. We express our heartfelt thanks and gratitude to the university for giving us an opportunity to undertake this project.

Abstract

In this paper, we wish to put forth the results of prediction-accuracy obtained on a churn dataset, that is, a data which has information about people and their churning decisions, by using three different machine learning algorithms, namely ANN, Logistic regression and KNN, and show which one performed the best out of these. The results obtained will be useful for industries who wish to do future prediction on similar data, by providing them with a clear answer to which algorithm works better on this type of data and why.

TABLE OF CONTENTS

1 INTRODUCTION

1.1 Background

Churn modelling, also known as customer attrition, is the organisation's loss of customer or clients from their customer base. The organisation can be anyone which deals with clients. In our case, it's a private bank whose customer attrition has been predicted by data mining algorithms. Organisations often use customer attrition analysis as one of their key business metrics because the cost of retaining existing customers is way less than acquiring new ones. Complex algorithms can also be used to find out if there are any co-relations among the demographic attributes of the customers and their churning likelihood, like being female, or being in a particular country, or having a particular salary, or things like that, although co-relation analysis is not the aim of our research, but this is also something which one can explore. Data analysts, after identifying the potential churners, can forward the list of customers to strategists, who then can look for strategies for keeping the customers in, like cold calling, providing them with attractive schemes, or doing anything which keeps their interest in the company going untill the company wants it to. Thus the aim of predicting is to:-

- Anticipate which customers are to leave, and often answer questions like: Who are they? or What is their behaviour(like their shopping patterns, their relations with the company officials, response to special offers etc.)
- Implement customised retention schemes, strategies so as to avoid their migration, increasing the capability to take right actions on such occasions.

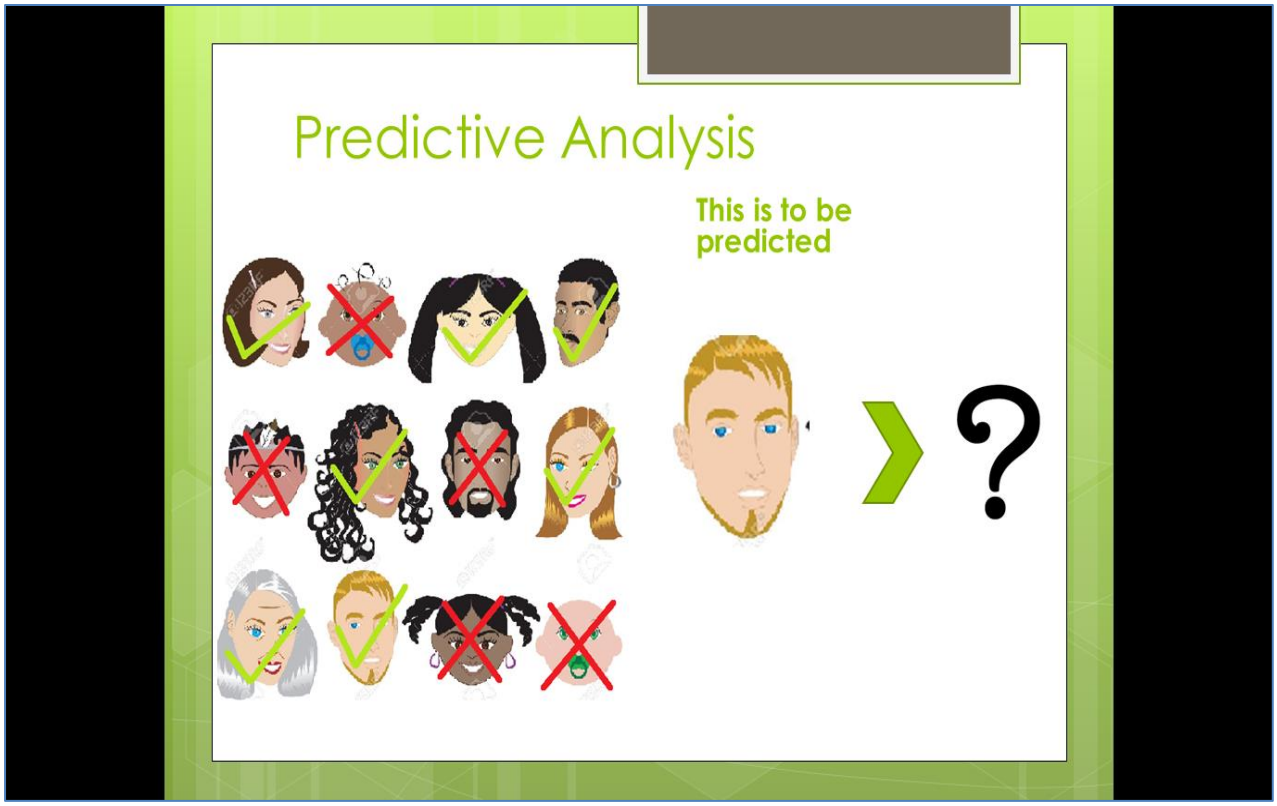
This is a great asset to keeping a company's customer base large, thus serving as a motivation of ours to pick up this particular topic for exploration and comparing algorithms to find which one best predicts the potential churners.

1.2 Motivation

Dire need of the churn modelling of the bank data by the industry is the reason and the motivation for the completion of this project. There is a high demand by the retail industry and bank industry to understand the behavior of their customers. This project can also be used by the airline industry to churn the customers from there dataset. As such the data analysis is an upcoming field and there are a variety of dimensions in which we could have worked but the reason for choosing the churn modelling is because this research hasn't been done before on churn data.

2 PREDICTIVE ANALYSIS DEMOSTRATION

2.1 A diagram showing what predictive analysis does. We have a sample data set of customers and their corresponding behavioural patterns with regards to their banking, and we have the answer to whether they exit their membership from the bank or continue it. Using this data, we need to find whether a particular employee for example John, will exit the membership from the bank or not.



3 SCREENSHOT OF DATA SET

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
6	RowNumb	Exited	P_Hat	TotalReco	Total	RandomSe	RandomSe	ModelSele	ModelSele	model2	exitedknn	P_hat	modelsel	modelselp	exitedlogit	modelsel	modelselp
7	10958	1	0.868685	1	0.10%	0.26	0.10%	1	0.40%	0.865561	1	1	1	0.38%	1	1	0.38%
8	10250	1	0.861297	2	0.20%	0.52	0.20%	2	0.80%	0.841125	1	1	2	0.77%	1	2	0.77%
9	10430	1	0.832155	3	0.30%	0.78	0.30%	3	1.20%	0.83711	1	1	3	1.15%	1	3	1.15%
10	10210	1	0.824039	4	0.40%	1.04	0.40%	4	1.50%	0.821941	1	1	4	1.54%	1	4	1.54%
11	10694	0	0.817353	5	0.50%	1.3	0.50%	4	1.50%	0.819837	1	1	5	1.92%	0	4	1.54%
12	10998	0	0.812318	6	0.60%	1.56	0.60%	4	1.50%	0.811865	1	1	6	2.31%	0	4	1.54%
13	10945	1	0.808891	7	0.70%	1.82	0.70%	5	1.90%	0.796419	1	1	7	2.69%	1	5	1.92%
14	10893	1	0.806505	8	0.80%	2.08	0.80%	6	2.30%	0.794847	1	1	8	3.08%	1	6	2.31%
15	10741	1	0.791722	9	0.90%	2.34	0.90%	7	2.70%	0.788734	1	1	9	3.46%	1	7	2.69%
16	10932	1	0.781573	10	1.00%	2.6	1.00%	8	3.10%	0.778494	1	1	10	3.85%	1	8	3.08%
17	10851	1	0.761677	11	1.10%	2.86	1.10%	9	3.50%	0.759869	0	1	10	3.85%	1	9	3.46%
18	10627	1	0.758891	12	1.20%	3.12	1.20%	10	3.80%	0.755093	0	1	10	3.85%	1	10	3.85%
19	10722	1	0.751711	13	1.30%	3.38	1.30%	11	4.20%	0.753781	1	1	11	4.23%	1	11	4.23%
20	10976	1	0.745922	14	1.40%	3.64	1.40%	12	4.60%	0.750912	1	1	12	4.62%	1	12	4.62%
21	10687	1	0.734067	15	1.50%	3.9	1.50%	13	5.00%	0.738894	1	1	13	5.00%	1	13	5.00%
22	10425	1	0.731534	16	1.60%	4.16	1.60%	14	5.40%	0.725545	1	1	14	5.38%	1	14	5.38%
23	10835	0	0.718687	17	1.70%	4.42	1.70%	14	5.40%	0.710941	1	1	15	5.77%	0	14	5.38%
24	10258	0	0.714648	18	1.80%	4.68	1.80%	14	5.40%	0.706889	1	1	16	6.15%	0	14	5.38%
25	10118	1	0.704092	19	1.90%	4.94	1.90%	15	5.80%	0.703609	1	1	17	6.54%	1	15	5.77%
26	10261	1	0.703917	20	2.00%	5.2	2.00%	16	6.20%	0.697357	1	1	18	6.92%	1	16	6.15%
27	10262	0	0.693596	21	2.10%	5.46	2.10%	16	6.20%	0.683457	1	1	19	7.31%	0	16	6.15%
28	10428	1	0.691122	22	2.20%	5.72	2.20%	17	6.50%	0.67399	1	1	20	7.69%	1	17	6.54%

4 DATA SET EXPLAINED

S.No	Attribute	Description
1	Row Number	Row number
2	Customer ID	Customer ID in the bank books
3	Surname	Family name of the customer (as an International standard)
4	CreditScore	Your trust score taken in terms of returning credit/loan
5	Geography	Country among Germany, or France
6	Gender	Male/Female
7	Age	Age in years
8	Tenure	Tenure in years with the bank
9	NumOfProducts	Number of products/services with the bank like loan, savings
10	HasCrCard	Has a credit card or not.
11	IsActiveMember	Is active or not, in terms of debit credit.
12	EstimatedSalary	Salary as estimated, as real salaries not available.
13	Exited	Our class parameter, as to exiting the bank or not.

As part of pre-processing the data, we have excluded Customer ID and surname as they are of no use. Geography is taken to be 1 for Germany and 0 for France and Spain as it was found by co-relation analysis that being or not being in Germany highly influences the exit decision. Gender is taken as 1 for female and 0 for male. All the parameters are then 'min-max' normalised between 0 and 1 using the formula $x' = (x - Min) / Max$, where x is the un-normalised value, Min is the minimum of all values of the attribute, max is the maximum.

5 FEATURE SELECTION USING P-VALUES

In general the p-value indicates how probable a given outcome or a more extreme outcome is under the null hypothesis. In our case of feature selection, the null

hypothesis is something like *this feature contains no information about the prediction target*, where *no information* is to be interpreted in the sense of the scoring method: If your scoring method tests e.g. univariate linear interaction then the null hypothesis says that this linear interaction is not present.

The p-value of a feature selection score indicates the probability that **this score or a higher score** would be obtained if this variable showed no interaction with the target.

Another general statement: **scores** are *better if greater*, **p-values** are *better if smaller* (and **losses** are *better if smaller*)

6 P-VALUES OBTAINED USING GRETL SOFTWARE

Model 1: Logit, using observations 1-10000

Dependent variable: Exited

Standard errors based on Hessian

	Coefficient	*Std. Error*	*z*	*p-value*	
const	−3.99238	0.232615	−17.1630	<0.0001	***
CreditScore	−0.0006743 8	0.00028021 5	−2.4067	0.0161	**
Age	0.0726405	0.00257405	28.2203	<0.0001	***
NumOfProducts	−0.095494	0.0475089	−2.0100	0.0444	**
IsActiveMember	−1.07253	0.0575976	−18.6210	<0.0001	***
Germany	0.746303	0.0650378	11.4749	<0.0001	***
Female	0.528301	0.054444	9.7036	<0.0001	***
log_balance	0.0690313	0.0139553	4.9466	<0.0001	***

Mean dependent var	0.203700	S.D. dependent var	0.402769
McFadden R-squared	0.152501	Adjusted R-squared	0.150919
Log-likelihood	−4284.015	Akaike criterion	8584.029
Schwarz criterion	8641.712	Hannan-Quinn	8603.554

Number of cases 'correctly predicted' = 8114 (81.1%)

f(beta'x) at mean of independent vars = 0.403

Likelihood ratio test: Chi-square(7) = 1541.75 [0.0000]

7 ODDS-RATIO

==========

Variable Odds-ratio 95.0% conf. interval

==========

CreditScore 0.9993 [0.999, 1.000]

Age 1.0754 [1.070, 1.081]

Tenure 0.9842 [0.966, 1.002]

NumOfProducts 0.9094 [0.828, 0.998]

IsActiveMember 0.3410 [0.305, 0.382]

Germany 2.1119 [1.859, 2.399]

Female 1.6934 [1.522, 1.884]

LogBalance 1.0715 [1.043, 1.101]

==========

8 ALGORITHMS USED

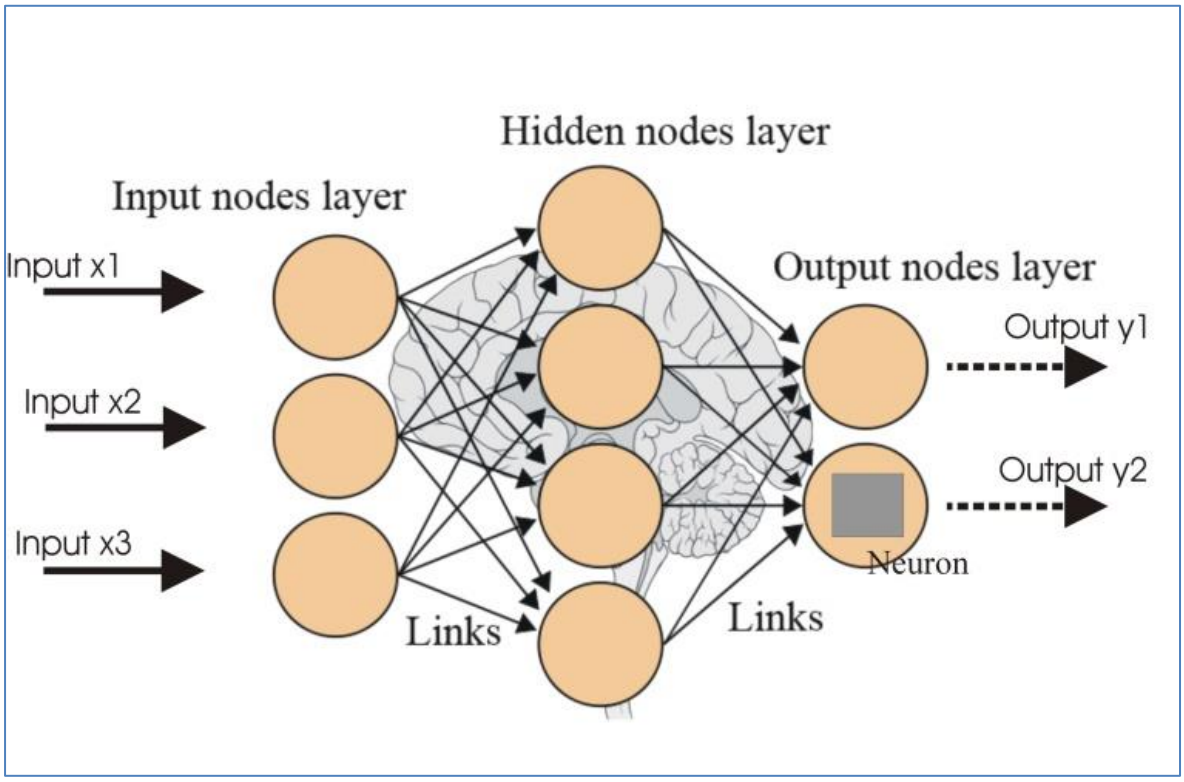
8.1 Artificial Neural Networks

Artificial neural networks are networks similar to human brain, used for mimicking the decision making process in human's heads. They have a similar structure to human brain, that is, having dendrites, which serve to get inputs from the raw data, like the brain gets from the outside world, process it using some pre-defined function, and give out an output. This is a basic perceptron, having only two layers – one input and one output. Wherein we have other more complex networks where there are many layers, and are called multi-layer perceptron and perform deep-learning. Such systems are self-learning and trained, rather than explicitly programmed, and perform well in areas wherein the solution/class is hard to be computed by traditional computer programs.

These neural networks work on the concept of "weights". Since not all inputs will have the same relevance or effect on the outcome, they are assigned a weight, or relevance number, depending on their influence on the decision to be made, just like our brains do. There's also a bias, so that in a case of all zero's we don't end up always with a zero. It's basically an adjustment. The aim of these models is to solve complex problems in the same way as the human brain, although they are more abstract. They are quite less complex in the number of connections than the human brain has and are closer to the computing power of nearly a worm. The outputs of neural networks are in binary digits)(0/1) and the value of the core and of the axon(inputs) are typically between 0 and 1. In order to train these artificial neural networks, a million cycles of iterations are needed to be done.

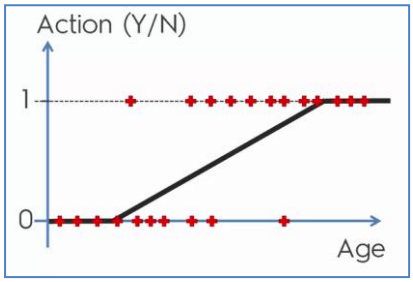
Neural networks have sets of multiple layers, and the signal (sum of inputs multiplied by corresponding weights) keeps traversing to and fro until it finds the best result. Back propagation is a technique used to adjust the relevance of the inputs on the output by resetting the weights on the succeeding neural units and this is often done in using training sets where the correct result is known. The results obtained are compared with the actual results and this is done with a large number of records until we get the most suitable weights of inputs.

8.1.1 DEMONSTRATION OF NEURAL NETWORKS



8.2 Logistic regression

When we are finding results from a data with a binary class label, we use logistic regression. Suppose we are trying finding out data with results as true and false. This can be done for the data set with classes like if the given customer will buy the product or not, will the given machine be able to manufacture the given product properly or not. In logistic regression we have independent variables and dependent variable. Independent variables are the attributes which are known about the data, and dependent variable/class is the one whose value is to be studies and predicted and has a 'logistic' value i.e. either a 1 meaning true/yes or a 0 meaning false/no. So the independent variable serves as the class of the dataset.

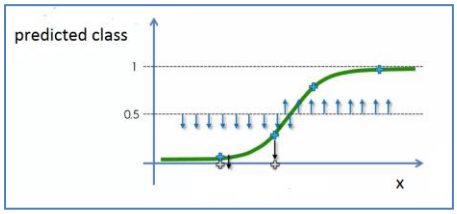


In logistic Regression, we calculate the probabilities, and plot them using a special mathematical graph called as a Sigmoid Function, an S-shaped graph, which is shown in the above picture, rounds of values(obtained by putting in regression formula) above 1 to 1, and below 0 to 0, since probabilities are always between 0 and 1.Above is the more suitable result as the results with the definite probabilities can be easily shown.

$Ln(^{p}/_{1-p}) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots\ldots + b_n x_n$

Where p = probability, b's are the coefficients and x's are the independent variables

An example - Now we are able to find out the probabilities of the two class labels for the different values of x. In this we are having 20 on the x- axis then the value for the same is .07 on the y- axis. This is same as saying that for a person with age 20 we will have .07 chances of him taking the action. This is almost negligible chance. Now might need to come up with more definite results which can definitely say if the given person will take the action or not like a person with the age of 20 with probability .07 will definitely not take the action same way the person with the age of 45 have probability of 78.8 which is almost equivalent to say that the person will take the action. In order to come up with this we will need a probability threshold such as .5 and any result above .5 is said to have taken action and any result with below .5 probabilities will not take the action.



8.3 K- Nearest Neighbour (KNN)

K-NN, or k-nearest neighbour, is a non-parametric classification technique, which classifies a new instance by finding out its distance to other instances whose class is known. It is a lazy, supervised learner. K-NN classifies objects of closest training examples in the feature space. K-NN is also known as :- I) instance based learning II) lazy learning III) case based reasoning IV) K- Nearest Neighbour V) example based reasoning VI) Instance based learning VII) memory based reasoning that means here a very little or no prior knowledge of the data distribution. The function is only to locally approximate and all other computation is deferred until classification. K-NN is very

simple classification and fundamental technique. K-NN is being used in many applications since 1970's, for example, in estimation of statistical data or pattern-recognition. K-NN can be divided into two types: I) without structure nearest neighbour technique – in which the entire data is divided into training and test sample data. From training data point to sample data point distance is calculated, and the point which has lowest distance between training point and sample point is called nearest neighbour. II) Structured nearest neighbour technique –this type of techniques are based on data structure for example ball-tree, axis – tree, OST, nearest future line and central line etc. Distance that will be used for calculating nearness can be – i) Manhattan Distance ii) Euclidean iii) Minskowaski Distance.

9 Cumulative Accuracy Profile (CAP) curve

Cumulative Accuracy profile (CAP) of a data analysis model is a curve which plots number of 'yes' predicted population in the x-axis and the total number of churners in the y-axis(in percentages), calculated beforehand. These customers in our case are the bank-churners. We then keep plotting these ratios on a graph at each record. The

resulting curve is called CAP curve. This validates the performance of our model in a way that the predicted values are plotted against the real values and thus the line is obtained. Processing the data for plotting this curve is as follows:-

1. Obtained result is arranged in ascending order. Thus all the 1's (predicted) appear before 0's and we take we take their corresponding actual values, sum them up, and take their ratio by the total number of customers. This gives the per cent of customers who were rightly predicted by our model.
2. We seldom take(and have taken in our results) an 'average curve', also called as a default curve, which keeps taking the taking X% of churners at each stage and plotting it on the graph. This X here is the percentage of total churners that we will be calculated beforehand.
3. We can also take a perfect curve, called as a crystal curve to compare the three curves.
4. The thumb rule is, the obtained curve should be better than the average curve, and can be less good than the perfect curve. The near it is to the crystal curve, the better our model is.
5. This can be quantified by taking the are under the obtained curve and the crytal curve,

Or if we are comparing two or more models, we can simply draw a vertical line at the 50% mark, which is what we have done to compare the three algo's.

10 CODES

10.1 CODE FOR ARTIFICIAL NEURAL NETWORKS

```
library(nnet)

churntrain<-read.csv(file="churn train.csv",head=TRUE)

churntrain<-sample(1:11000,10000)

churntrain<-read.csv(file="churn test.csv",head=TRUE)

show(churntest)

class<- churntrain$Exited
```

```
class.ind(churn$CreditScore)

ideal <- class.ind(seeds$Class)

show(class)

help(nnet)

show( class[churntrain,])

churnANN = nnet(churntrain[-11], class, size=5, entropy=TRUE)

show(churnANN)

predict(churnANN, churntest, type="class")

table(predict(seedsANN, churntest[-11], type="class"),churntest$Exited)

//

library(nnet)

churntrain<-read.csv(file="churn train.csv",head=TRUE)

churntest<-read.csv(file="churn test.csv",head=TRUE)

show(churntest)

class<- churntrain$Exited

class.ind(churn$CreditScore)

ideal <- class.ind(seeds$Class)

show(class)

help(nnet)

show( class[churntrain,])

churnANN = nnet(churntrain, class, size=1, entropy=TRUE)

show(churnANN)
```

```
show(predict(churnANN, churntest[-11], type="class"))



table(predict(churnANN, churntest, type="class"),churntest)

//

churntrain<-read.csv("Churn-Modelling.csv",head=TRUE)

churntest<-read.csv("Churn-Modelling-Test-Data.csv",head=TRUE)

maxs <- apply(churntrain[1:10], 2, max)

mins <- apply(churntrain[1:10], 2, min)

scaled.data <- as.data.frame(scale(churntrain[1:10],center = mins, scale = maxs - mins))

print(head(scaled.data,2))

maxs2 <- apply(churntrain[1:11], 2, max)

mins2 <- apply(churntrain[1:11], 2, min)

scaled.data2 <- as.data.frame(scale(churntrain[1:11],center = mins2, scale = maxs2 - mins2))

print(head(scaled.data2,2))

maxs1 <- apply(churntest[1:10], 2, max)

mins1 <- apply(churntest[1:10], 2, min)

scaled.datatest <- as.data.frame(scale(churntest[1:10],center = mins1, scale = maxs1 - mins1))

print(head(scaled.datatest,2))

feats<-names(scaled.data)

f5 <- paste(feats,collapse=' + ')
```

```
f5 <- paste('Exited ~',f5)

show(f5)

f5 <- as.formula(f5)

library(neuralnet)

nn <- neuralnet(f5,scaled.data2,hidden=3,linear.output=FALSE)

predicted.nn.values <- compute(nn,scaled.datatest,type="raw")

print(predicted.nn.values$net.result)
```

10.2 CODE FOR KNN

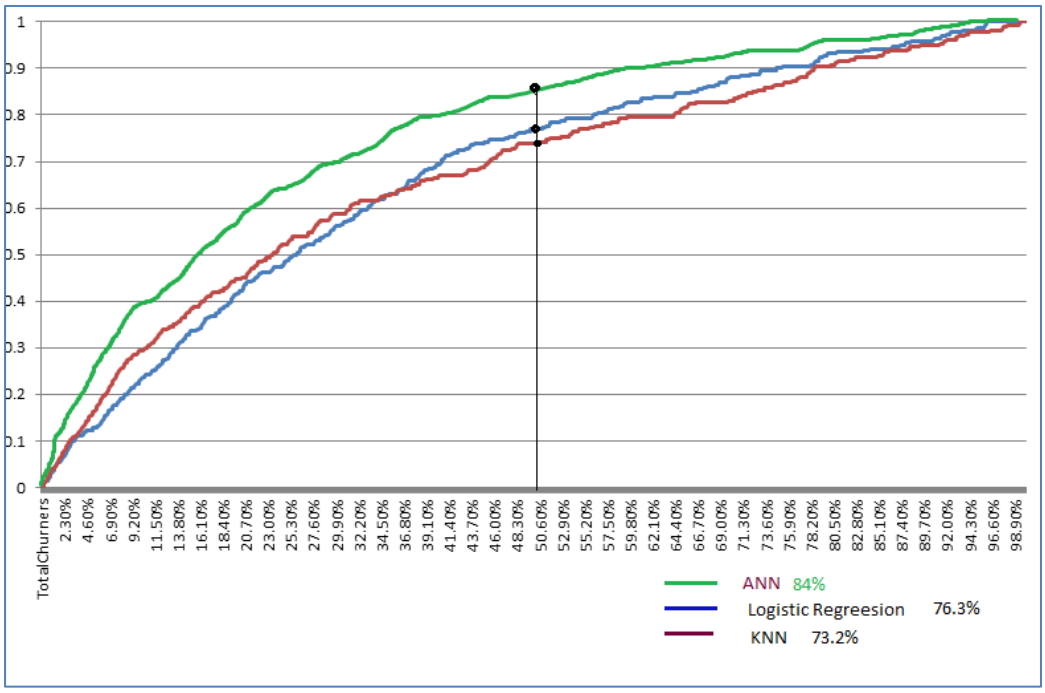
```
//importing churn data having 10k records into sas data file named churndata here, getnames=yes here means data set has field names contained, dbms=csv is used to

give type of file

PROC IMPORT DATAFILE = "C:\Users\User\Desktop\geodemographic model\Churn-Modelling-Copy.csv"

OUT = churndata

DBMS = csv

REPLACE;

GETNAMES=YES;

RUN;


//importing churn test data set having 1000 records into sas file named churntestdata

PROC IMPORT DATAFILE = "C:\Users\User\Desktop\geodemographic model\Churn-Modelling-Test-Data-Copy.csv"

OUT = churntestdata
```

```
DBMS = csv

REPLACE;

GETNAMES=YES;

RUN;


//running classification method KNN with k=5, using SAS PROC discrim here, testout is sas data file for storing output on test data which later will be converted to

excel file . class exited means the class to be classified here is the field 'exited' in our data. the line var bla --bla means data from bla to bla will be analysed

to make the knn model

proc discrim data=churndata

method=npar k=5

testdata=churntestdata

TESTOUT=result2

;

class exited;

var Creditscore -- EstimatedSalary;

run;


//proc for printing results in SAS

PROC PRINT DATA = result2;

OPTIONS;

RUN;

PROC EXPORT DATA= WORK.RESULT OUTFILE=
"C:\Users\User\Desktop\geodemographic model\knn_result.xls"

DBMS=EXCEL2000 REPLACE;
```

```
SHEET="knn";

RUN;
```

11 RESULTS

The results obtained shows that neural networks perform the best among the three with accuracy of 84.0%. Logistic Regression being the second best performer with an accuracy of 76.3%, followed by KNN with 73.2%.



12 FUTURE SCOPES

This project can have a variety of future aspects on which it can work with few improvisation. The biggest scope is the fact that this model in being applied on the churn data of the bank. The same can be applied on the retail industry or the airline industry

Moreover we can increase the dimensionality of this project by increasing the number of algorithms in this project. We have done this project with k- nearest neighbor, logistic algorithm and artificial neural networks. We can increase the scope of this project by using the other algorithms such as SVM, or the binary tree etc.

13 REFERENCES

1. S B Imandoust et al. Int. Journal of Engineering Research and Applications www.ijera.com Vol. 3, Issue 5, Sep-Oct 2013, pp.605-610

2. Patrick Verlinde G´erard Chollet CNRS URA-820 Ecole Nationale Sup´erieure de T´el´ecommunications/TSI Department F75634 Paris, France

3. Koh, Hian Chye, Wei Chin Tan, & Chwee Peng Goh. " A Two-step Method to Construct Credit Scoring Models with Data Mining Techniques." *International Journal of Business and Information* [Online], 1.1 (2006): n. pag. Web. 8 Mar. 2017

4. Keramati, A., Ghaneei, H. & Mirmohammadi, S.M. Financ Innov (2016) 2: 10. doi:10.1186/s40854-016-0029-6

5. International Journal of Computer Applications (0975 – 8887) Volume 27– No.11, August 2011