RESEARCH ARTICLE OPEN ACCESS

Data Analysis of Consumer Complaints in Banking Industry using k means clustering and sentiment analysis

S.Jagadeesan M.Sc(CS)., MCA., M.Phil(CS)., ME(CSE).,*, P. Shanmugapriya**
* Assistant Professor, Department of MCA, Nandha Engineering College (Autonomous), Erode, Tamilnadu, India.
Email: jagadeesan12398@gmail.com
** Final MCA, Department of MCA, Nandha Engineering College (Autonomous), Erode, Tamilnadu, India.
Email: priyasmp999@gmail.com

****

ABSTRACT
This paper specialise in exploring and analyzingConsumer Finance Complaints data, to seek out what percentage similar complaints are there in reference to an equivalent bank or service or product. These datasets fall into the complaints of Credit reporting, Mortgage, Debt Collection, personal loan and Banking Accounting. By using data processing techniques, cluster analysis also as predictive modeling is applied to get valuable information about complaints in certain regions of the Country. The banks that are receiving customer complaints filed against them will analyse the complaint data to supply results on where the foremost complaints are being filed, what products/ services are producing the foremost complaints and other useful data. Our model will assist banks in identifying the situation and kinds of errors for resolution, resulting in increased customer satisfaction to drive revenue and profitability.

Key Terms — **Consumer, Complaint, analysis, clustering, predictive.**

****

I. 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 may be a critical a 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 way 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, 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 a hybrid approach of hierarchal and k-means clustering.

II. LITERATURE REVIEW

The number of studies has been conducted regarding the services to customers and their awareness. As such, we have reviewed some of them.

Kamakodi (2007) concluded 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) 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) took 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.

Jain and Jain (2006) demonstrated 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) discusses 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.

Bhaskar (2004) computed that expansion of banking is directly proportional to the quality of services provided by the banks and satisfaction is regarded highly as customers feedback is the only thing to lean on, when it comes to the highly competitive banking industry. Arguably, India's banking industry is highly thriving and depends heavily on customer morale and loyalty.

Furthermore, Hasanbanu (2004) stated how the rural India is unaware about various schemes and benefits offered by the banks in order to ensure financial welfare. The majority of rural population is inaccessible to the web domain services of the banks and continue to prefer local moneylenders charging ligh interest rates. The study was conclusive and based on the data provided by the RBI, however, it is based on the questionnaire and surveys. Although Singh (2004) spurred about the reality of banks in terms of providing customer support and found out that the customers are influenced by the banks location and the minutest detail of the banking details including the banking interest rates as well as attitudes and customer support provided by the personnel.

III. METHODOLOGY AND PROCEDURE

A. Hierarchical Clustering

Probably the foremost applied method in economy is agglomerative hierarchical cluster analysis. It is based on a proximity matrix which includes the similarity evaluation for all pairs of objects. It means that various similarity or dissimilarity measures for different types of variables (quantitative, qualitative and binary)can be used. Moreover, different approaches for evaluation of the cluster similarity (single linkage, complete linkage, average linkage, Ward’s method, etc.) can also be applied.

Given: A set X of objects{x1,x2,. xn}

A distance function dist(c1,c2)

for i=1 to n
ci = {xi}
end forC={c1,...,cn}
l=n+1

while c.size >1 do

-(cmin1,cmin2)=minimum dist(ci,cj)for all ci,cjinc

- remove cmin1 and cmin2 from c

- add{cmin1, cmin2} to c

- l = l+1

end while

B. K-Clustering

In k-clustering the set of objects is divided to a certain number (k) clusters. We can distinguish different approaches from different points of view. The first classification is for hard and fuzzy clustering. In the first one, an object is assigned exactly to at least one cluster. The result is a membership matrix for objects and clusters with ones (the object is assigned to the cluster) and zeroes (the object is not assigned to the cluster). In the second approach membership degrees are calculated for all cluster-object pairs. Moreover, some other approaches to expressing uncertainty in cluster analysis have been proposed.

1. Initialize cluster centroids µ1,µ2,.... µk € R

2. Repeat until convergence: {

For every i, set c(i) :=arg minj||x(i) - µj

For each j, set

C. Multi-linear Regression

As a predictive analysis, the multiple rectilinear regression is employed to elucidate the connection between one continuous variable and two or more independent variables. The independent variables can be continuous or categorical.Relevant to understand the correlation between our variables and against the single response

D. Outlier Analysis

In data processing , anomaly detection (also outlier detection) is that the identification of things , events or observations which don't conform to an expected pattern or other items during a dataset. Instead, a cluster analysis algorithm could also be ready to detect the micro clusters formed by these patterns.

On top of all this, we've also performed eclat and apriori algorithm as well as Topic modeling.

Each week the Consumer Financial Protection Bureau sends thousands of consumers' complaints about financial products and services to companies for response. Those complaints are then compiled into a large dataset. The data specifically contains complaint information from Americans who have general debts such as student loans, mortgages, credit cards, consumer loans, and etc. Learning about this data set we have set out to analyze it and find patterns that help understand the finance complaints characteristics. Excel and Jupyter are the tools we plan on using to explore the data. Excel will be used to visually explore the data and to determine what parts of the data are going to be most useful.

Then Jupyter will be used to write code in R to clean, reduce and draw preliminary relationships in the data. We will clean the data by auto filling certain blank columns of each observation. Then we will remove certain columns that had irrelevant information. Then we remove observations that are missing a value in a crucial attribute column. We will stick to the basic outline of pre-processing data. Removal of observations rather than auto-filling indices will be more preferable so that the dataset's size will be reduced. Then we will continue on to performing modeling on the data in such as way as to reveal to us any pattern or correlation that can help solve or isolate certain complaints.

IV. RESULT:

By performing Hierarchical Clustering and K-Means clustering, we got a better insight by having 5 clusters. Figure1 and Figure2 shows

clusters inhaving the highest density while the rest of the have a lower density in comparison. The plots below show us the results for our clustering model. Cluster 2 & 3 have 5 elements in their cluster while clusters 1,4& 5 have significantly more elements in their clusters.

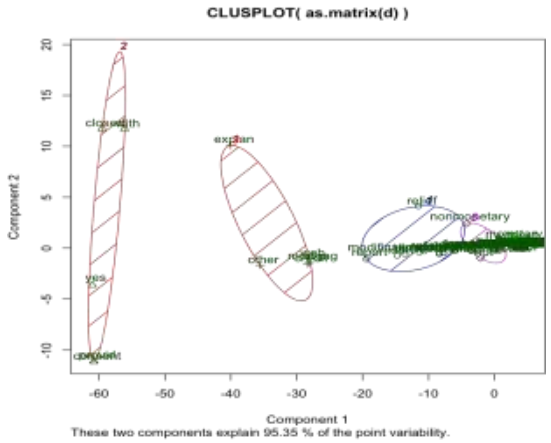

Figure 1. Clustering of frequently appeared words

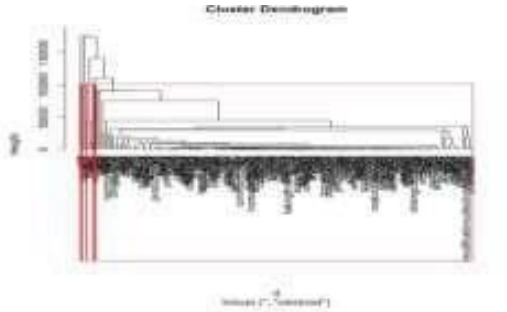

Figure 2. Clustering Dendrogram of the mostfrequent words with marked clusters

Reflecting on our cluster modeling results, we evaluated the performance of our model. Our clusters were graphed to depict the relationship between the various features against the known tags, we focus on the top five companies in our data set. However, the types of these five companies are different. Three of the companies are banks: Bank of America, Wells fargo, and JP morgan. The other two companies are consumer credit reporting agency: Experian and Equifax. Therefore, as can be seen in Figure3, the main product complaints for Bank of America, Wells fargo, and JPmorgan have to do with mortgage, credit cards, Bank accounts or service. While the main product complaints of Experian and Equifax have to do with credit reporting.

CONCLUSION

The results show what problems customers are having with specific problems in particular regions of the country. This valuable information will show where companies will need to invest in to improve their overall performance in the view of their customer. This will lead to improved customer satisfaction. By maximizing customer satisfaction, you can increase the opportunity for repeat sales to customers. Customer satisfaction also helps to increase customer loyalty, reducing the need to allocate marketing budget to acquire new customers. Satisfied customers can also recommend your products or services to other potential customers, increasing the potential for extra revenue and profit. Future research can be done to collect more complaint data so that we can perform analysis for other products, services and companies.

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 Services- Understanding through a Sample Survey of consumers of Present Day Banks in India Contributors, Banknet Publications, 4: 30-43. Mishra JK, Jain M (2007). Constituent Dimensions of Customer
[4] 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 various sorts 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.