Credit card Segmentation and Recommendation system

Surendra Tanniru
REVA Academy for Corporate Excellence, REVA University
Yelahanka,Bangalore-560064, India
surendraT.BA08@race.reva.edu.in

Ratnakar Pandey
REVA Academy for Corporate Excellence, REVA University
Yelahanka,Bangalore-560064, India
ratnakarpandey@race.reva.edu.in

Shinu Abhi
REVA Academy for Corporate Excellence, REVA University
Yelahanka,Bangalore-560064, India
shinuabhi@reva.edu.in

***Abstract*— Credit cards are an integral part of the growing economy in the Banking, and Financial industry. Credit card utilization adds a prominent value to the banks. Scraping the credit card data will help in identifying interesting patterns and characteristics among different features of the cards that in the future can be used by the banking and financial services to further increase their strength by credit card issuance and acquiring new users. For a user it is difficult to acquire the best card to fulfill most of their needs from the vast availability of credit cards. Thus, recommending a card that suits the customer's needs is important. This further helps banks to identify and analyze different demographic characteristics and payment methods of customers. It also helps to identify potential customers and to implement target marketing. This study is driven to develop and identify the credit card offers and group them into clusters for recommending better card that serves the need of the customer. The data is collected from Banks by scraping their websites. There are 189 credit cards that contain 65 variables. The most important techniques used are Selenium Web driver, Beautiful soup, Principal Component Analysis (PCA), K-means, and K-Nearest Neighbor (KNN). Results indicated how credit cards are grouped based on how different offers and benefits it provides. Moreover, results revealed there are many offers available from banks but not every offer is necessary for analysis and grouping. Further, KNN is applied to identify the closest similar card to provide the credit card recommendation.**

Index Terms- Credit card, web scraping, clustering, recommendation, target marketing

I. INTRODUCTION

India has become one of the fastest growing economies in the world. Access to formal credit is still very minimal compared to many countries. The banking and financial services industry is helping to close this gap by announcing new payment options to bring formal credit closer to the customers. One of their key focuses is credit cards.

The penetration of credit cards in India is very less as compared to various transaction mediums like Unified payments Interface (UPI), debit cards, Point of Sale (POS), internet banking, etc. According to the data on cards published by the Reserve Bank of India (RBI), credit cards show strong momentum on e-commerce platforms. There are 7.36 million credit cards in the country. Credit card spending by users on e-commerce platforms is higher than spending on POS. On the other hand, the transaction count was slightly lower on e-commerce platforms compared to offline transactions. This shows that the average value of payments on online platforms via credit cards was significantly higher.

There are 31 credit card issuers in India. They have issued about 62 million cards so far. As per RBI data released in March 2021, the top 6 issuers account for 81% of the total market share.

More number of Millennials acquiring a credit card and using them aggressively is an important factor for the growth of credit cards in India. As per the study conducted by CRIF, the Millennials holds 14 percent of all new credit cards issued in FY20, and the same stands at just 1.6 percent in FY16. Following the digital upsurge in India, credit cards are being used for utility bills, educational loans, healthcare, insurance, groceries, electronics, clothing and fuel. The pandemic has moved most credit cardholders to online transactions, which also helped in identifying more opportunities to issue more credit cards.

Credit card spending increased 48% year on year to Rs1.07 trillion. There are also increased partnerships between banks and FinTech companies and are issuing cards to them that offer exciting offers and consumer experiences. As per the report by Motilal Oswal (MOFSL) [1], the spending on credit cards boosted because of the increase in number of transactions on e-commerce websites. The vast availability of credit cards from different banks is becoming difficult to acquire the best card to fulfill most of our needs. In this study, we try to analyze the offers and benefits provided by different cards from different banks and will try to provide a better selection of a card that will fulfill most of the customer's needs.

Selecting a card for our needs by creating a better ROI is difficult. Thus, recommending a card that suits the customer's needs by spending less and earning more is important. This will help banks identify and analyze different demographic characteristics and payment methods of customers. It also helps to identify potential customers, and to implement target marketing for specific cards based on their characteristics.

The main goal of this study is to improve ROI by recommending the best credit card.

This study involves the following.

1. Web Scraping
2. PCA for dimensionality reduction
3. K-means to cluster the cards with similar offers and benefits
4. KNN to identify the most similar card for recommendations

WebDriver is used natively with in the browser, either locally or on a remote machine using the Selenium server. It is used as automation to crawl over the websites and scrape the data [2]. Beautiful Soup is used on Hyper Text Markup Language (HTML) queries and Extensible markup Language (XML) queries. It helps to take HTML and XML codes based on tags. The tags are taken based on ID and class, also these are obtained as object, and here we can do several operations. It hovers through the tags with in HTML or XML queries and extract data from HTML, which helps in web scraping [3].

This study is driven to develop and identify the credit card offers and group them into clusters for recommending better card that serves the need of the customer. The data set contains 65 variables and a total of 189 records. The most important techniques that have been used are Selenium Web driver, Beautiful soup, Principal Component Analysis (PCA), and K-means.

II. Literature Review

There are few studies done to identify the credit card offers and to identify the customer behavior based on the usage of cards. Web Services Group, Samsung R&D Institute India done a study [4] on credit card offers provided by online aggregators and recommends the best coupons and offers available.

A study by Wei Li [5] segments the credit card users and recommends a targeted marketing which is based on the real data of a Chinese commercial bank's credit card, the credit card customers are grouped into four classifications by K-means. The analysis is done on the customers transactions rather than offers and benefits provided by the credit cards. The clustering is based on customer's income and consumption habits. A similar study [6] is done by Sarween Zara, on credit card-holder's behavior in order to predict the market segmentation based on their income.

In a study done by Aihua Li [7], that talks about whether a card-holder is a defaulter or not based on his credit card transaction usage. The dataset used is from a major US bank with 65 attributes such as over limit fee, over charge fee and other information etc., Used PCA for dimensionality reduction and MCLP for classifying the card holders into good or bad customer that identify defaulter.

Dimensionality is basically the number of features associated to any data. Features refer to columns in any tabular data. To analyse any dataset, the complexity of the analysis increases with the number of features increasing. Dimensionality reduction is a process that helps in reducing the number of variables. These variables cover most of the variance with in the data, and are called principal variables. According to Wold [8], principal component analysis (PCA) helps in solving problems with more number of dimensions.

For data extraction and featuring process, the clustering algorithms are used frequently. These algorithms help in classifying the data with identical characteristics and group them in to one. There are many different techniques for clustering data, some of which group the data into more than one group. Some of them group the features in a probabilistic way rather than a categorical way. Others group these features together hierarchically. The K-Means algorithm groups the data and classifies them according to distance between two features. Ideally, it checks the distance between the data points from the centre of the group [9].

K-means has different techniques. One being Silhouette score. Silhouette score evaluates the quality of clusters and checks how well the data is grouped with other samples of data which are similar to each other. It is calculated for each and every sample of all the groups. Silhouette score calculates the intra-cluster distance for each sample with other samples within the same cluster. It also evaluates the inter-cluster distance to the sample of one cluster to the sample in the next nearest cluster. The range of silhouette score is [-1, 1]. If the Silhouette score is 1, then the clusters are very close and nicely separated. A value of 0 means, the clusters overlap each other. The negative value indicates that the clusters may not have formed correctly.

III. Methodology

The scope of this study is limited to creating the corpus of credit card data and clustering them based on the offers and benefits provided by each card. Further, the study is extended to recommend the credit card which provides similar benefits to the existing card.

Data is collected by scraping card details on different bank websites. Selenium Web driver and beautiful soup are used to scrape the data from bank official websites. The data is further used to identify the information like offers and benefits and also the terms and conditions of the card.

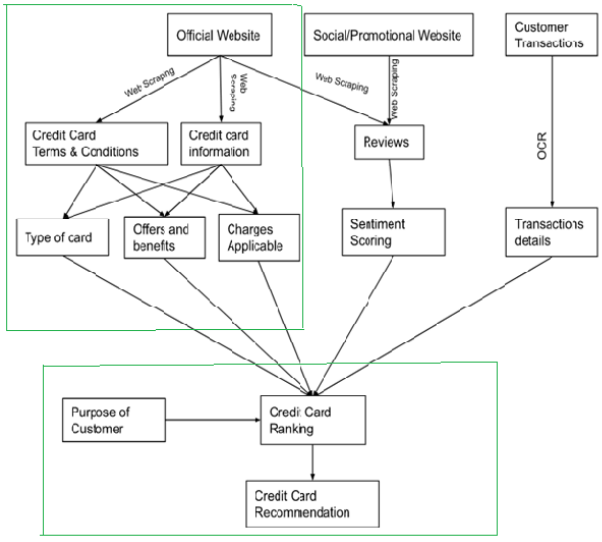

Fig. 1 System Design

PCA will help in dimensionality reductions and gives us the features that provide most variance.

K-means is applied on the data with the features obtained from PCA to cluster them based on the offers they provide.

KNN is used to identify the similar card for recommendation. We used Euclidean distance to identify the similar instances between two cards.

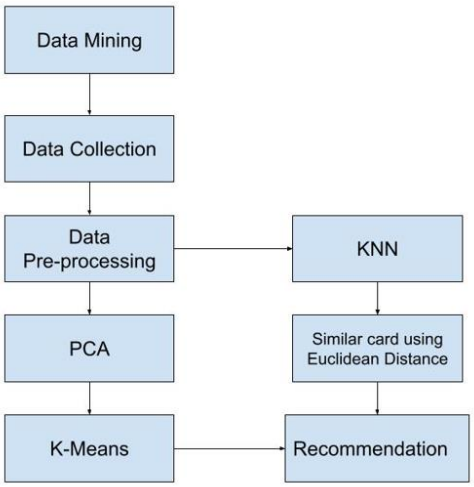

Fig. 2 Data Flow chart

A. *Data Collection:*

Selenium web driver is used to scrape the offers and benefits provided by each card from different bank website. Further the Beautiful Soup is used to parse the HTML and identify the required tags for card names and its offers and benefits. The scraped card details like, card name, Cashback, Fuel, Annual benefits, welcome benefits etc., are added to a dictionary and populated to a dataframe.

B. *Data Pre-processing:*

For further analysis, the data is then converted to numerical form based on the offer. The numerical form is done manually by replacing the textual data with its number available in it which represents the percentage or money provided by that offer. The data set contains 189 unique cards and a total of 65 different offers and benefits like Cashback, Fuel, Annual benefits; welcome benefits etc. provided by 189 cards. These 65 different offers and benefits are represented as features.

Out of 65 different offers and benefits, not every card provides the same offer. Some card provides vast variety of offers while some provides less and it is important feature for that particular card. Principal component Analysis is used to obtain the important features that cover the maximum variance of the data.

IV. Findings and Key Insights

For the pre-processed data, it is very important to reduce the dimensions to have a better analysis.

A. *Principal Component Analysis:*

PCA is used for dimensionality reduction which resulted 90% of variance from 17 Features.

The Fig. 3 represent the variance towards the number of features from PCA. 17 features cover the variance of about 90% in the data.

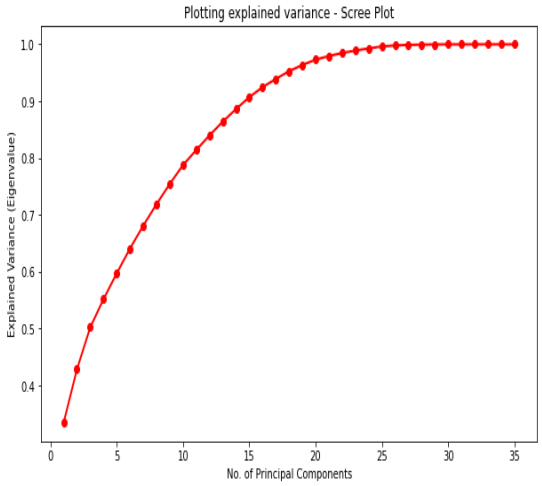

Fig. 3 Principle Component Analysis

Fig. 4 represent the correlation between the features that cover 90% of the variance. The correlation between the features is zero.

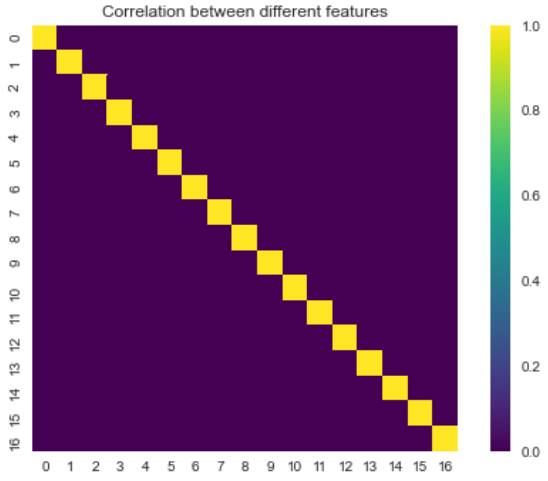

Fig. 4 Correlation of variables after PCA

The features that are obtained from PCA analysis are then used to group the 189 cards into clusters using K-means.

B. *K-Means Clustering:*

After applying PCA on the data, the problem with the huge number of dimensions is reduced. To perform the clustering and decrease group variability within offers and benefits, the k-means algorithm was utilized.

The idea of clustering the data is to define clusters where the total intra-cluster variation is minimal. The inter-cluster distance within the samples of the same cluster should be very less. And, identifying the number of clusters is also

important. Elbow Method is a technique that plots the calculated sum of squares of the distances of the groups up to their respective centres and helps in identifying the number of clusters to be formed. This is also called as the inertia of the clusters.

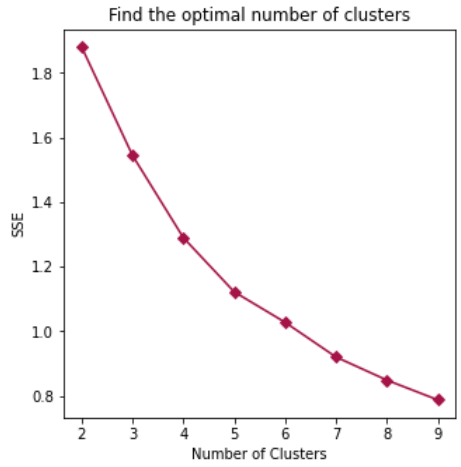

Fig. 5 Elbow Method

Figure 7.3 represents the Elbow method. The change in the curve is called the 'Elbow' which represents the optimal number of clusters to be formed which is 6 in this case.

C. K-Means - Silhouette Method:

Further, the Silhouette method helps in evaluating the quality of clusters where the clusters are grouped well and no overlapping of samples with the other clusters. The Silhouette score is calculated for a different number of clusters to identify the correct number of clusters.

Cluster	Silhouette Score
Cluster 3	0.770
Cluster 4	0.787
Cluster 5	0.798
Cluster 6	0.851
Cluster 7	0.872
Cluster 8	0.881

TABLE I: Silhouette Scores for each cluster

The silhouette score for clusters 7, 8 and 9 looks good, but one of the groups is having a negative range which says the clusters can be incorrect. The only possible group with all positive range of groups is cluster 6. Both, Elbow method and silhouette method indicate 6 as the optimal number of clusters.

Cluster 1: Low Annual Fee Low Cashback

The cards in this cluster provide low cashback and charge less Annual Fees. This Cluster has more than 50% of total cards.

Cluster 2: High Annual Fee High Cashback Medium Lounge

The cards in this cluster provide High cashback along with better Lounge benefits. These cards also charge Higher in Annual Fees.

Cluster 3: Medium Annual Fee Medium Cashback

The cards in this cluster provide medium cashback and also charge fairly to the card holders.

Cluster 4: Medium Annual Fee High Cashback

The cards in this cluster provide High cashback but charges fairly to the card holders. This cluster seems to be better cards with medium Fees but provides a high cashback.

Cluster 5: Zero Annual Fees Medium Cashback Medium Lounge

The cards in this cluster don't charge anything from the cardholders but provide a fair to medium cashback as well as some Lounge benefits.

Cluster 6: Medium Annual Fee Low Cashback High Lounge

The cards in this cluster provide lesser cashback but provide better lounge benefits. These cards have a medium Annual Fee.

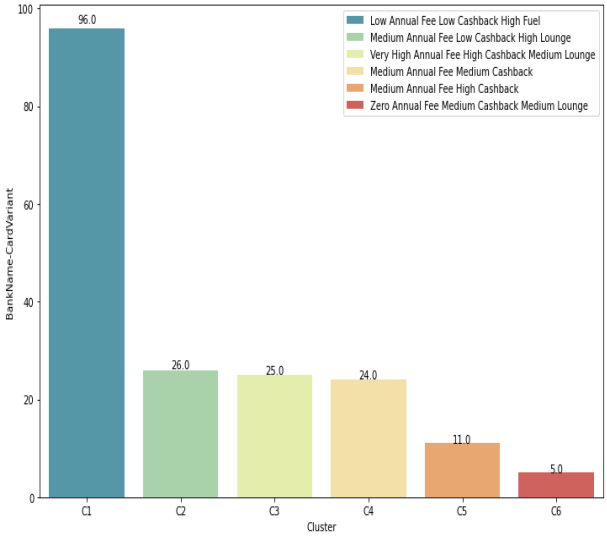

Fig. 6 Credit card Clusters

D. K-Nearest Neighbor (KNN):

In KNN, the features of the training dataset are stored, and the predictions for the test patterns are made by comparing the training dataset with the K most similar instances that are neighbors. The predictions are then combined to identify the most similar instance from these K instances. The Euclidean distance or Manhattan distance is used to determine which instance is most similar to the input test data. Euclidean distance is calculated as the square root of the sum of the squared differences between an existing and new point across all input variables.

The value of k is determined based on the data. The effect of noise on the classification reduces with the increase of value of k. But this will reduce the boundaries between the variables less distinct. The One nearest neighbour is the most intuitive nearest neighbour type classifier. 1-KNN assigns a sample to the sample of its closest neighbour in the feature space and the test dataset sample is grouped to the cluster of its single nearest neighbor.

The results from 1-KNN provides insights of how cards from different banks provide similar offers and also how cards from same bank provide similar offers from its own fleet of cards. This helps in banks to identify the cards and

either discontinue the low-earning cards or promote the cards that get high profits. Similarly, it also helps customers to choose a substitute card from another bank if they are not satisfied with the card they hold.

V. Conclusion

In this study, the credit cards from different banks are grouped to form 6 different clusters wherein cards in each cluster provide same offers and benefits. For analysis, PCA is used for dimensionality reduction which reduced the features which are offers and benefits of various credit cards from 65 to 17. These features are mostly focused on Annual Fee, Cashback, Fuel and Lounge benefits. The offers can be extended to a higher range of benefits to cover different offers provided by various cards. This helps in identifying more clusters and recommend better suitable cards that cater different needs. Further, KNN is used to identify the most similar card for an existing credit card. For this study, the 1-nearest neighbor is used and can be extended to 2 and more neighbors which will help in identifying competitive cards among different financial institutions

References

[1] Aggarwal, N. (2022). Digital Payments Tracker Technology. Mumbai: Motilal Oswal.

[2] Ram Sharan Chaulagain, S. P. (2017). Cloud Based Web Scraping for Big Data. 2017 IEEE International Conference on Smart Cloud, 6.

[3] Zaza, S. (2015). Mining and Exploration of Credit Cards Data in UAE. *2015 Fifth International Conference on e-Learning*

[4] Web Services Group, Samsung R&D Institute India. (2016). Best Offer Recommendation Service. *2016 Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), Sept. 21-24, 2016, Jaipur, India*, 6.

[5] Wei Li, X. W. (2010). Credit Card Customer Segmentation and Target Marketing Based on Data Mining . 2010 International Conference on Computational Intelligence and Security, 4.

[6] Sarween Zaza, M. A.-E. (2015). Mining and Exploration of Credit Cards Data in UAE. *2015 Fifth International Conference on e-Learning*, 5.

[7] Aihua Li, Y. S. (2006). A Data Mining Approach to Classify Credit Cardholders' Behavior. *Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)*, 5.

[8] S. Wold, K. E. (n.d.). Principal component analysis, Chemometrics and intelligent laboratory systems 2 (1-3) (1987) 37–52.

[9] Lorran Santos Rodrigues, M. d. (2022). Application of DEA and Group Analysis using K-means; compliance in the context of the performance evaluation of school networks. Sciencedirect, 10.

[10] Chadha, S. (2021, December). What is fuelling India's credit card splurge despite the rise of UPI payments? Retrieved 2022, from https://timesofindia.indiatimes.com/business/india-business/decoded-in-charts-what-is-fuelling-indias-credit-card-splurge-despite-the-rise-of-upi-payments/articleshow/88612339.cms

[11] Gandhi, M. (2021, August). The credit industry in India. Retrieved 2021 2021, 2021, from The changing landscape of India's credit industry: https://www.pwc.in/industries/financial-services/fintech/dp/the-changing-landscape-of-indias-credit-industry.html

[12] J. MacQueen, e. a. (n.d.). Some methods for classification and analysis of multivariate observations, in: Proceedings of the fifth Berkeley symposium.

[13] Kiran Gajanan Javkar, S. H. (2016). Best Offer Recommendation Service. 2016 Intl. Conference on Advances in computing, Communications and Informatics(ICACCI) (p. 7). Jaipur: IEEE.

[14] Sumit Agarwal, J. C. (2008). Learning in the Credit Card Market. NBER Working Paper No. 13822, 37.