www.ijiarec.com

ISSN:2348-2079

Volume-9 Issue-2

International Journal of Intellectual Advancements and Research in Engineering Computations

Data analysis of consumer complaints in banking industry using k-mean clustering and sentiment analysis

[1]P.V.Nandhini, [2]S.Narmadha, [3]Y.Priyanka,
[4]Mr.T.Viswanathkani M.E.,
UG Scholar's, Department of Computer Science and Engineering, Vivekanandha College of Engineering For Women [Autonomous], Tiruchengode - 637 205, Tamilnadu, India.[1,2,3]

Assistant Professor, Department. of Computer Science and Engineering, Vivekanandha College of Engineering For Women [Autonomous], Tiruchengode - 637 205, Tamilnadu, India[4]

ABSTRACT

A consumer's complaints present bank or reporting agency with an opportunity to identify and rectify specific problems with their current product or service. The banks that are receiving customer complaints filed against them will analyze the complaint data to provide results on where the most complaints are being filed, what products/ services are producing the most complaints and other useful data. This project assists banks in identifying the location and types of errors for resolution, leading to increased customer satisfaction to drive revenue and profitability. This project finds a correlation between complaints, companies and consumers to refine company applications to better accommodate consumer needs using k-means clustering. In addition, using SVM classification, the complaints sentiment values are analyzed and classified into positive or negative reviews. The project is designed using R Studio 1.0 as front end. The coding language used is R version 3.4.4.

Keywords: Consumer compliant, Support Vector Machine, Location identifications.

INTRODUCTION

As we are aware that in today's modern era people are more into business, so receiving a complaint from a consumer happens almost every day. A consumer's complaints present bank or reporting agency with an opportunity to identify and rectify specific problems with their current product or service. Service complaints management is a critical part of business management.

A good complaint-management strategy will result in best customer relationship outcome with minimal human-resource investment and so hope to find a correlation between complaints, companies, and consumers to refine company applications to better accommodate consumer needs. Increasingly companies are recognizing the value of a customer complaint in that it is feedback on their experience, and an opportunity to not only resolve a problem for that particular customer but perhaps also for a much larger number of customers and that leads to inevitable amounts of data that has to be analyzed and specific functions are used to aggregate the analysis results. Clustering is regarded as a crucial unsupervised learning problem, that tries to search for similar structures among an unlabeled data set .These similar structure are data sets, usually referred to as clusters. the information within every cluster is comparable (or close) to components within its cluster, and is dissimilar to (or additional from) parts that belong to alternative clusters.

The mining techniques' goal is to detect the intrinsic grouping of a data set. In hierarchical clustering, a treelike cluster structure (dendrogram) is created through recursive partitioning (divisive methods) or combining (agglomerative) of existing clusters, whereas in k-means clustering divides a cluster of k points with reference to a centroid,

Author for correspondence:
Department of Computer Science and Engineering, Vivekanandha College of Engineering For Women [Autonomous], Tiruchengode - 637 205, Tamilnadu, India.

which helps if we are aware of the data points that are probable and output relevant. We hope to find a correlation between complaints, companies and consumers to refine company applications to better accommodate consumer needs using k-means clustering.

Customer behavior and complaints is an important issue that needs to be addressed and resolved in both public and private sectors of service providers. Customer complaints can be used as an accurate measure of how successful a service is, especially with the transformation of information technology into concepts and ideas. The competitive advantage of some companies is not theservice offered but rather the attention to customer complaints and resolution. In fact, there is the risk to the company when the client is silent and not complaining. When the client is faced with the problem, usually he/she has two options to go with: either grumble from the company and disconnect from them permanently or complain. Therefore, it is very important to rely on a mechanism that helps the Customer Relationship Management (CRM) division in each company to analyze and handle complaints. In this paper, we analyze customer complaints dataset from a public service provider, the Metropolitan Transportation Authority (MTA) public transportation service provider, which provides services such as subway, bus and rail.

Literaturesurvey

Jain and Jain (2006)Customer Satisfaction in Retail Banking Services NICE, J. Bus. Stud., 1(2):95-102demonstrated that the banking sector, both private and public have suffered radical as well as revolutionary changed due to the liberalization act of 1991. Retail banking is the consumer preferred choice which articulates itself responses received from 200 customers of HDFC bank, ICICI bank and some other banks in the city of Varanasi, Uttar Pradesh and he looked upon the schemes offered by the banks, quantized satisfaction in different types of services, expectations about these schemes and the height of segmentation among the services offered.

Singh (2006)Customer Management in Banks Vinimaya, 37(3): 31-35discusses CRM approaches in various banks. He emphasized on how the management targets customers in order to gain insight and gives out value added services and products. Web as provided a smooth user experience, giving access to the various features used by the customers thereby achieving customer satisfaction. Management has to strive to ensure end to end delivery and ensure customer satisfaction which is essential to the banks in terms of maintaining high regards and loyalty obtained from customers.

Kamakodi (2007) Customer Preferences on e-Banking Services- Understanding through a Sample Survey of Customers of Present Day Banks in India Contributors, Bank net Publications, 4: 30-43concluded that modern day generation is influenced by the computation features used by banks and so the banks study about factors influencing their preferences. Residence relocation, salary fluctuation and unavailability banking based services are reasons enough to change bank.

Uppal and Kaur(2007)Customer Service in Banks- An Empirical Study', Bankers Conference Proceedings, pp. 36-42 determined how consumer's awareness of web domains used by banks and gave some measures to make these applications more successful. They concluded that the limitation about today's web domain application is spreading the awareness about the varied features offered.

Mishra and Jain (2007) Constituent Dimensions of Customer Satisfaction: A Study of Nationalized and Private Banks Prajnan, 35(4)390-398took up dimensions of consumer satisfaction in national and private banks. The study talks about how satisfaction is the foremost asset to the organization, which provides unmatched competitive edge that helps achieving loyalty of a customer. They also spoke how high level of customer satisfaction leads to loyalty. The study observed ten factors and five areas of satisfaction for both national and private sector bank.

Chetna Sethi and Garima Mishra(2013)"A Linear PCA based hybrid K-Means PSO algorithm for clustering large dataset," International Journal of Scientific & Engineering Research, Volume 4, Issue 6, June-2013, pp.1559-1566.Proposed a Linear PCA based hybrid K-Means clustering and PSO algorithm (PCA-K-PSO). In (PCA-K-PSO) algorithm the fast convergence of K-Means algorithm and the global searching ability of Particle Swarm Optimization (PSO) are combined for clustering large data sets using Linear PCA. Better clustering results can be obtained with PCA-K-PSO as compared to ordinary PSO. This was effectively developed in order to make its use for efficient clustering of high- dimensional data sets.

N. Bouhmala, A. Viken, J. B. Lonnum, (2015)"Enhanced Genetic Algorithm with K-Means for the Clustering Problem", International Journal of Modelingand Optimization, pp. 150-154, 2015.Combined Genetic Algorithm and K- Means to improve the quality of clusters formed and speed up their search process. The performance of GAKM is tested over the datasets such as iris, glass, etc., and that has been taken from Machine learning repository. The experimental results have proved that GAKM converges faster while comparing to standard Genetic Algorithm. Though this algorithm failed to capture the best quality of clusters, it is unsuitable for the maximizing both homogeneity and heterogeneity within same clusters and with different clusters respectively.

These datasets fall under the complaints of Credit reporting, Mortgage, Debt Collection, Consumer Loan and Banking Accounting. By using data mining techniques, cluster analysis as well as predictive modeling is applied to obtain valuable information about complaints in certain regions of the Country.

Arin Brahma,David M. Goldberg, Nohel Zaman,MarianoAloiso(2021)Automated mortgage origination delay detection from textual conversations January 2021For modern mortgage firms, the process of setting up and verifying a new loan, known as origination, is complex and multifaceted. The literature notes that this process is rife with delays that can stunt the firm’s business opportunities, but no modern analytical techniques have been developed to address the problem. In this paper, we suggest the use of text analytic and machine learning techniques to predict likely delays. In collaboration with a large national mortgage firm, we derive a large dataset of transcripts from employees’ communications pertaining to potential loans.

Existing System

In existing system, K-Means clustering of bank customer data is used to analyze and group the customers. However, from their reviews/complaints the seriousness of the complaint type could not be identified. The sentiment values could not be found out to measure the exactness of the complaints. Only the customers can be grouped into three, four and more groups based on their similarity among the complaint records.

Disadvantages

- The existing system does not give the estimated sentiment prediction of the subject based on the text reviews/complaints sent by the customers.
- Sentiment analysis is not carried out so that the review is judged as either positive or negative.
- Percentage of positive/negative reviews could not be found out.
- Words are not given with exact sentiment numerical values and so classification such as positive or negative is not accurate.

Proposed System

In proposed system, all the existing methodology is carried out. The proposed approach takes input from the data set created by accumulating all the text messages send by the customers. All the messages may be from different banking reviews like loan processing delay, feedbacks etc. It also gives a highly efficient method of finding the sentiment of the customer by analyzing the text reviews and also processing emoticons. Using sentiment values for all words, the overall sentiment value for the paragraphs are calculated and so the polarity among the reviews is found out.

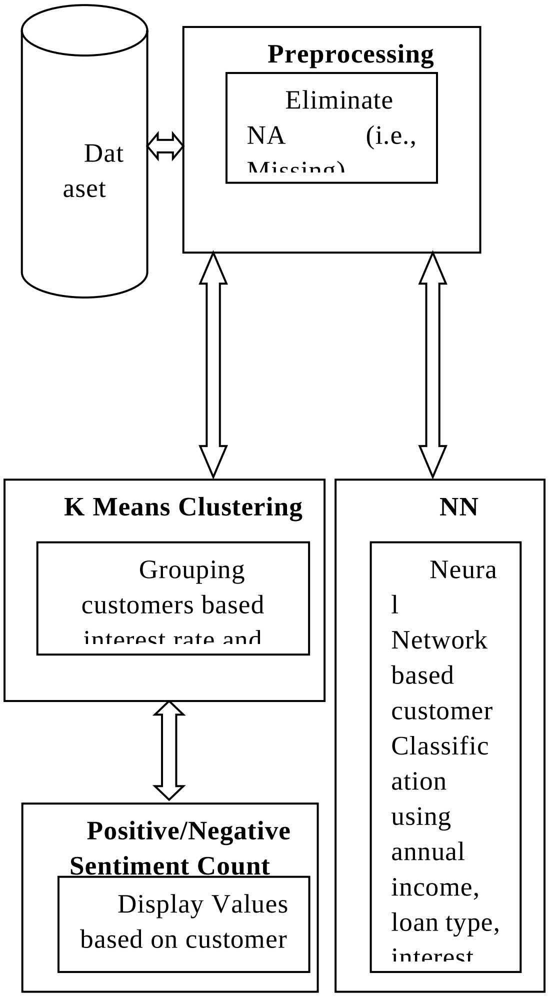
This project work focus on analyzing and predicting if the given loan will be fully repaid or not using Artificial Neural Network. This data set fall under the complaints of Annual income, Interest rate, Loan status, Duration term, Verification status etc.,.

Prediction is made for better accuracy and for the practical users. In this paper prediction is made for if the loan will be fully repaid or not, and also analyzing what kind of situation does a customer is unable to pay loan. This prediction will lead to know more about the customer whether the loan can be given to them or not.

Advantages

- The proposed system gives the estimated sentiment prediction of the subject based on the text reviews/complaints sent by the customers.
- Sentiment analysis is carried out so that the review is judged as either positive or negative.
- Percentage of positive/negative reviews can be found out.
- Various words are given with exact sentiment numerical values and so classification such as positive or negative is accurate.
- Neural network helps to classify the given loan request details into one of the predefined applied loans.

System Model


Dat
aset
Preprocessing
Eliminate NA (i.e., Missing)
K Means Clustering
Grouping customers based interest rate and
NN
Neural Network based customer Classification using annual income, loan type, interest
Positive/Negative Sentiment Count
Display Values based on customer

Methodology

Modules

- **K-MEANS CLUSTERING**
- **POSITIVE/NEGATIVE SENTIMENT COUNTS**
- **NEURAL NETWORK BASED CLASSIFICATION**
- **SIMILARITY PERCENT FINDING**

K-Means Clustering

In this module, K-Means clustering of bank customer data are used to analyze and group the customers. The customer data set is taken with CustomerId, AnnualIncome, LoanType, InterestRate, DurationinMonths, LoanAmountRequested and LoanStatus columns of which InterestRateand DurationinMonths are taken as X and Y Axis for KMeans clustering. By default 3 is given for K, but we can give any number to cluster the data.

Positve/Negative Sentiment Counts

In this module, sentiment values for various words are given in the tabular data text file and fetched in a data frame. Then the customer review data set is taken and filled in a data frame. For each record, the sentiment words presence is found out and the values are summated both for positive words as well as negative words. Then overall sum for both positive and negative counts are prepared and displayed. This will assist in overall customer feedback analysis.

Neural Network Based Classification

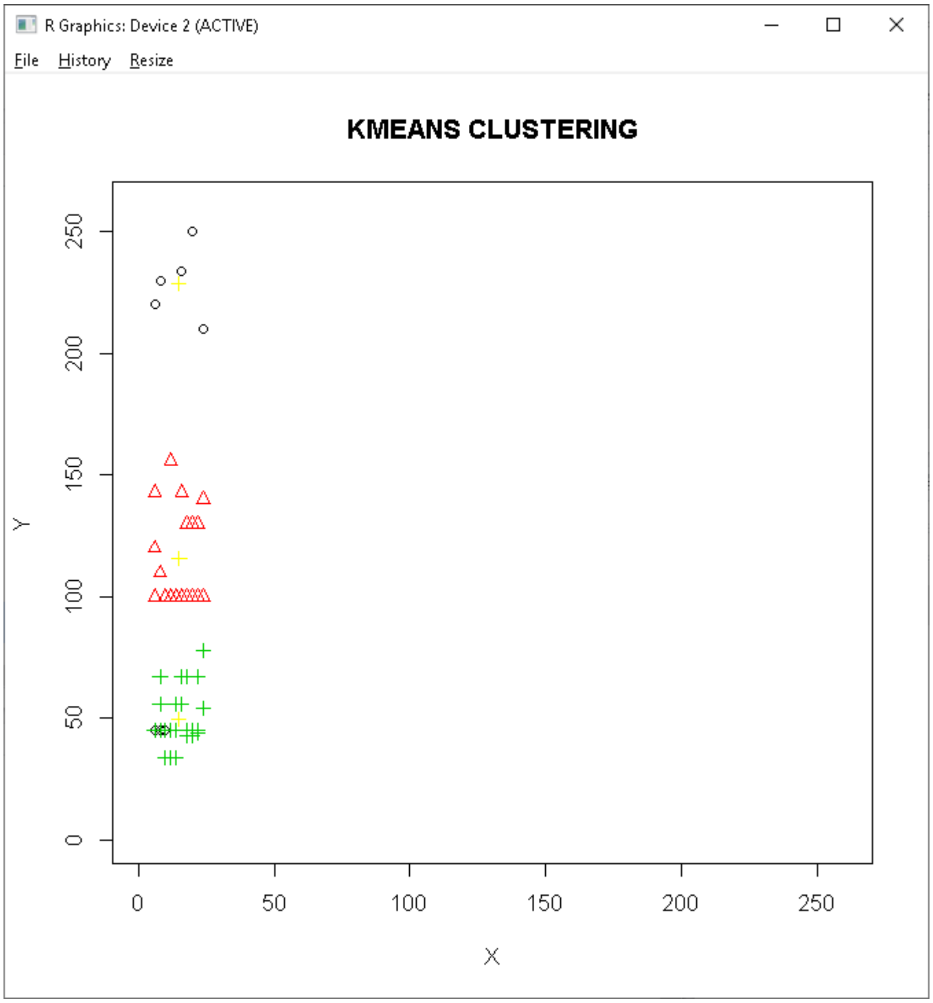
In this module, the load data set is taken which contains CustomerId, AnnualIncome, LoanType, InterestRate, DurationinMonths, LoanAmountRequested and LoanStatus values. All the columns are factorized i.e., unique values are found out and filled in a vector. Then hot encoding vector is found out for those record values. This is carried out for all columns and is concatenated for each record vector. These hot encoding strings are added as a separate column to the data frame and used as input to neural network model.

The input neuron count becomes the length of the data set records. The hidden layer neurons are set to 500 and output layer neuron to 2 [1 for approved, 2 for rejected]. Then weight and bias finding work is carried out.

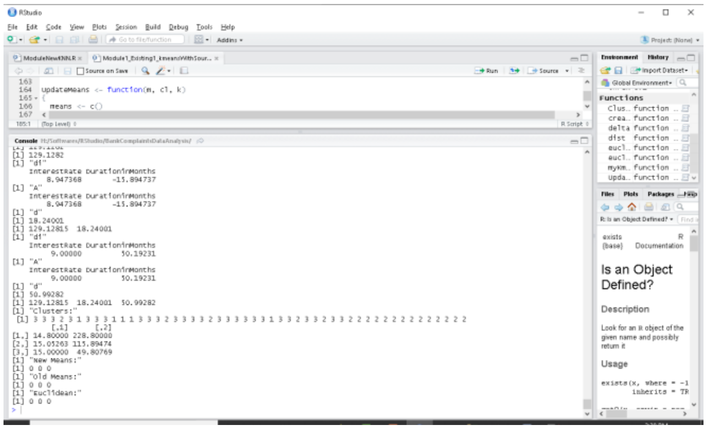
Similarity Percent Finding

To find the nearest matching record values among the given training record set, hot encoding vector comparison is also made so that similarity percent of the test record in the overall training dataset records are also found out. The test record is classified into maximum similarity matching record.

EXPERIMENTAL RESULTS



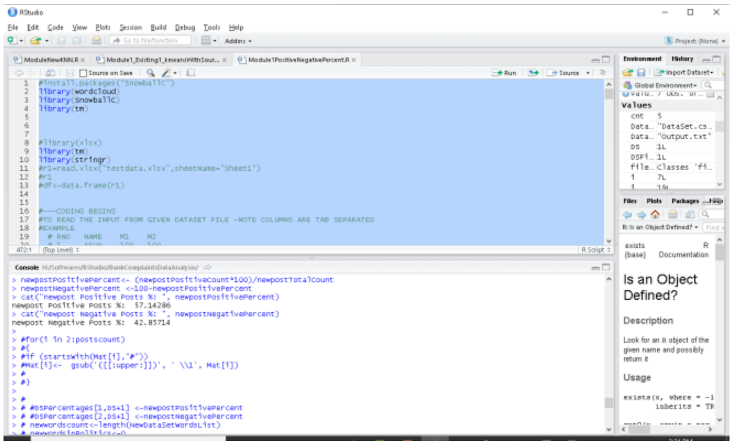
Records Grouped In Clusters


RStudio
File Edit Code View Plots Session Build Debug Tools Help
ModuleNewKNN.R
Module1_Existing1_kmeansWithSour...
Source on Save
Run
Source
Console H:/Softwares/RStudio/BankComplaintsDataAnalysis/
Functions
Is an Object Defined?
Description
Usage

```
updateMeans <- function(m, cl, k)
{
  means <- c()
```

```
[1] 129.1282
[1] "d1"
   InterestRate DurationinMonths
       8.947368       -15.894737
[1] "A"
   InterestRate DurationinMonths
       8.947368       -15.894737
[1] "d"
[1] 18.24001
[1] 129.12815  18.24001
[1] "d1"
   InterestRate DurationinMonths
        9.00000         50.19231
[1] "A"
   InterestRate DurationinMonths
        9.00000         50.19231
[1] "d"
[1] 50.99282
[1] 129.12815  18.24001  50.99282
[1] "Clusters:"
 [1] 3 3 3 2 3 1 3 3 3 1 1 1 3 3 3 2 3 3 3 3 2 3 3 3 3 3 3 1 3 3 2 3 3 2 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
         [,1]      [,2]
[1,] 14.80000 228.80000
[2,] 15.05263 115.89474
[3,] 15.00000  49.80769
[1] "New Means:"
[1] 0 0 0
[1] "Old Means:"
[1] 0 0 0
[1] "Euclidean:"
[1] 0 0 0
>
```

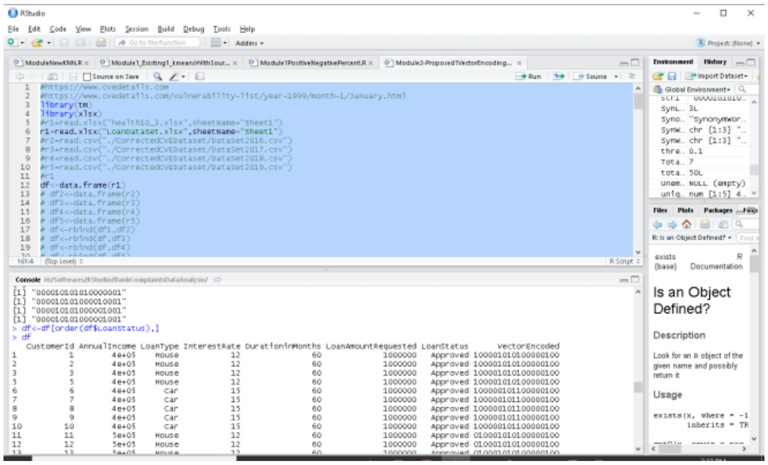
Positive/Negative Posts Percentage


RStudio
File Edit Code View Plots Session Build Debug Tools Help
ModuleNewKNN.R
Module1_Existing1_kmeansWithSour...
Module1PositiveNegativePercent.R
Source on Save
Run
Source
Console H:/Softwares/RStudio/BankComplaintsDataAnalysis/
Values
Is an Object Defined?
Description
Usage

```
#install.packages("Snowballc")
library(wordcloud)
library(Snowballc)
library(tm)



#library(xlsx)
library(tm)
library(stringr)
#r1=read.xlsx("testdata.xlsx",sheetName="Sheet1")
#r1
#df<-data.frame(r1)


#---CODING BEGINS
#TO READ THE INPUT FROM GIVEN DATASET FILE -NOTE COLUMNS ARE TAB SEPARATED
#EXAMPLE
  # RNO    NAME     M1     M2
```

```
> newpostPositivePercent<- (newpostPositiveCount*100)/newpostTotalCount
> newpostNegativePercent <-100-newpostPositivePercent
> cat("newpost Positive Posts %: ", newpostPositivePercent)
newpost Positive Posts %:  57.14286
> cat("newpost Negative Posts %: ", newpostNegativePercent)
newpost Negative Posts %:  42.85714
>
> #for(i in 2:postscount)
> #{
> #if (startsWith(Mat[i],"#"))
> #Mat[i]<-  gsub('([[:upper:]])', ' \\1', Mat[i])
> #
> #}
>
> #
> # #DSPercentages[1,DS+1] <-newpostPositivePercent
> # #DSPercentages[2,DS+1] <-newpostNegativePercent
> # newwordscount<-length(NewDataSetWordsList)
```

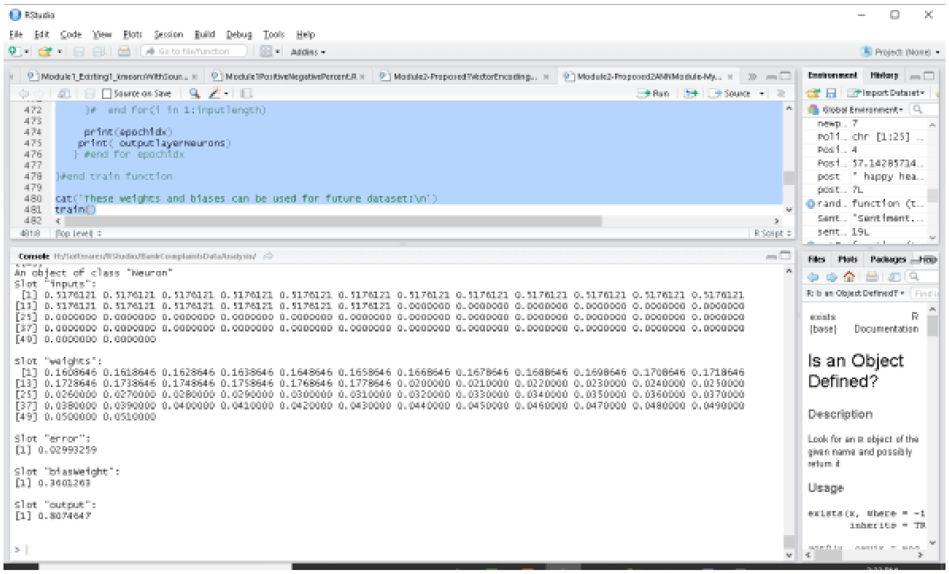
Vector Encoded Values


RStudio
File Edit Code View Plots Session Build Debug Tools Help
ModuleNewKNN.R
Module1_Existing1_kmeansWithSour...
Module1PositiveNegativePercent.R
Module2-ProposedTVectorEncoding...
Source on Save
Run
Source
Console H:/Softwares/RStudio/BankComplaintsDataAnalysis/
Is an Object Defined?
Description
Usage

```
#https://www.cvedetails.com
#https://www.cvedetails.com/vulnerability-list/year-1999/month-1/January.html
library(tm)
library(xlsx)
#r1=read.xlsx("health10_3.xlsx",sheetName="Sheet1")
r1=read.xlsx("LoanDataset.xlsx",sheetName="Sheet1")
#r2=read.csv("./CorrectedCVEDataset/DataSet2016.csv")
#r3=read.csv("./CorrectedCVEDataset/DataSet2017.csv")
#r4=read.csv("./CorrectedCVEDataset/DataSet2018.csv")
#r5=read.csv("./CorrectedCVEDataset/DataSet2019.csv")
#r1
df<-data.frame(r1)
# df2<-data.frame(r2)
# df3<-data.frame(r3)
# df4<-data.frame(r4)
# df5<-data.frame(r5)
# df<-rbind(df1,df2)
# df<-rbind(df,df3)
# df<-rbind(df,df4)
```

```
[1] "00001010101010000001"
[1] "00001010100000010001"
[1] "00001010100000001001"
[1] "00001010100000001001"
> df<-df[order(df$LoanStatus),]
> df
   CustomerId AnnualIncome LoanType InterestRate DurationinMonths LoanAmountRequested LoanStatus       VectorEncoded
1           1        4e+05    House           12               60             1000000   Approved 10000101010000000100
2           2        4e+05    House           12               60             1000000   Approved 10000101010000000100
3           3        4e+05    House           12               60             1000000   Approved 10000101010000000100
5           5        4e+05    House           12               60             1000000   Approved 10000101010000000100
6           6        4e+05      Car           15               60             1000000   Approved 10000010110000000100
7           7        4e+05      Car           15               60             1000000   Approved 10000010110000000100
8           8        4e+05      Car           15               60             1000000   Approved 10000010110000000100
9           9        4e+05      Car           15               60             1000000   Approved 10000010110000000100
10         10        4e+05      Car           15               60             1000000   Approved 10000010110000000100
11         11        5e+05    House           12               60             1000000   Approved 01000101010000000100
12         12        5e+05    House           12               60             1000000   Approved 01000101010000000100
```

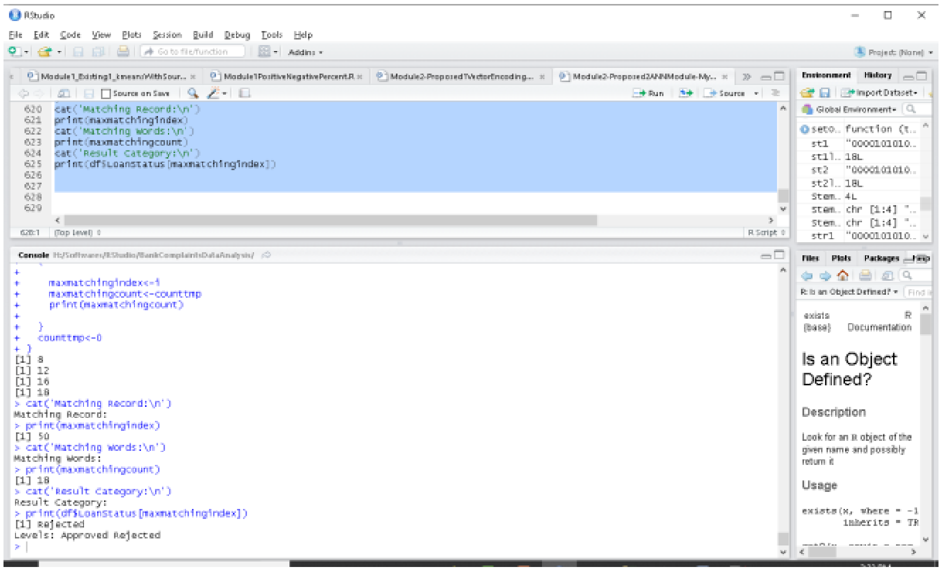
Neural Network Model [Weights And Biases]



```
    }#  end for(i in 1:inputlength)

     print(epochidx)
    print( outputlayerNeurons)
   } #end for epochidx

}#end train function

cat('These weights and biases can be used for future dataset:\n')
train()
```

```
An object of class "Neuron"
Slot "inputs":
 [1] 0.5176121 0.5176121 0.5176121 0.5176121 0.5176121 0.5176121 0.5176121 0.5176121 0.5176121 0.5176121 0.5176121 0.5176121
[13] 0.5176121 0.5176121 0.5176121 0.5176121 0.5176121 0.5176121 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
[25] 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
[37] 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
[49] 0.0000000 0.0000000

Slot "weights":
 [1] 0.1608646 0.1618646 0.1628646 0.1638646 0.1648646 0.1658646 0.1668646 0.1678646 0.1688646 0.1698646 0.1708646 0.1718646
[13] 0.1728646 0.1738646 0.1748646 0.1758646 0.1768646 0.1778646 0.0200000 0.0210000 0.0220000 0.0230000 0.0240000 0.0250000
[25] 0.0260000 0.0270000 0.0280000 0.0290000 0.0300000 0.0310000 0.0320000 0.0330000 0.0340000 0.0350000 0.0360000 0.0370000
[37] 0.0380000 0.0390000 0.0400000 0.0410000 0.0420000 0.0430000 0.0440000 0.0450000 0.0460000 0.0470000 0.0480000 0.0490000
[49] 0.0500000 0.0510000

Slot "error":
[1] 0.02993259

Slot "biasweight":
[1] 0.3601263

Slot "output":
[1] 0.8074647

> |
```

Matching Record In The Training Data For The Given Test Data



```
cat('Matching Record:\n')
print(maxmatchingindex)
cat('Matching words:\n')
print(maxmatchingcount)
cat('Result category:\n')
print(df$Loanstatus[maxmatchingindex])
```

```
+
+     maxmatchingindex<-i
+     maxmatchingcount<-counttmp
+     print(maxmatchingcount)
+
+   }
+   counttmp<-0
+ }
[1] 8
[1] 12
[1] 16
[1] 18
> cat('Matching Record:\n')
Matching Record:
> print(maxmatchingindex)
[1] 50
> cat('Matching words:\n')
Matching words:
> print(maxmatchingcount)
[1] 18
> cat('Result category:\n')
Result category:
> print(df$LoanStatus[maxmatchingindex])
[1] Rejected
Levels: Approved Rejected
> |
```

CONCLUSION

The project has explored various concepts of data mining of bank customers' data set and the results show what problems customers are having with specific problems in particular loan provisions. This valuable information will show where companies will need to invest in to improve their overall performance in the view of their customers. This will lead to improved customer satisfaction. By maximizing customer satisfaction, the opportunity for repeat sales to customers can be increased. Customer satisfaction also helps to increase customer loyalty, reducing the need to allocate marketing budget to acquire new customers. Satisfied customers may also recommend your products or services to other potential customers, increasing the potential for additional revenue and profit.

REFERENCES

[1]. Goyal S, Thakur KS (2008). A Study of Customer Satisfaction Public and Private Sector Banks of India Punjab, J. Bus. Stud., 3(2): 121- 127.
[2]. Uppal RK (2007). Customer Service in Banks- An Empirical Study', Bankers Conference Proceedings, pp. 36-42.
[3]. Kamakodi N (2007). Customer Preferences on e-Banking ServicesUnderstanding through a Sample Survey of Customers of Present Day Banks in India Contributors, Banknet Publications, 4: 30-43.
[4]. Mishra JK, Jain M (2007). Constituent Dimensions of Customer Satisfaction: A Study of Nationalized and Private Banks Prajnan, 35(4): 390-398.
[5]. Jain AK, Jain P (2006). Customer Satisfaction in Retail Banking Services NICE, J. Bus. Stud., 1(2):95-102.
[6]. Singh SB (2006). Customer Management in Banks Vinimaya, 37(3): 31- 35.
[7]. Bhaskar PV (2004). Customer Service in Banks IBA Bulletin, 36(8): 9- 13.
[8]. Hasanbanu S (2004). Customer Service in Rural Banks: An Analytical Study of Attitude of Different types of Customers towards Banking Services IBA Bulletin, 36(8): 21-25.
[9]. Singh S (2004). An Appraisal of Customer Service of Public Sector Banks IBA Bulletin, 36(8): 30-33.
[10]. Shankar AG (2004). Customer Service in Banks IBA Bulletin, 36(8): 5- 7.
[11]. Ganesh C, Varghese ME (2003). Customer Service in Banks: An Empirical Study'.Vinimaya, 36(2): 14-26.