

Identification of Customer Behaviors for Banking Market Strategy using Machine Learning

Submitted in partial fulfilment of the requirements for the award of
Bachelor of engineering degree in Computer Science and Engineering

By

SANTOSH KUMAR (17SCS0419)

SAMAR PRATAP (17SCS0410)



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING**

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY (DEEMED TO BE UNIVERSITY)

Accredited with Grade “A” by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI,

CHENNAI – 600 119

APRIL - 2021



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited with "A" grade by NAAC
Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this project report is the bonafide work of **SANTOSH KUMAR (17SCS0419) & SAMAR PRATAP (17SCS0410)** who carried out the project entitled **"IDENTIFICATION OF CUSTOMER BEHAVIOUR FOR BANKING MARKET STRATEGY USING MACHINE LEARNING"** under my supervision from **September 2020** to **April 2021**.

Internal Guide
Ms. R. Aishwarya M.E.,

Head of the Department
Dr.L.Lakshmanan M.E., Ph.D
Dr. S. Vigneswari, M.E., Ph.D.,

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I SANTOSH KUMAR hereby declare that the Project Report entitled “**IDENTIFICATION OF CUSTOMER BEHAVIOUR FOR BANKING MARKET STRATEGY USING MACHINE LEARNING**” is done by me under the guidance of **Ms. R. Aishwarya M.E.**, Department of Computer Science and Engineering at Sathyabama Institute of Science and Technology is submitted in partial fulfilment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering.

DATE:

PLACE: CHENNAI

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala, M.E., Ph.D., Dean**, School of Computing, **Dr. S. Vigneswari, M.E., Ph.D.**, and **Dr. L. Lakshmanan, M.E., Ph.D.**, Heads of the **Department** of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Ms. R. Aishwarya M.E., Assistant Professor**, for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

ABSTRACT

Customer segmentation is the process of classifying a customer into groups of individuals that are common in specific ways relevant to marketing, such as interests and spending habits. Customer segmentation is a marketing strategy in improving customer relationships. Companies employing customer segmentation operate under the fact that every customer is different and that their marketing efforts would be better served if they target specific, identifiable customers with shared needs. Customer segmentation is done using k-means clustering with elbow method to cluster customers into an effective number of clusters. With k-means scaled data, PCA analysis is done to depict the insights properly. Principal component analysis (PCA) is a method of multivariate statistical analysis. BP neural network is a multilayer feed forward neural network. In this study, we combined the advantages of principal component analysis and BP neural network and presented a customer segmentation model through analysing the retail consumer behaviour. Taking the results of PCA as the input of BP neural network, we used the VIP customer data of a retail business and verified the validity of the model. Although the main purpose of k-Means clustering is to partition samples into several homogeneous clusters by minimizing within-cluster errors, it has been shown that a relaxed solution of k-means can be recovered in a PCA-guided manner. With the PCA insights, we can clearly generate customer profiles.

TABLE OF CONTENTS

ABSTRACT	v
LIST OF FIGURES	viii

CHAPTER NO.	TITLE	PAGE NO.
1.	INTRODUCTION	
	1.1 INTRODUCTION	1
	1.2 DATA ANALYSIS	2
2.	AIM AND SCOPE	
	2.1 AIM OF PROJECT	3
	2.1.1 OBJECTIVES	3
	2.2 SCOPES	4
	2.2.1 ADVANTAGES	5
3.	SYSTEM DESIGN & METHODOLOGY	
	3.1 EXISTING SYSTEM	6
	3.2 PROPOSED SYSTEM	6
	3.3 ENVIRONMENTS AND TOOLS	7
	3.4 MODULES	8
	3.4.1 K-MEANS	8
	3.4.2 PRINCIPAL COMPONENT ANALYSIS (PCA)	8
	3.4.3 ELBOW METHOD	9

4.	SOFTWARE DEVELOPMENT METHODOLOGY	
	4.1 ARCHITECTURAL DESIGN	10
5.	RESULTS AND DISCUSSION	
	5.1 SCREENSHOTS	12
6.	CONCLUSION AND FUTURE WORK	
	6.1 CONCLUSION	16
	REFERENCES	17
	APPENDIX	
	A. LETTER OF ACCEPTANCE	18
	B. PLAGIARISM REPORT	19
	C. JOURNAL PAPER	20
	D. SOURCE CODE	25

LIST OF FIGURES

FIGURE NO.	FIGURE NAME	PAGE NO.
4.1.	ARCHITECTURE DIAGRAM	10
5.1.	HEATMAP OF THE DATASET	12
5.2	PLOT OF ELBOW METHOD	13
5.3	HISTOGRAM FOR DIFFERENT CLUSTERS	14
5.4	FINAL RESULT OF CLUSTERED CUSTOMERS USING PCA	15

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

With the increasingly fierce market competition, product is no longer the core of competition, customers are the key to winning decision. The management of the business is moving from "product-centric" to "customer-centric" that is customer relationship management (CRM). To retail enterprises, customer-centric, consumer personalized and the differences consumer demand became increasingly obvious. Through the analysis of different consumer behaviour, retail enterprises divide customers into different characteristics customer groups, which will contribute to the relevance and effectiveness of corporate marketing activities, enhance customer satisfaction and loyalty, and enable enterprises access to a stable development environment and competitive advantage. Customer segmentation is one of the CRM fundamental tasks, which has been more widely used in some industries and areas, including tourism, catering, retail trade, network marketing, network services and other e-commerce, etc. Alex. Berson used decision trees and clustering technology for customer segmentation [11]. Carrie M. Heilman studied the issue of customer classification using customer purchase data [12]. Jaesoo Kim used the neural networks in tourism industry customer classification [13]. Meng Xiaolian, Yang Yu proposed a customer identifying model based on customer value in commercial banks [14]. For the retail business, customer segmentation through certain methods will help the good implementation of customer relationship management. The customer segmentation based on the purchase behaviour may be an effective method of choice. In this study, we give a set of customer segmentation index system through the analysis of retail consumer behaviour, combine the advantages of principal component analysis and BP neural network. We screen the customer segmentation indicators through the principal component analysis and take the results as the input of BP networks. At last, a neural network model is constructed to realize the customer segmentation.

1.2 DATA ANALYSIS

Cluster analysis can be used define groups on several dimensions *simultaneously*. This type of analysis is very powerful but can be tricky to interpret. Our recommendation is to begin with simple crosstabs to determine what groupings make the most sense, then decide if its multivariate analysis is needed. We often find that managers respond well to cross tab tables (indexes and tables) because they are telegraphic and easy to understand.

A common way to use crosstab analysis is to compare the characteristics of a brand's most frequent customers to less frequent customers to address the question '*What do we have to do to make less frequent customers more like the frequent ones?*' One of our retail clients was surprised to learn the only meaningful difference between their most frequent customers and customers who come once or twice a year is their perceptions of the store's merchandise selection. In nearly every other way, they are similar – demographically, in terms of category involvement, even the other stores they shopped. By better communicating its depth and breadth of assortment, this client will attract less frequent customers to shop more often.

CHAPTER2

AIM AND SCOPE

2.1 AIM OF PROJECT

The aim of the project is to offers financial services to the general public. Typical services offered by retail banks include checking and savings accounts, personal loans, credit card access, and mortgage loans. Customers are segmented into different clusters and on that base customer profiles are created for better marketing strategies. These customer labels will help the banks to understand the customers they are handling and manage them accordingly.

2.1.1 OBJECTIVES

The Objectives are -

- To explore link between customer segmentation and marketing campaign and also to enhance by using PCA (Principal component analysis) insights.
- The Profiles of customers in each group could be analysed by marketers to make better strategies for each group.
- To establish better customer relationship management strategies.
- Markets can be more tightly defined as a people who have a true need or want for the company's offering, the ability to pay for it, and the willingness and authority to buy it. The total number of buyers must be large enough to be profitable for the company.
- Markets can also be a place such as a shopping centre. This identification of markets is useful for marketing decision-making purposes because factors such as product features, price, location of facilities, and promotional design are affected by geographic factors.

- The primary types of markets are consumer markets, industrial markets, institutional markets, and reseller markets. These categories are not always clear-cut, and, in some industries, a business may be in a different category altogether or may even encompass multiple categories

2.2 SCOPE

The premise of customer segmentation is that to maximize sales to a large population of customers, it is best to divide it into logical subgroups. The assumption is that by dividing one large, amorphous mass into subgroups, you can fine-tune your product, messaging, support, or distribution channels to meet the specific needs of unique customer groups. Thus, the goal is to use a customer segmentation model to improve marketing success and optimize marketing ROI.

Segmentation models vary from basic to complex, and the approaches to developing and applying them is a topic for an entire book itself. But here are some examples:

Example 1: A telecommunications company selling mobile phone services might segment its market based on complexity of needs. One customer group might only need voice service, and very little volume at that. Another group might primarily use mobile texting services. Yet another might be a heavy user of mobile phones for voice, text, email, and web browsing. By identifying distinct patterns in customer needs, the company can optimize product bundles and target them at the correct audiences.

Example 2: A hotel chain that caters to families might simply segment its market based on income level and travel frequency. The chain might find that a group of moderate income-frequent travellers exists that is swayed by certain loyalty program rewards. It might find another important group exists—that is more swayed by on-site amenities. With this information, the hotel can optimize its offerings and loyalty programs to appeal to each group's unique needs.

Market segmentation studies are fairly involved. They are often multi-phased projects that can get very complex, especially for global organizations.

In fact, market segmentation studies are among the most complex types of projects and have a higher risk of failure as compared to other studies. Even when completed as planned, many clients find it challenging to truly apply the results. Market segmentation can be done with both consumer and business populations. That said, the studies for B2B populations are very different than those for B2C populations.

Take this into consideration when selecting potential agencies for this type of project. A firm that has strong experience in consumer segmentation does not necessarily have B2B segmentation expertise.

2.2.1 ADVANTAGES

- Determine appropriate banking policies.
- Develop customized marketing campaigns.
- Design an optimal policy strategy.
- Choose specific banking features for deployment.
- Prioritize new policy development efforts.

CHAPTER 3

SYSTEM DESIGN & METHODOLOGY

3.1 EXISTING SYSTEM

- Existing system consists of only k-means clustering.
- Number of clusters in existing system are vague.
- Lack of PCA (Principal component analysis) which is an efficient way of clustering high-dimensional data.

3.2 PROPOSED SYSTEM

- In the proposed system, the analysis of customer is clear and were segmented efficiently.
- It contains all the insights which we need to cluster the customers or users.
- Elbow method is used to find the affective number of clusters needed to cluster all the customers.
- The data set which we used is high-dimensional.
- So, to depict the clusters properly and efficiently we used PCA (Principal component analysis).
- We can clearly see how customers are segmented and on what base they are segmented.

3.3 ENVIRONMENTS AND TOOLS

1. scikit-learn- The scikit-learn project started as scikits.learn, a Google Summer of Code project by David Cournapeau. Its name stems from the notion that it is a "SciKit" (SciPy Toolkit), a separately-developed and distributed third-party extension to SciPy.
2. Seaborn- Seaborn is a library that uses Matplotlib underneath to plot graphs. It will be used to visualize random distributions.
3. NumPy- NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
4. Pandas- pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.
5. Matplotlib- Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK.
6. python 3.8- Python is an interpreted, high-level and general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant indentation.
7. Jupyter notebook- Project Jupyter is a project and community whose goal is to "develop open-source software, open-standards, and services for interactive computing across dozens of programming languages". It was spun off from IPython in 2014 by Fernando Pérez.

3.4 MODULES

3.4.1 K-MEAN CLUSTERING

It generally partitions n -observations into k -clusters. Clusters created based on central mean value like (student, current employee, retired people, former, senior citizen) and (deposit, withdraw, transactions) as the central mean value. K -means often divides a large cluster into several small groups or merges small adjacent clusters into a larger one to get minimal SSE. Clustering is the process of dividing a set of physical or abstract objects into groups of similar objects. The K -means algorithm, as one of the most popular clustering algorithms, was first used by Macqueen in 1967, and it has been used extensively in various fields including data mining, statistical data analysis, and other business applications.

The literature shows that one of the major applications of K -means is customer segmentation. The K -means algorithm is widely used to effectively identify valuable customers and develop pertinent marketing strategies. In particular, Cheng & Chen used a RFM model and K -means to perform customer relationship management, and experimental results demonstrate that the model they proposed is an effective method in customer value analysis. Khalili-Damghani et al. proposed a hybrid soft computing approach based on clustering, rule extraction, and decision tree methodology to predict segmentation of new customers of customer-centric companies. This approach was applied in two case studies in the fields of insurance and telecommunication, respectively, aiming to predict potentially profitable leads and to outline the most influential features available to customers during such prediction.

3.4.2 PRINCIPAL COMPONENT ANALYSIS (PCA)

Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss. It

does so by creating new uncorrelated variables that successively maximize variance. PCA is a method that reduces the dimensionality of a dataset, by finding a new set of variables, smaller than the original set of variables. This efficient reduction of the number of variables is achieved by obtaining orthogonal linear combinations of the original variables – the so-called Principal Components (PCs). PCA is useful for the compression of data and to find patterns in high-dimensional data.

3.4.3 ELBOW METHOD

The elbow method is based on the observation that increasing the number of clusters can help to reduce the sum of within-cluster variance of each cluster. This is because having more clusters allows one to capture finer groups of data objects that are more like each other. To define the optimal clusters, Firstly, we use the clustering algorithm for various values of k . This is done by ranging k from 1 to 10 clusters. Then we calculate the total intra-cluster sum of square. Then, we proceed to plot intra-cluster sum of square based on the number of clusters. The plot denotes the approximate number of clusters required in our model. The optimum clusters can be found from the graph where there is a bend in the graph.

Work of Elbow method:

1. Compute K-Means clustering for different values of K by varying K from 1 to 10 clusters.
2. For each K , calculate the total within-cluster sum of square (WCSS).
3. Plot the curve of WCSS vs the number of clusters K .
4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

CHAPTER 4

SOFTWARE DEVELOPMENT METHODOLOGY

4.1 ARCHITECTURAL DESIGN

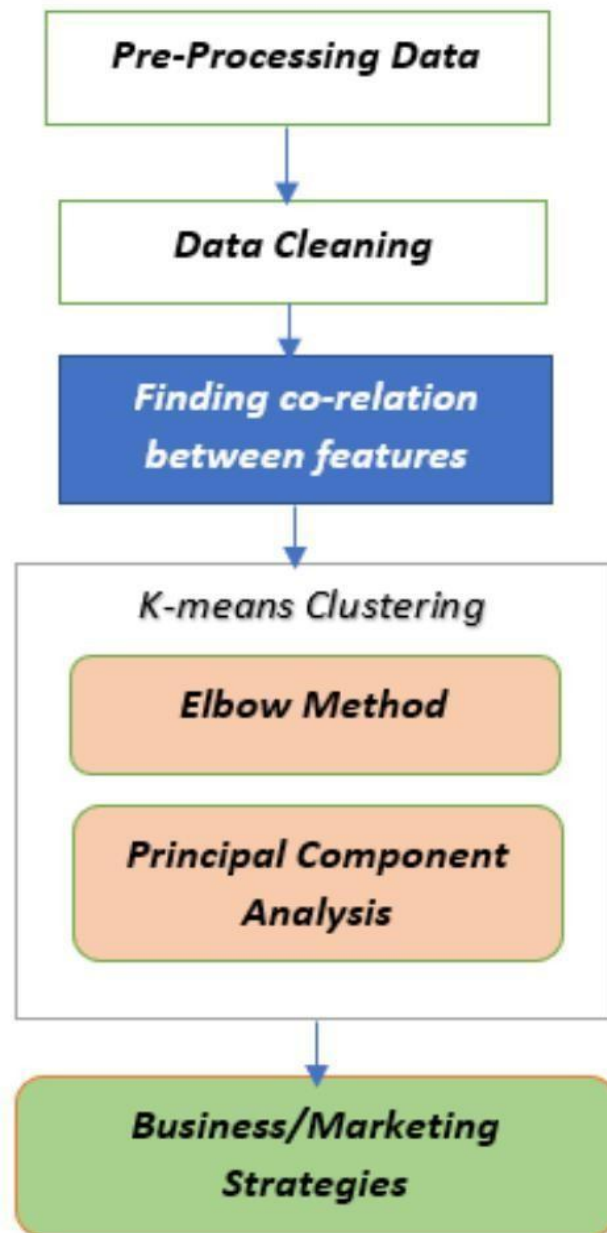


Figure 4.1 Architecture Diagram

CHAPTER 5

RESULTS AND DISCUSSION

Customer segmentation has highly improved the performance of different businesses over the years, the fact that the consumer platform is changing dramatically has proven to be a headache for various businesses. That's why segmenting the customers into different groups has some incredible benefits for improving the market results. Earlier, the business markets used to be localized, thereby limiting the customer pool of specific needs, tastes, choices, price preferences, and so on. However, in today's scenario, this localized market has overcome the boundaries and taken the shape of a global village, where interaction between different levels of the business isn't restricted anymore. With the globalization of the market, the customer pool is welcoming more new variations in each factor. And perhaps this is the reason why businesses need to study the differences in the consumer behaviourism in order to have detailed knowledge about the market.

The performance of the proposed system is evaluated by working on the banking transactions data set of the customers. The dataset consists of 18 features with 8950 points. The missing values in the dataset are filled with mean value of the column during data pre-processing. Post processing of data, heat-Map is used to find correlation between the instances (Figure 2). Elbow method is used to find number of clusters needed to implement k-means (Figure 3). K-means is used to partitions n-observations into k-clusters. Clusters are created based on the central mean value. The k-means insights were depicted using histograms (figure 4). The results indicate that unsupervised dimension reduction is closely related to unsupervised learning. Several implications are discussed. On dimension reduction, the result provides new insights to the observed effectiveness of PCA-based data reductions, beyond the conventional noise-reduction explanation that PCA, via singular value decomposition, provides the best lowdimensional linear approximation of the data (figure 5). On learning, the result suggests effective.

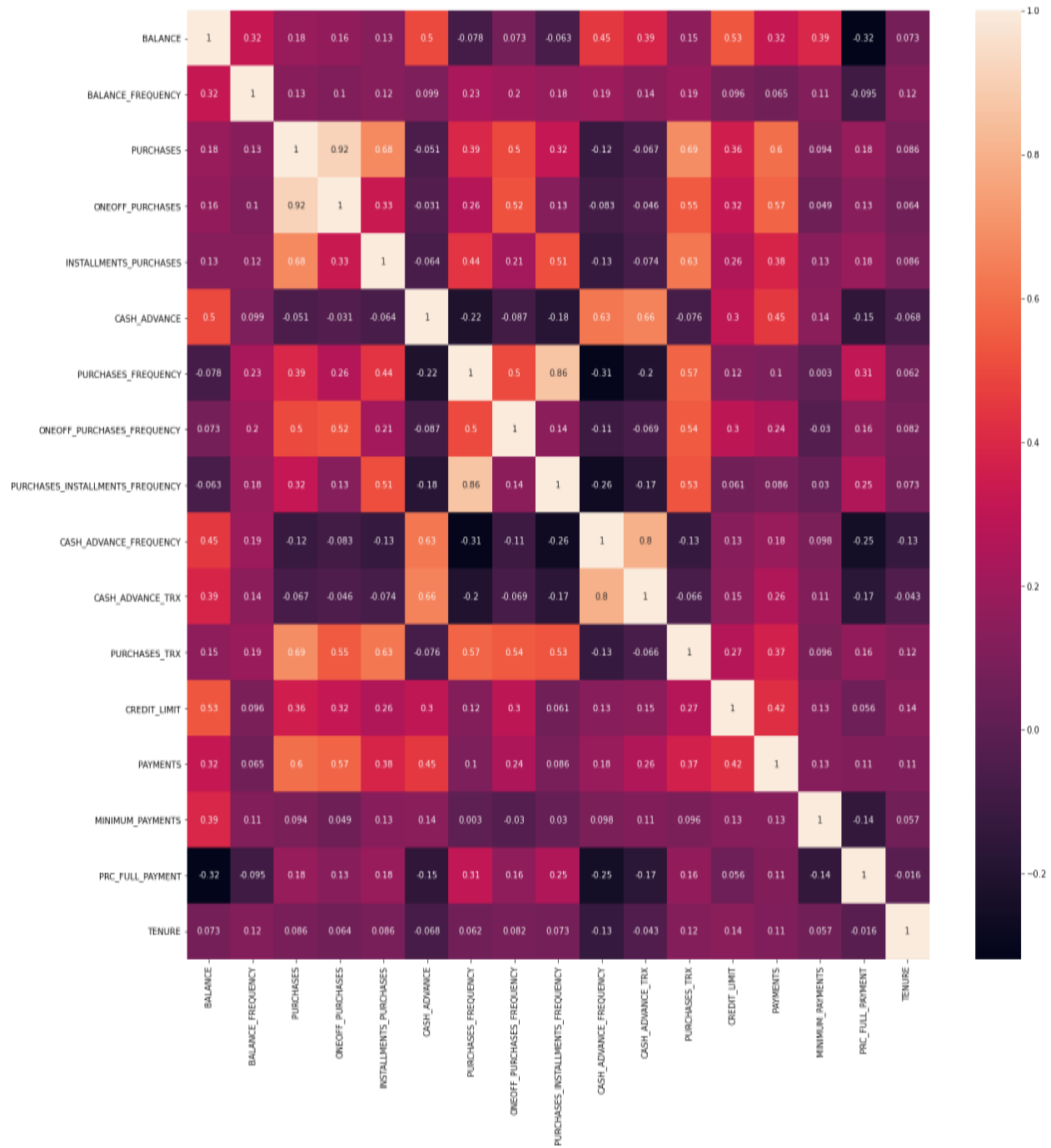


Figure 5.1. HeatMap of the dataset

Finding the right number of clusters

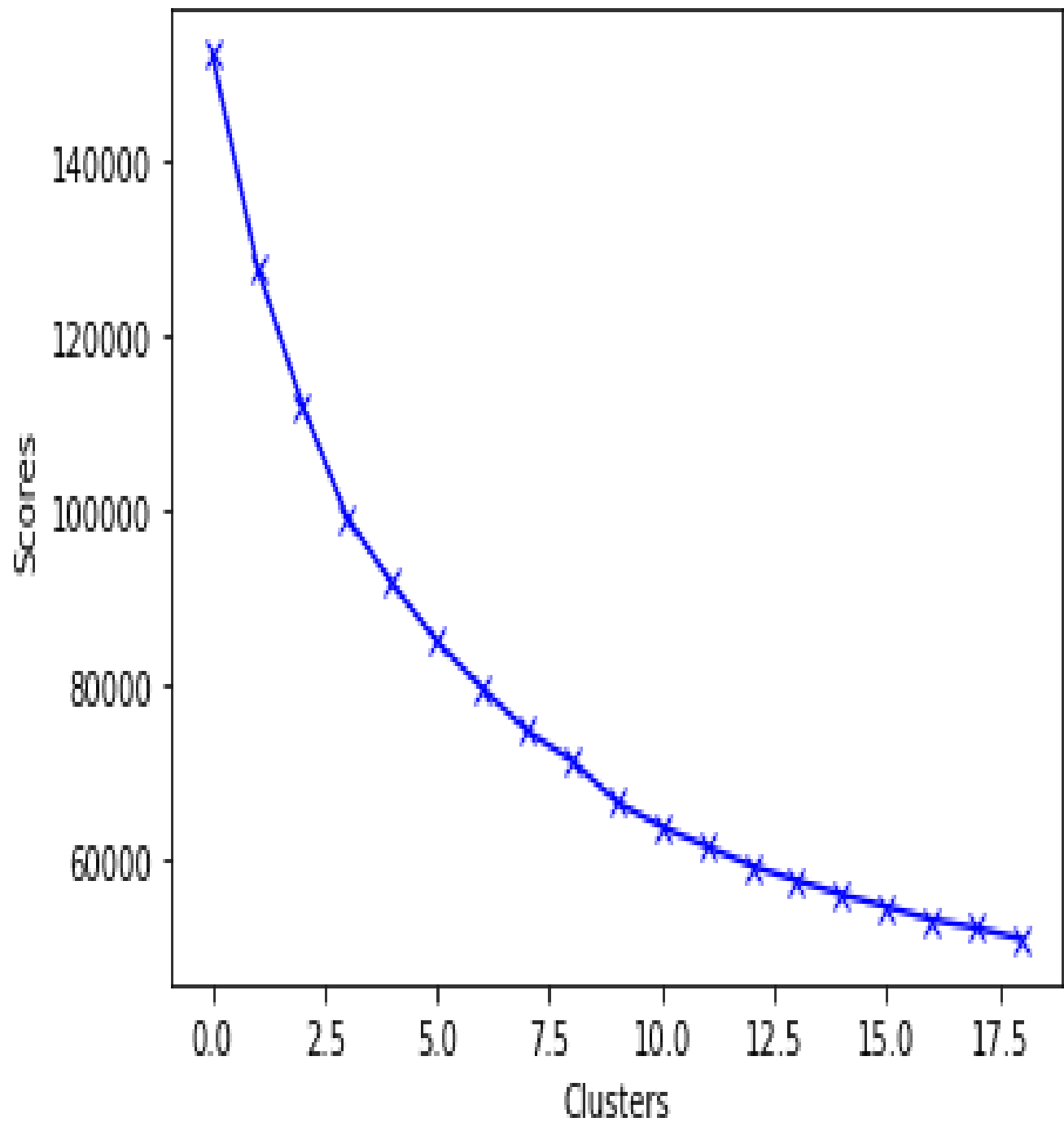


Figure 5.2 Plot of elbow method

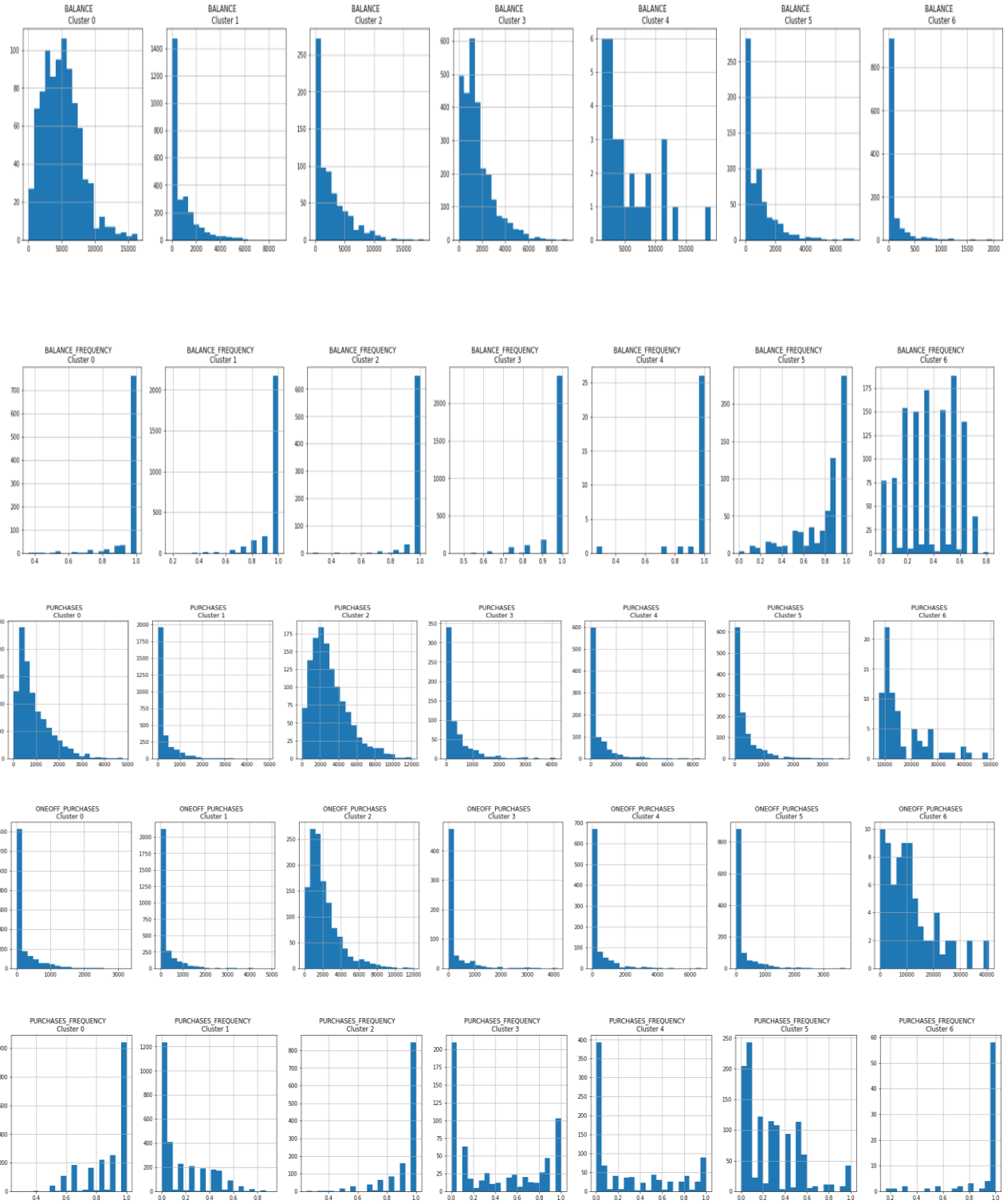


Figure 5.3 Histograms for different clusters

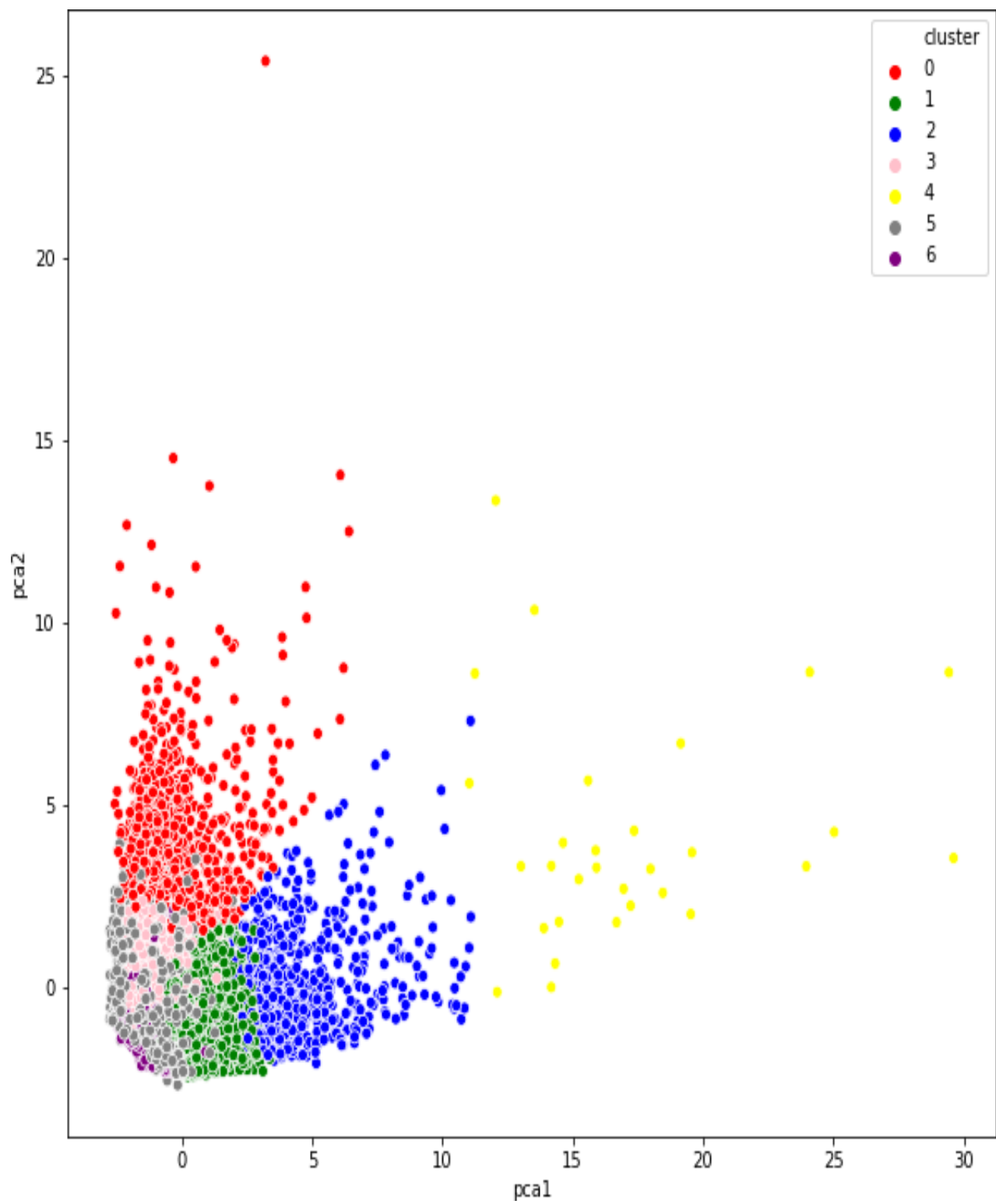


Figure 5.4 Final result of clustered customers using PCA

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 CONCLUSION

Segmenting the customers will deepen the relationships with customers. Finding new customers for the enterprise is vital, meanwhile retaining the existing clients [8]. In our proposed work clustering is done using k-means. Elbow method is used to find the number of clusters needed for clustering. Further, PCA (Principal Component Analysis) is implemented to plot the clusters in an effective manner. The results were analyzed, and customer profiles are generated using the insights of PCA. The implemented algorithm is efficient because it can cluster high-dimensional data. The final customer clusters (profiles) can be analysed by bankers and accordingly they can manage the banking strategies based on the spending habits of the customers. Future work includes analysis of customer profiles deeply such as the products which are bought frequently by the members of each clusters. This would help in providing better promotional offers to specific products.

REFERENCES

- [1] H. Su-li, 2010 "The customer segmentation of commercial banks based on unascertained clustering,". doi: 10.1109/ICLSIM.2010.5461416.
- [2] He and Li (2016) suggested a three-dimensional approach to improving the customer lifetime (CLV).
- [3] Jiang and Tuzhilin (2009) identified that both customer segmentation and buyer targeting are necessary to improve the marketing performances.
- [4] Kolarovszki et al., 2016. Group-specific marketing is common and much needed from the traditional mass marketing perspective. Customer segmentation is a part of various activities under Customer Relationship Management (CRM) value chain (Kolarovszki et al., 2016, Khalili-Damghani et al., 2018, Khajvand et al., 2011).
- [5] Lu et al. (2014) analysed the customer churn prediction. The authors used logistic regression and isolated the transactional data for creating a new distinct prediction model. doi: <https://doi.org/10.1016/j.jksuci.2018.09.004>
- [6] Khalili-Damghani et al., 2018. The proposed hybrid method is to predict the new entities for the customer centric companies.
- [7] Sheshasaayee and Logeshwari (2017) designed a new integrated approach by segmentation with the RFM and LTV (Life-Time Value) methods.
- [8] Tong, L., Wang, Y., Wen, F., Li, X., Nov. 2017. The research of customer loyaltyimprovement in telecom industry based on NPS data mining. China Commun.14(11),260–268.<https://doi.org/10.1109/CC.2017.8233665>.
- [10] Christy, A.J., et al. RFM ranking – An effective approach to customer segmentation. Journal of King Saud University –Computer and Information Sciences (2018), <https://doi.org/10.1016/j.jksuci.2018.09.004>

APPENDIX

A. LETTER OF ACCEPTANCE



5th International Conference on
Intelligent Computing and Control Systems
ICICCS 2021 | May 06 - 08, 2021
iciccs.com/2021/ | iccs.conf19@gmail.com

Letter of Acceptance

TO
Santosh Kumar, Samar Pratap, Ms. R. Aishwarya, Ms. R. Yogitha, G. Kalaiarasi
Sathyabama Institute of Science and Technology, Chennai, India,

Paper ID- ICICCS198

Subject: Acceptance to the International Conference on Intelligent Computing and Control Systems [ICICCS 2021], 06-08, May 2021, Madurai, India.

Dear Author,
We are happy to inform you that your paper titled "**Identification of Customer behaviors for Banking Market Strategy using Machine Learning**" has been accepted for the oral presentation at the International Conference on Intelligent Computing and Control Systems [ICICCS 2021] to be held on **6 – 8, May 2021** at Vaigai College of Engineering, Madurai, Tamil Nadu, India.

With the evident of its previous publications, ICICCS 2021 is also dedicated for the publication in **IEEE Xplore Digital library**. As a result of the peer review process, the technical conference program committee is pleased to inform you that your paper is shortlisted and accepted for the presentation in conference event and formally accepted for inclusion in IEEE Xplore. We appreciate if you could submit the final version of manuscript at your earliest convenience, in order to ensure a novel and timely publication of your research paper.

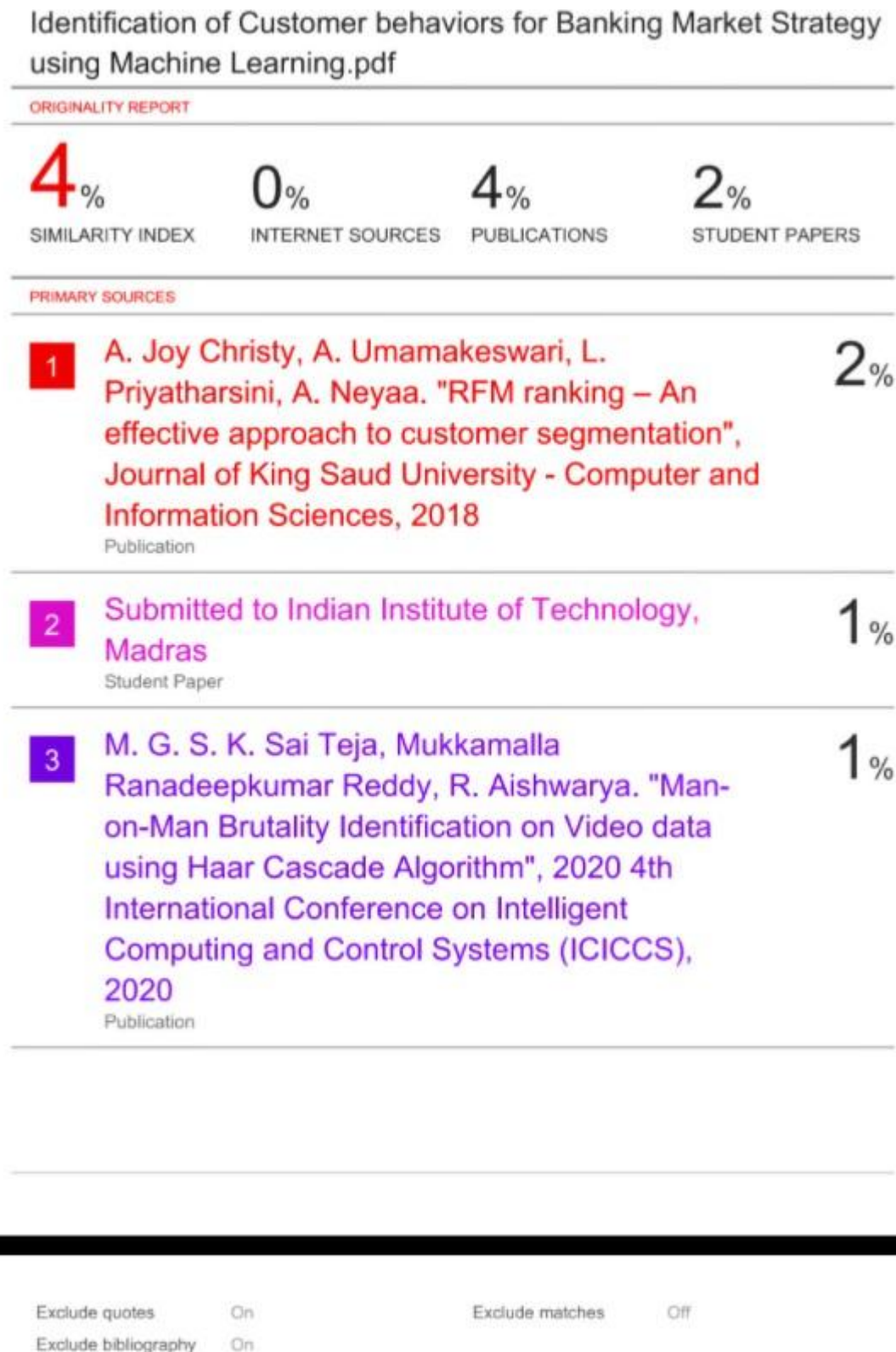
Once again, on behalf of the conference committee we extend our warm gratitude to welcome you at our conference site.
Yours' Sincerely,


Prof. P. Sugumaran
Vice Principal,
Vaigai College of Engineering,
Madurai, India.

Our Previous Publications
[ICICCS 2020 Publication Link](#)
[ICICCS 2019 Publication Link](#)
[ICICCS 2018 Publication Link](#)
[ICICCS 2017 Publication Link](#)

Proceedings by


B. PLAGIARISM REPORT



Identification of Customer behaviors for Banking Market Strategy using Machine Learning

Santosh Kumar¹, Samar Pratap², Ms. R. Aishwarya M.E.³

¹Sathyabama Institute of Science and Technology, Chennai, India, kumar.san0808@gmail.com

²Sathyabama Institute of Science and Technology, Chennai, India, samarpratap2050@gmail.com

³Sathyabama Institute of Science and Technology, Chennai, India, aishwarya.cse@sathyabama.ac.in

ABSTRACT

Customer segmentation is a process of classifying the customer into clusters of individuals that are standard in definite ways appropriate to the market, such as spending interests and habits. Customer clustering is a marketing technique in enhancing client bonding. Enterprises employing customer segmentations function on the fact that each and every customer is distinct and their marketing strategies and proceedings would be better organized if they target individual customers with collaborative needs. Customer segmentation is done using k-means clustering with elbow method to cluster customers into an effective number of clusters. With k-means scaled data, PCA analysis is done to depict the insights properly. However, the primary objective of k-Mean's clustering is to divide attributes into various similar segments by reducing intra-cluster errors; it has been demonstrated that an effective result of k-means can be redeemed in a PCA-directed manner. With the PCA insights, we can clearly generate customer profiles.

Keywords-Customer Segmentation, k-means, elbow method, Principal Component Analysis (PCA).

1. INTRODUCTION

In this decade, an enormous urge in the contention amidst companies to retain in the business has been seen. Many banks fail to get financial gains. The gains of the banks can be enhanced by the customer clustering model. Maintaining old customers is more vital than acquiring new clients.

The objective of this study is-

- To identify potential clients by providing them proper and prompt data to business units in Banking Firms.
- The Customers Profiles in each cluster could be evaluated by bankers to make improved tactics for every cluster.

- To increase the profitability of the firm and to know how to maintain the customer owned.

Marketing has become an important process in the banking firms which makes banks pay more attention to enhance the market strategies. To obtain favorable outcomes over their market, they have to determine, comprehend, and aim the money-making customers from other non-profitable members. It is not possible for the banks to give one-to-one attention to every customer. To overcome this problem, the paper proposes a k-means associating method which uses the Elbow technique to get the exact number of clusters needed to segment the customers. On top of this, k-means scaled data are used to analyze the customers and depict them into appropriate groups by using Principal Component Analysis (PCA). The output insights can be used to generate the customer profiles.

The generated profiles can be analyzed by bank marketers to increase the profitability and to set market strategies.

2. LITERATURE SURVEYH.

Su-li, 2010 [1]. On the basis of the actuality of China's commercial banks, the paper designed the CLV assessment indicator system, furthermore, the unsupervised learning has been applied to set up the customer clustering model of commercial banks.

Shah and Singh (2012). Recommended new fragmenting calculation that executes the same the K-implies strategy. The proposed strategy doesn't give a genuine arrangement in any cases; however, it diminishes the bunching blunder standard.

Lu et al. (2014). Analysed the client money forecast [4]. The creators utilized coordinated relapse and distant value-based information for making another diverse suspicion model. With this exploratory execution, it is seen that the clients with enormous money esteems can be

recognized furthermore, can be continued utilizing singular business techniques.

He and Li (2016) proposed a three-dimensional system to improve the client's lifetime esteem (CLV), the satisfaction of the clients and their practices. The creator has understood that the clients are not quite the same as one another thus their necessities are additionally not the equivalent. Grouping makes a difference in discovering final proposal, supposition and supporting a great update.

Sheshasaayee and Logeshwari (2017). Arranged a new brought together strategy by bunching with the RFM qualities and LTV. These two acquired a two-stage strategy with the initial step being the segment strategy and the 2nd step is to do division. The Author's point is to do K-implies division after the two-stage demonstrating and afterward utilizing a neural organization to escalate their grouping.

3. PROPOSED WORK

In the proposed system, the analysis of customers is clear and was segmented efficiently. It contains all the insights which we need to cluster the customers. Elbow technique is used to get the effective number of segments needed to cluster all the customers. The data set which we used is high-dimensional. So, to depict the clusters properly, PCA (Principal component analysis) is implemented. We can clearly see how customers are segmented and on what base they are segmented. And on the basis of k-means scaled data, customer profiles can be generated and these generated profiles can be analysed by the bank marketers to increase profitability. Here are the modules which were used in our proposed system-

3.1 Pre-processing: The dataset is pre-processed and cleansed. Missing values were filled with the mean value of the column. Removal of unnecessary columns to make the dataset more presentable.

3.2 HeatMap: To find the correlation between features, heatmap is plotted to get the correlation matrix between features. Heatmap shows the correlation between the features. It can be used to find out how one column is related to another.

3.3 Elbow method: Used to track down the ideal number of Clusters utilizing the Elbow Method. A heuristic strategy of portrayal and authentication of consistency in group examination intended to help get the

reasonable number of bunches. Assuming the given diagram resembles an arm, the "elbow" on that arm is the estimation of k that should be thought of. Utilizing this suitable number of bunches are found.

3.4 K-Means Clustering: It divides n-observations into k-clusters. Clusters are depicted on the basis of central mean value. K-means generally splits a large cluster into various small segments or transforms small bordering segments into a larger cluster to acquire minimal Sum of Squares (SSE) error.

3.5 PCA: PCA stands for Principal Component Analysis. It is a method for decreasing the largeness of such datasets, optimizing interpretability but at the same time scaling down information loss. PCA effectively does that by implementing novel uncorrelated variables that successively optimize variance. PCA insights shows how customers are segmented. And using PCA we can generate the customer profiles. And hence the generated profiles can be used by bank marketers to increase the profitability and market strategies.

4. SYSTEM ARCHITECTURE

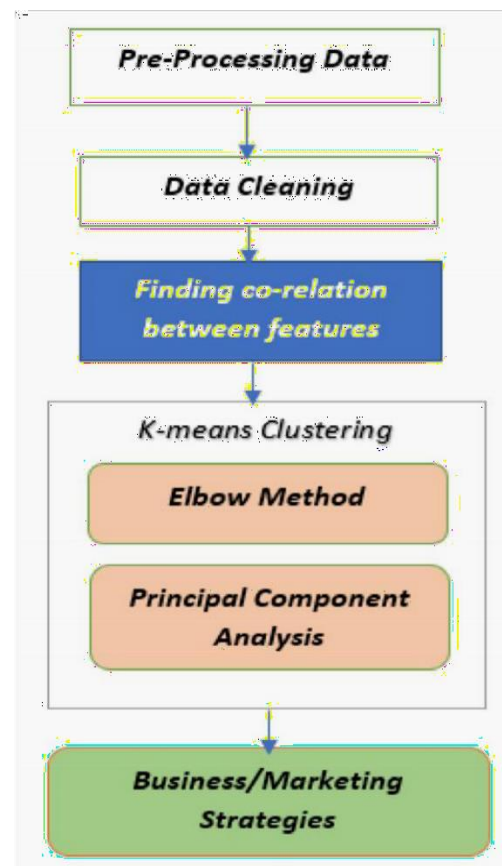


Figure 1. Proposed System Architecture

5. EXPERIMENTAL RESULTS

The potential of the proposed system is assessed by operating on the banking transactions dataset of the customers. The dataset consists of 18 features with 8950 points. The missing values in the dataset are filled with the mean value of the column during data pre-processing. Post processing of data, heatmap is used to find correlation between the instances (Figure 2). Elbow technique is used to get the appropriate number of clusters needed to implement k-means (Figure 3). K-means is used to partition n-observations into k-clusters. Clusters are established based on the mean value of the attribute. The k-means insights were depicted using histograms (figure 4). The outcomes specify that proportion decrement is firmly linked to unrul study. Various inferences are explored. Upon Dimension deduction, the outcome gives naïve understanding to the adhered potency of PCA info dimension deductions, over the traditional noise-reduction demonstration that PCA, through single value degradation, equips the prime less-dimensional straight approximation (figure 5). On studying, the assessment demonstrates operative methods for K-means data clustering.

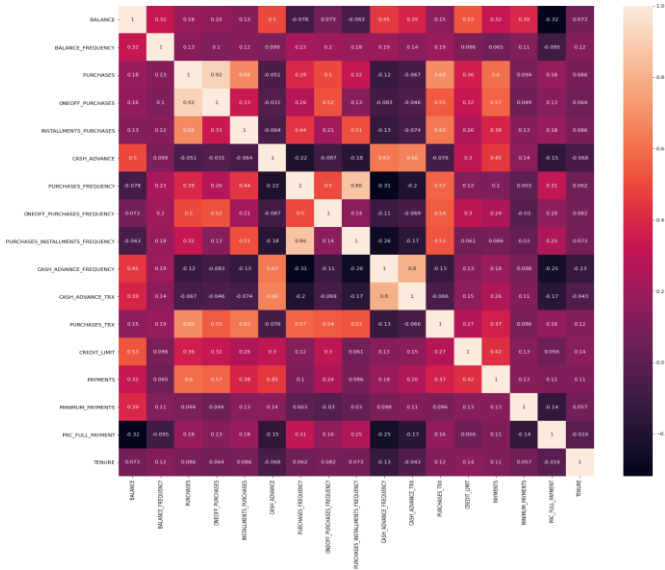


Figure2. HeatMap of the dataset

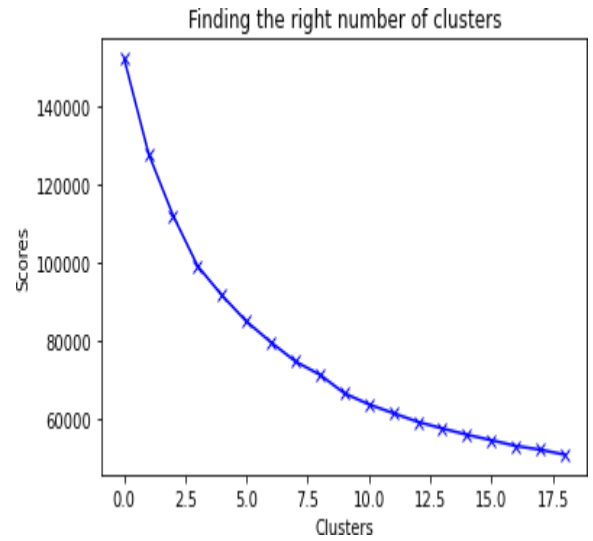


Figure 3. Plot of elbow method

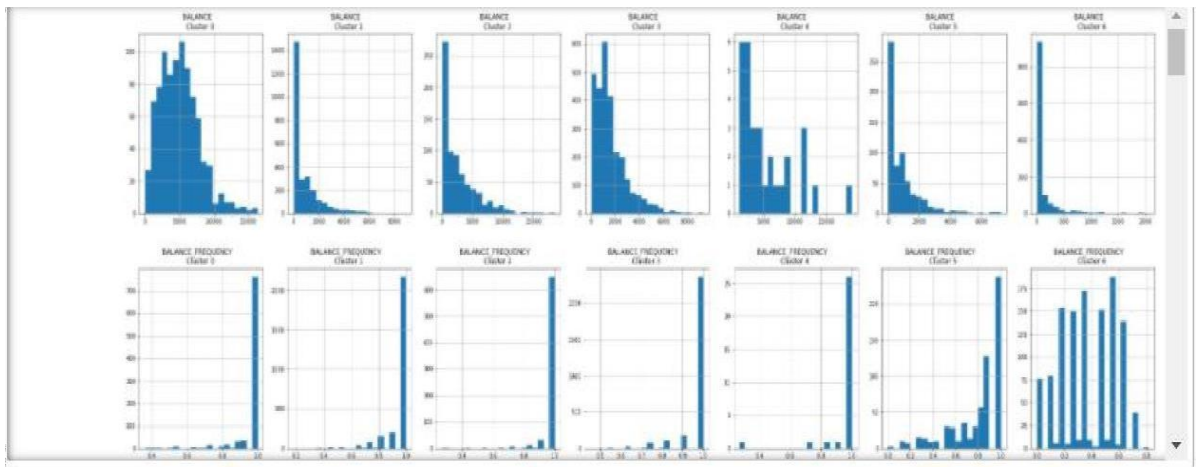


Figure 4. Histogram for different clusters

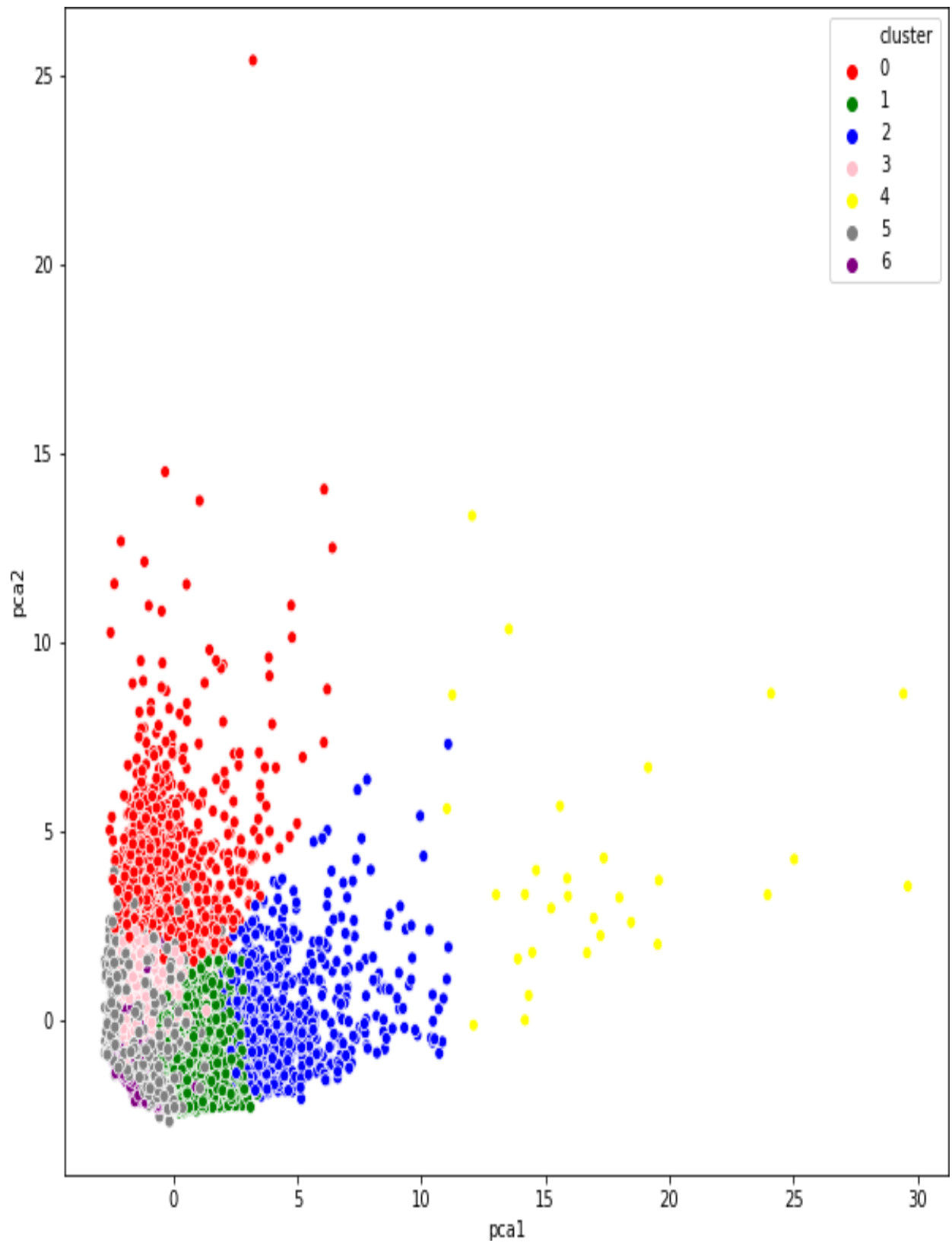


Figure 5. Result of clustered customers using PCA

6. CONCLUSION

Clustering the members will intensify the connection with the customers. Searching new clients for the association is critical, then proceeding with the current customers. In our proposed work clustering is done using k-means. Elbow technique is used to get the appropriate number of clusters needed for clustering. Further, PCA (Principal Component Analysis) is implemented to plot the clusters in an effective manner. The results were analysed, and customer profiles are generated using the insights of PCA. The implemented algorithm is efficient because it can cluster high-dimensional data. The final customer clusters (profiles) can be analysed by bankers and accordingly they can manage the banking strategies based on the spending habits of the customers. Future work includes analysis of customer profiles deeply such as the stocks which are traded regularly by the customers of each and every segment. Deep analysis would help in giving effective propagative offers to distinct investments

REFERENCES

- [1] H. Su-li, 2010 "The customer segmentation of commercial banks based on unascertained clustering,". doi: 10.1109/ICLSIM.2010.5461416.
- [2] Kolarovszki et al., 2016. Gathering explicit banking is actually something similar and much needed from the customary style banking POV. Client grouping is a piece of numerous exercises under the CRM esteem chain.
- [3] Khalili-Damghani et al., 2018. Recommended technique is basically for accepting the new elements for the client driven organizations.
- [4] Lu et al. (2014) dissected the client account suspicion. They utilized oversight relapse and detached the exchanged information for making another distinctive suspicion model.
- [5] Tong, L., Wang, Y., Wen, F., Li, X., Nov. 2017. The exploration of client loyalty improvement in telecommunication firms dependent on NPS information mining. China Commun.14(11),260–268. <https://doi.org/10.1109/CC.2017.8233665>.

D. Source Code2

Import Libraries and Load Dataset

```
import pandas as pd import numpy as np import seaborn as
sns import matplotlib.pyplot as plt from
sklearn.preprocessing import StandardScaler, normalize from
sklearn.cluster import KMeans
from sklearn.decomposition import PCA

creditcard_df = pd.read_csv('Dataset.csv')

# CUSTID: Identification of Credit Card holder
# BALANCE: Balance amount left in customer's account to make purchases
# BALANCE_FREQUENCY: How frequently the Balance is updated, score between
0 and 1 (1 = frequently updated, 0 = not frequently updated)
# PURCHASES: Amount of purchases made from account
# ONEOFFPURCHASES: Maximum purchase amount done in one-go
# INSTALLMENTS_PURCHASES: Amount of purchase done in installment
# CASH_ADVANCE: Cash in advance given by the user
# PURCHASES_FREQUENCY: How frequently the Purchases are being made, score
between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased) #
ONEOFF_PURCHASES_FREQUENCY: How frequently Purchases are happening in on
e-go (1 = frequently purchased, 0 = not frequently purchased)
# PURCHASES_INSTALLMENTS_FREQUENCY: How frequently purchases in installmen
ts are being done (1 = frequently done, 0 = not frequently done) #
CASH_ADVANCE_FREQUENCY: How frequently the cash in advance being paid
# CASH_ADVANCE_TRX: Number of Transactions made with "Cash in Advance"
# PURCHASES_TRX: Number of purchase transactions made
# CREDIT_LIMIT: Limit of Credit Card for user
# PAYMENTS: Amount of Payment done by user
# MINIMUM_PAYMENTS: Minimum amount of payments made by user
# PRC_FULL_PAYMENT: Percent of full payment paid by user
# TENURE: Tenure of credit card service for user
creditcard_df.head()

creditcard_df.info()
# 18 features with 8950 points
creditcard_df.describe()
# Mean balance is $1564
# Balance frequency is
frequently updated on
average ~0.9
# Purchases average is $1000
# one off purchase average is ~$600
# Average purchases frequency is around 0.5
```

```
# average ONEOFF_PURCHASES_FREQUENCY, PURCHASES_INSTALLMENTS_FREQUENCY, and
CASH_ADVANCE_FREQUENCY are generally low
# Average credit limit ~ 4500
# Percent of full payment is 15%
# Average tenure is 11 years
```

Visualize and Explore Dataset

```
# Let's see if we have any missing data, luckily we don't have many!
sns.heatmap(creditcard_df.isnull(), yticklabels = False, cbar = False, cma
p="Blues");
creditcard_df.isnull().sum()
```

```
# Fill up the missing elements with the mean values
creditcard_df.loc[(creditcard_df['MINIMUM_PAYMENTS'].isnull() == True), 'M
INIMUM_PAYMENTS'] = creditcard_df['MINIMUM_PAYMENTS'].mean()
creditcard_df.loc[(creditcard_df['CREDIT_LIMIT'].isnull() == True), 'CREDI
T_LIMIT'] = creditcard_df['CREDIT_LIMIT'].mean()
sns.heatmap(creditcard_df.isnull(), yticklabels = False, cbar = False,
cma p="Blues");
```

```
# Let's drop Customer ID since it has no meaning here
creditcard_df.drop("CUST_ID", axis = 1, inplace= True)
creditcard_df.head()
```

```
len(creditcard_df.columns)
```

```
# distplot combines the matplotlib.hist function with seaborn kdeplot()
# KDE Plot represents the Kernel Density Estimate
# KDE is used for visualizing the Probability Density of a continuous vari
able.
# KDE demonstrates the probability density at different values in a contin
uous variable.
```

```
# Mean of balance is $1500
# 'Balance_Frequency' for most customers is updated frequently ~1
# For 'PURCHASES_FREQUENCY', there are two distinct group of customers #
For 'ONEOFF_PURCHASES_FREQUENCY' and 'PURCHASES_INSTALLMENT_FREQUENCY' m
ost users don't do one off purchases or installment purchases frequently
# Very small number of customers pay their balance in full 'PRC_FULL_PAYME
NT'~0
# Credit limit average is around $4500
```

```

# Most customers are ~11 years tenure
plt.figure(figsize=(10,50)) for i in
range(len(creditcard_df.columns)):
    plt.subplot(17, 1, i+1)
sns.distplot(creditcard_df[creditcard_df.columns[i]], kde_kws={"color":
"b", "lw": 3, "label": "KDE"}, hist_kws={"color": "g"})
plt.title(creditcard_df.columns[i])
plt.tight_layout()

# obtain the correlation matrix between features
correlations = creditcard_df.corr() f, ax =
plt.subplots(figsize = (20, 20))
sns.heatmap(correlations, annot = True)

# 'PURCHASES' have high correlation between oneoff purchases, 'installment
purchases, purchase transactions, credit limit and payments.
# Strong Positive Correlation between 'PURCHASES_FREQUENCY' and 'PURCHASES
_INSTALLMENT_FREQUENCY'

```

Find the Optimal Number of Clusters using Elbow Method

- The elbow method is a heuristic method of interpretation and validation of consistency within cluster analysis designed to help find the appropriate number of clusters in a dataset.
- If the line chart looks like an arm, then the "elbow" on the arm is the value of k that is the best.
- Source:
 - [https://en.wikipedia.org/wiki/Elbow_method_\(clustering\)](https://en.wikipedia.org/wiki/Elbow_method_(clustering))
 - <https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-means/>

```

# Let's scale the data first scaler
= StandardScaler()
creditcard_df_scaled =
scaler.fit_transform(creditcard_df
) creditcard_df_scaled.shape
creditcard_df_scaled

# Index(['BALANCE', 'BALANCE_FREQUENCY', 'PURCHASES', 'ONEOFF_PURCHASES',
#       'INSTALLMENTS_PURCHASES', 'CASH_ADVANCE', 'PURCHASES_FREQUENCY',
#       'ONEOFF_PURCHASES_FREQUENCY', 'PURCHASES_INSTALLMENTS_FREQUENCY',
#       'CASH_ADVANCE_FREQUENCY', 'CASH_ADVANCE_TRX', 'PURCHASES_TRX',

```

```

#         'CREDIT_LIMIT', 'PAYMENTS', 'MINIMUM_PAYMENTS', 'PRC_FULL_PAYMENT'
#
#         'TENURE'], dtype='object')
scores_1 =
[]
range_values = range(1,
20)
for i in
range_values:
    kmeans = KMeans(n_clusters = i)
    kmeans.fit(creditcard_df_scaled)
    scores_1.append(kmeans.inertia_)
plt.plot(scores_1, 'bx-') plt.title('Finding the
right number of clusters') plt.xlabel('Clusters')
plt.ylabel('Scores') plt.show()

```

From this we can observe that, 4th cluster seems to be forming the elbow of the curve.

However, the values does not reduce linearly until 8th cluster.

Let's choose the number of clusters to be 7 or 8.

Apply K-Means Method

```

kmeans = KMeans(7)
kmeans.fit(creditcard_df_scaled)
labels = kmeans.labels_
#cluster_centre is the arithmetic mean of all the points belonging to a cl
uster
kmeans.cluster_centers_.shape
    cluster_centers = pd.DataFrame(data = kmeans.cluster_centers_,
columns = [ creditcard_df.columns])
cluster_centers
# In order to understand what these numbers mean, let's perform inverse tr
ansformation cluster_centers = scaler.inverse_transform(cluster_centers)
cluster_centers = pd.DataFrame(data = cluster_centers, columns = [creditca
rd_df.columns]) cluster_centers

```

First Customers cluster (Transactors): Those are customers who pay least amount of intrerest charges and careful with their money, Cluster with lo west balance (\$104) and cash advance (\$303), Percentage of full payment = 23%

Second customers cluster (revolvers) who use credit card as a loan (most lucrative sector): highest balance (\$5000) and cash advance (~\$5000), low purchase frequency, high cash advance frequency (0.5), high cash advance transactions (16) and low percentage of full payment (3%)

```

# Third customer cluster (VIP/Prime): high credit limit $16K and highest p
percentage of full payment, target for increase credit limit and increase s
pending habits
# Fourth customer cluster (low tenure): these are customers with low tenur
e (7 years), low balance
labels.shape # Labels associated to each data
point labels.max() labels.min() y_kmeans =
kmeans.fit_predict(creditcard_df_scaled) y_kmeans

# concatenate the clusters labels to our original dataframe
creditcard_df_cluster = pd.concat([creditcard_df, pd.DataFrame({'cluster':
labels})], axis = 1) creditcard_df_cluster.head(10)

# Plot the histogram of various clusters for i in creditcard_df.columns:
plt.figure(figsize = (35, 5)) for j in range(7):
plt.subplot(1,7,j+1) cluster =
creditcard_df_cluster[creditcard_df_cluster['cluster'] == j]
cluster[i].hist(bins = 20) plt.title('{} \nCluster {}'.
'.format(i,j))

plt.show()

```

Apply Principal Component Analysis and Visualize the Results

```

# Obtain the principal components pca =
PCA(n_components=2) principal_comp =
pca.fit_transform(creditcard_df_scaled) principal_comp

# Create a dataframe with the two components pca_df =
pd.DataFrame(data = principal_comp, columns = ['pca1', 'pca2'])
pca_df.head()

# Concatenate the clusters labels to the dataframe pca_df =
pd.concat([pca_df, pd.DataFrame({'cluster': labels})], axis = 1)
pca_df.head()
plt.figure(figsize=(10,10)) ax = sns.scatterplot(x="pca1", y="pca2", hue
= "cluster", data = pca_df, palette
=['red', 'green', 'blue', 'pink', 'yellow', 'gray', 'purple']) plt.show()

```