REVA UNIVERSITY
Bengaluru, India

A Project Report on

Credit Card Segmentation and Recommendation System

Submitted in Partial Fulfilment for Award of Degree of

Master of Business Administration

In Business Analytics

Submitted By

Surendra Tanniru

R20MBA09

Under the Guidance of

Ratnakar Pandey

Heading Data science and Analytics
for CX, Amazon

REVA Academy for Corporate Excellence - RACE

REVA University

Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Bengaluru - 560 064

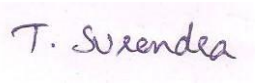
race.reva.edu.in

August, 2022

REVA UNIVERSITY
Bengaluru, India

Candidate's Declaration

I, **Surendra Tanniru** hereby declare that I have completed the project work towards the first year of Master of Business Administration in Business Analytics at, REVA University on the topic entitled **Credit Card Segmentation and Recommendation System** under the supervision of **Ratnakar Pandey.** This report embodies the original work done by me in partial fulfilment of the requirements for the award of degree for the academic year **2022.**



Place: Bengaluru
Date: 20th August, 2022

Name of the Student: Surendra Tanniru
Signature of Student:

REVA UNIVERSITY
Bengaluru, India

Certificate

This is to certify that the project work entitled **Credit Card Segmentation and Recommendation System** carried out by **Surendra Tanniru with SRN R20MBA09,** a bonafide student of REVA University, is submitting the first-year project report in fulfilment for the award of **MBA in Business Analytics** during the academic year 2022. The Project report has been tested for plagiarism and has passed the plagiarism test with a similarity score of less than 15%. The project report has been approved as it satisfies the academic requirements in respect of the project work prescribed for the said degree.

Signature of the Guide
Name of the Guide
Mr. Ratnakar Pandey

Signature of the Director
Name of the Director
Dr. Shinu Abhi

External Viva
Names of the Examiners

1. Harsh Vardhan, Chief Digital Technology Architect, Capgemini
2. Pradeepta Mishra, Director – AI, L&T InfoTech
3. Rajib Bhattacharya, Director – Data & Analytics, Cargill

Place: Bengaluru
Date: 20-August-2022

Acknowledgement

I am highly indebted to **Dr. Shinu Abhi,** Director, and Corporate Training for their guidance and constant supervision as well as for providing necessary information regarding the project and for their support in completing the project.

I would like to thank my project guide **Mr. Ratnakar Pandey** and my team member **Mr. Ravikumar** for the valuable guidance provided to understand the concept and execute this project. It is gratitude towards all other mentors for their valuable guidance and suggestion in learning various data science aspects and for their support. I am thankful for my classmates for their aspiring guidance, invaluable constructive criticism, and friendly advice during the project work.

I would like to acknowledge the support provided by Hon'ble Chancellor, **Dr. P Shyama Raju**, Vice Chancellor **Dr. M. Dhananjaya**, and Registrar **Dr. N. Ramesh**. It is sincere thanks to all members of the program office of RACE who are supportive of all requirements from the program office.

It is my sincere gratitude towards my parents, and my family for their kind co-operation and encouragement which helped me in the completion of this project.

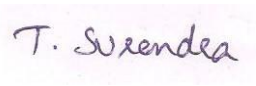
Place: Bengaluru
Date: 20-August-2022

REVA UNIVERSITY
Bengaluru, India

Similarity Index Report

This is to certify that this project report titled **Credit Card Segmentation and Recommendation System** was scanned for similarity detection. Process and outcome are given below.

Software Used: Turnitin
Date of Report Generation: 15th August, 2022
Similarity Index in %: 8%
Total word count: 3886
Name of the Guide: Ratnakar Pandey



Place: Bengaluru
Date: 20-August-2022
Verified by: M N Dincy Dechamma

Name of the Student: Surendra Tanniru
Signature of Student

Signature
Dr. Shinu Abhi,
Director, Corporate Training

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	PCA	Principal Component Analysis
2	KNN	K-Nearest Neighbour

List of Figures

List of Tables

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.

The data is collected from Banks by scraping their websites. Principal component Analysis (PCA) and K-means have been applied to the collected data, 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.

Keywords: Credit card, web scraping, clustering, recommendation, target marketing

Contents

Chapter 1: Introduction

India has become one of the fastest growing economies in the world with a growth rate of 8.2%. Access to formal credit is still very patchy 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.

Traditionally, India has been a debit card market. The penetration of credit cards in India is only 5.55% as compared to various transaction mediums like UPI, debit cards, POS, internet banking, etc. According to the data on cards published by the Reserve Bank of India, 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 INR 68327.73 million, but spending on vending machines (POS) is limited to INR 38773.42 million.

On the other hand, the transaction count was slightly lower on e-commerce platforms (at 11.03 Cr) compared to offline transactions (at 11.34 Cr). 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. (Gandhi, 2021).

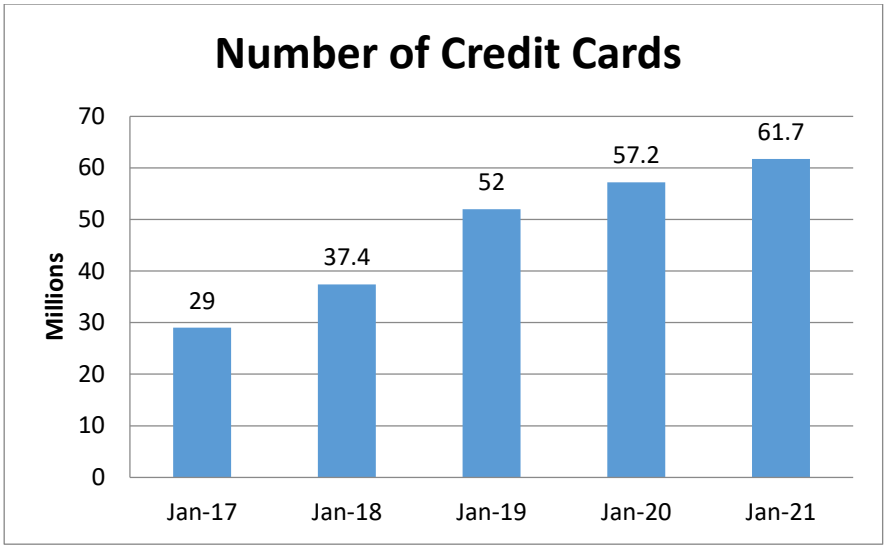


Figure No. 1.1: Credit cards count YOY

There are many key factors driving credit card growth in India. One of them is the increase in acquiring users from rural areas by issuing more cards. A study conducted by TransUnion states that the recent upswing in new credit card issuance is mostly from towns and non-metropolitan areas. It also mentions that non-metros have a growth of 23 percent year on year whereas metros accounted a growth of 10 percent.

More number of Millennials acquiring a credit card and using them aggressively is another 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.

India’s credit-card user base surged at the fastest pace in the past 2 years. Credit card spending increased 48% year on year to Rs1.07 trillion. Following the Indian government's digital push, there are many providers offering users many perks in the name of cashback and coupons while making payments. In order to meet the demand of customers, there are also increased partnerships between banks and FinTech companies and are issuing cards to them that offer exciting offers and consumer experiences (Chadha, 2021).

As per data released by the RBI, the BFSI in India reported a net increase of 1.9m credit cards until March 2022. The total credit card base stood at 73.6 million, up 18.7% year on year (YoY).

Figure No. 1.2: Number of Cards added by top banks

With HDFC having few hiccups in recent times, Axis Bank issued 433,966 cards whereas HDFC issued 263,864 credit cards making Axis bank the largest issuer.

As per the report by analysts at Motilal Oswal (MOFSL), the spending on credit cards boosted because of the increase in number of transactions on e-commerce websites (Aggarwal, 2022).

Monthly spending per card has been increased by 21% which is huge when compared to pre-Covid levels. The average number of transactions per card is increased to 3 from 2.6.

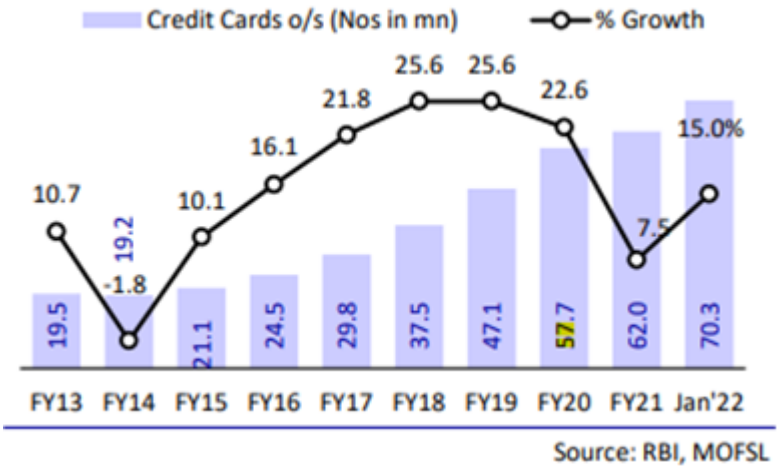


Figure No. 1.3: Outstanding cards grew 15% YoY in Jan’22

The banking system reported net additions of ~1.3m credit cards in Jan’22 (+86% YoY). This has taken the total credit card base to 70.3m (+15.0% YoY– highest in the past 19 months).

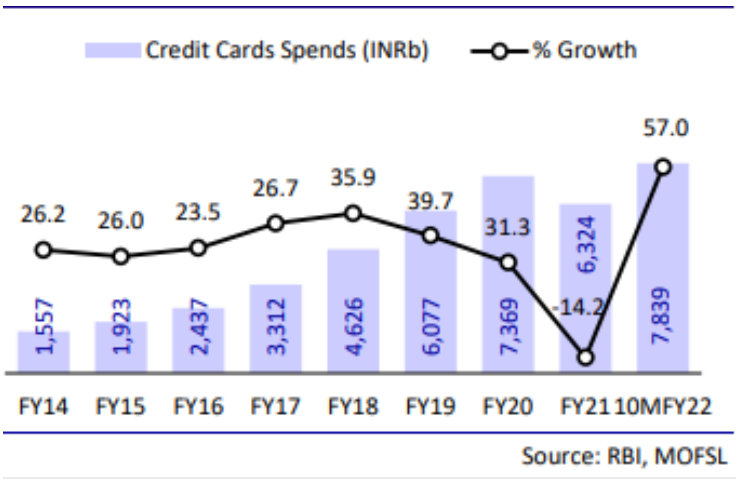


Figure No. 1.4: Credit card spends

Credit card spending increased by 57% over 10MFY22 but they have dropped 7% MoM to INR880b in Jan'22. The same grew strongly by 35% y/y (+14% two-year CAGR).

According to the RBI report, the number of credit card transactions is increasing YOY and the value of these transactions is also increasing accordingly.

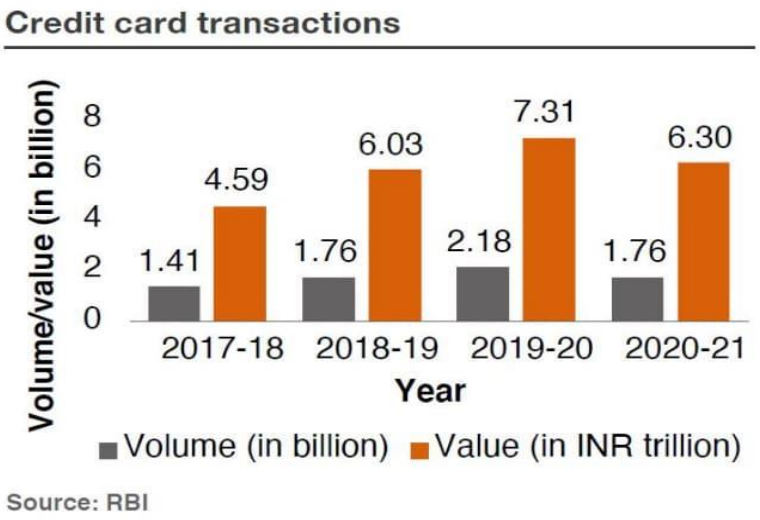


Figure No. 1.5: Volume Vs Value

Even though the credit card transactions in the volume are less compared to debit card and UPI transactions, the value of credit card transactions is high compared to other transactions.

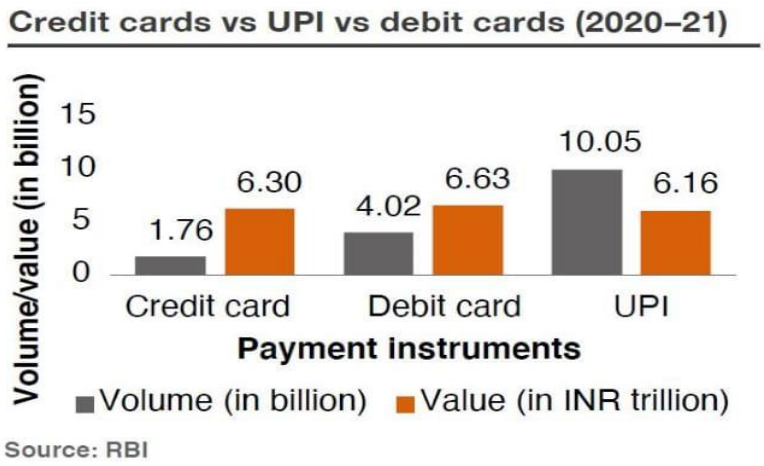


Figure No. 1.6: Volume Vs Value

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 (Wei Li, 2010).

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

Chapter 2: Literature Review

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.

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. (Ram Sharan Chaulagain, 2017). Beautiful Soup is used on HTML and XML queries. It identifies the tags in HTML and XML codes based on ID and class and extracts the data (Zaza, 2015).

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 (Web Services Group, Samsung R&D Institute India, 2016) on credit card offers provided by online aggregators and recommends the best coupons and offers available.

A study by Wei Li (Wei Li, 2010) 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 is done by Sarween Zara (Sarween Zaza, 2015), on credit card-holder's behaviour in order to predict the market segmentation based on their income.

In a study done by Aihua Li (Aihua Li, 2006), 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. This is also called as multivariate analysis. 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 (S. Wold), principal component analysis (PCA) helps in solving problems with more number of dimensions. It applies linear approximation on the data table which will be the product between 2 matrices. This product of matrix will contribute to the variance of the dataset.

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. The K-Means algorithm groups the data and classifies them according to distance between two features. (J. MacQueen). Ideally, it checks the distance between the data points from the centre of the group (Lorran Santos Rodrigues, 2022)

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.

K-Nearest Neighbour is a supervised learning method that classifies the data onto different categories based on the training dataset. It takes all the features in the training data and classifies them into different groups and categories based on similarity of these features in the data. KNN makes no assumptions about the underlying data distribution, which makes it a non-parametric technique. There is no fixed number of parameters and these parameters grow along with the data. (Kiran Gajanan Javkar, 2016).

Chapter 3: Problem Statement

There are number of credit cards available from different banks and is becoming difficult to acquire a best card to fulfill most of our needs. One has to consider many factors while selecting a credit card which varies from cashback to different offers provided. User may also need to look at discounts when doing any payments and the fees associated to it. The main factor that needs to consider while choosing a credit card are the charges and annual fees which comes under terms & conditions. One has to check the services it offers and the credit limit. (Sumit Agarwal, 2008). There is huge number of credit cards available in the market which caters different needs of the customers. Customers trying to avail a card find it difficult to select the perfect credit card for their use.

Selecting a card for our needs by generating a better ROI is difficult. Thus, recommending a card that suits the customer needs by spending less and gaining more is important.

In this project, ***machine learning techniques like PCA, K-means are applied on the scraped credit card data from bank websites and cluster the cards based on different offers and benefits they provide. Further, KNN is applied to recommend the card that provides similar offers*** which will fulfill most of the customer needs.

Chapter 4: Objectives of the Study

The scope of this project is to analyze different credit cards that are scraped from bank websites and apply machine learning techniques to identify the most common offers and benefits and group the cards in to clusters based on the similar offers and benefits they provide.

In order to analyze the cards data, we have to create a corpus of credit card data by Web scraping (using Selenium web driver) the Offers and Benefits of the credit card from official websites and summarize the Offers and Benefits based on different categories/needs and then build a Credit card recommendation system.

The Objectives of the study are,

- Create a corpus of data by scraping the credit card information including the offers, benefits, and terms and conditions from bank official websites.

- Group the cards in to clusters based on the similar offers and benefits they provide.

- A recommendation engine to recommend similar credit cards based on the credit card categories.

- Simple Chat Bot, which suggests the best suitable credit card based on the user questions.

Chapter 5: Project 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 one which we hold or want to compare.

Figure No. 5.1 represents the larger scope of the study, but the scope of this is limited to web scraping of the bank websites and clustering them. Further, the clusters are used to recommend the most similar card for the recommender system.

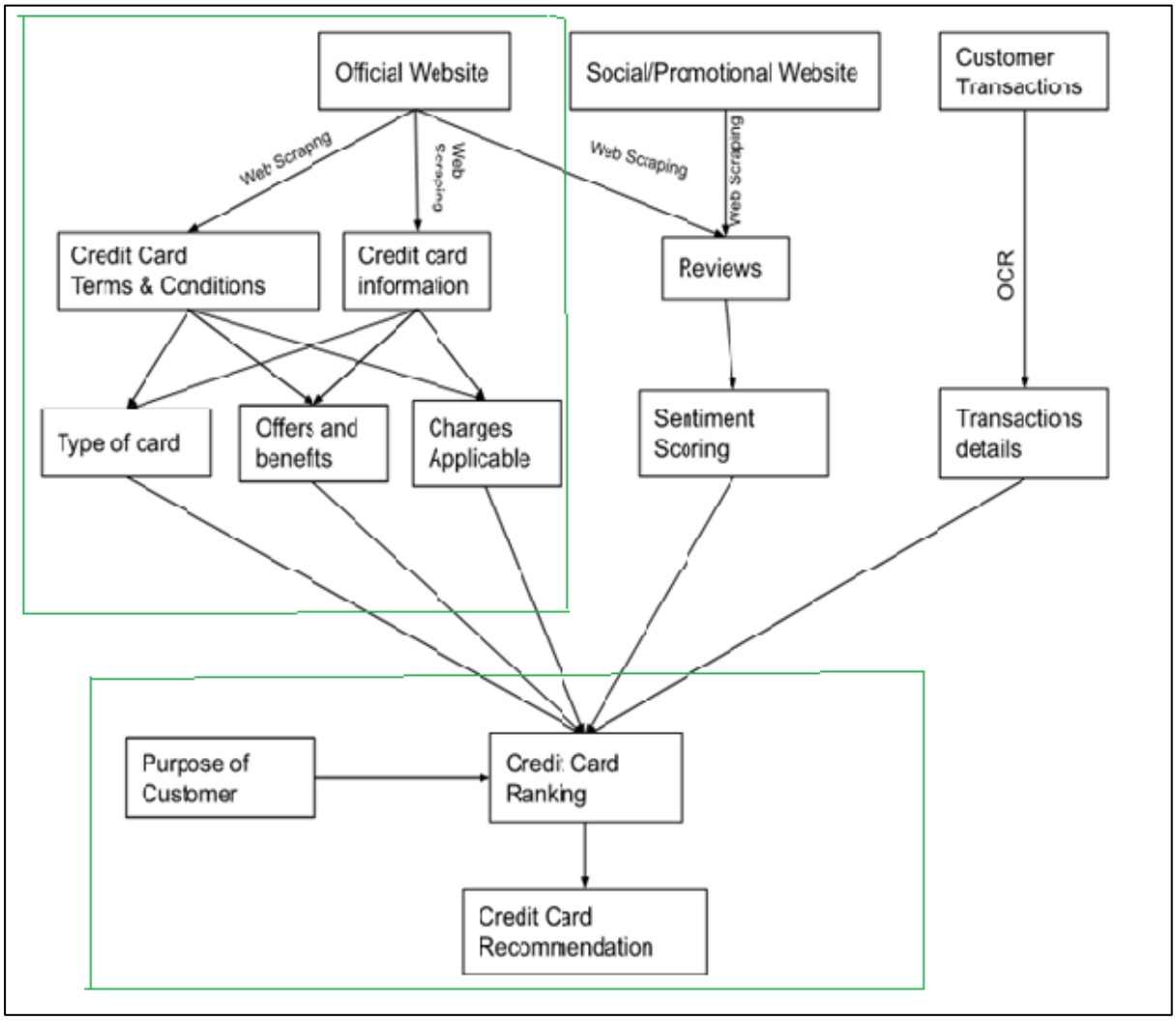


Figure No. 5.1: System Design

Figure No. 5.2 represents the Data flow diagram of the project.

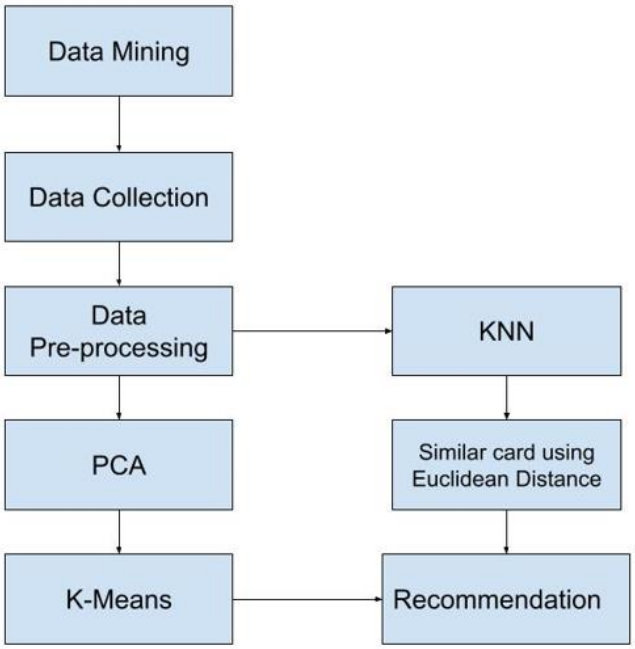


Figure No. 5.2: Data Flow chart

Data is collected by scraping card details on different bank websites. Since there are different kinds of offers and benefits provided by each card, it is not an easy task to analyze every offer they provide.

Principal component Analysis (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.

K-Nearest Neighbor (KNN) is used to identify the similar card for recommendation. We used Euclidean distance to identify the similar instances between two cards.

Chapter 6: Data Collection and Pre-processing

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. Figure No. 6.1 represents the scraped Credit cards data

CardVariant	WelcomeBenefits	AnnualBenefits	AirportLoungeBenefits	Cashback
Flipkart Axis Bank Credit Card	Rs. 1000 worth of joining and activation benefits on your Flipkart Axis Bank		Enjoy 4 complimentary lounge visit	5% cashback on Flipkart
Axis Bank Privilege Credit Card	Get 12500 EDGE REWARD Points, re	Annual fees reversal on achieving spe	2 Complimentary access per calendar quarter to select dome	
AXIS Bank SELECT Credit Card	Get Amazon voucher worth Rs 2000	Priority Pass Membership renewal on	Get complimentary Priority Pass m	Earn 10 Axis EDGE points
AXIS Bank MY ZONE Credit Card			Enjoy 1 complimentary access to select airport lounges withi	
AXIS Bank Magnus Credit Card	Choose between one complimentary	Annual fee of Rs 10,000 + Taxes waive	Enjoy 8 complimentary end-to-end VIP services at the airport	
IndianOil Axis Bank Credit Card	Earn 100% cashback up to INR 250 c	Spend more than INR 50,000 in a year and you will be eligible for annual fee waiver		

Figure No. 6.1: Scraped Credit cards data

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.

Data Pre-processing:

For further analysis, the data is then converted to numerical form based on the offer. Figure No. 6.2 shows the pre-processed data.

BankName-CardVariant	AnnualFee2ndYear	MinimumSpend	CashWithdrawalFe	JoiningFee	AnnualMinSpend	LoungeBenefits	FuelSurcharge	FuelDiscount/C
HDFC-6E Rewards XL-IndiGo HDFC Bank	2500	0	2.5	1	0	0	0	0
HDFC-6E Rewards-IndiGo HDFC Bank	700	0	2.5	1	0	0	0	0
ICICI-Accelero ICICI Bank Credit Card	499	125000	0	0	1	1	1	2.5
SBI-Air India SBI Platinum Card	1499	0	2.5	1	0	8	1	0
SBI-Air India SBI Signature Card	4999	0	2.5	1	0	0	1	0
SBI-Allahabad Bank SBI Card ELITE	4999	0	2.5	1	0	8	1	0
SBI-Allahabad Bank SBI Card PRIME	2999	0	2.5	1	0	8	0	0
SBI-Allahabad Bank SimplySAVE SBI Car	499	100000	2.5	0	1	0	1	0

Figure No. 6.2: Pre-processed Credit cards data

There are 189 cards and 65 different offers and benefits, but 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.

Chapter 7: Data Modelling and Analysis

PCA is used for dimensionality reduction and after applying PCA, it provided the 90% of variance from 17 Features. The Figure No. 7.1 represents the variance towards the number of features from PCA. 17 features cover the variance of about 90% in the data.

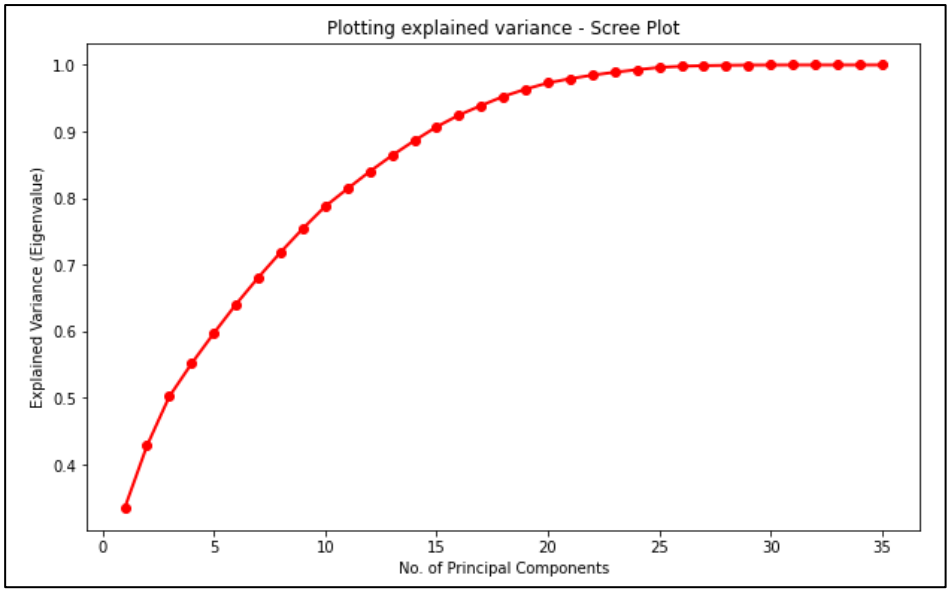


Figure No. 7.1: Principle Component Analysis

Figure No. 7.1 represents the correlation between the features that cover 90% of the variance. The correlation between the features is zero.

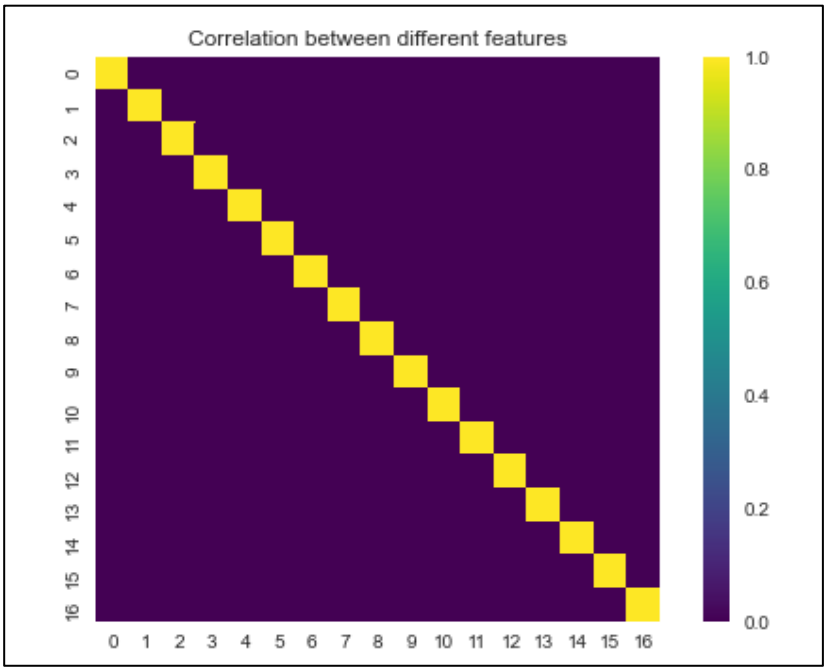


Figure No. 7.2: Correlation of variables after PCA

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

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. Below Figure No. 7.3, represents the optimal number of clusters after applying Elbow method.

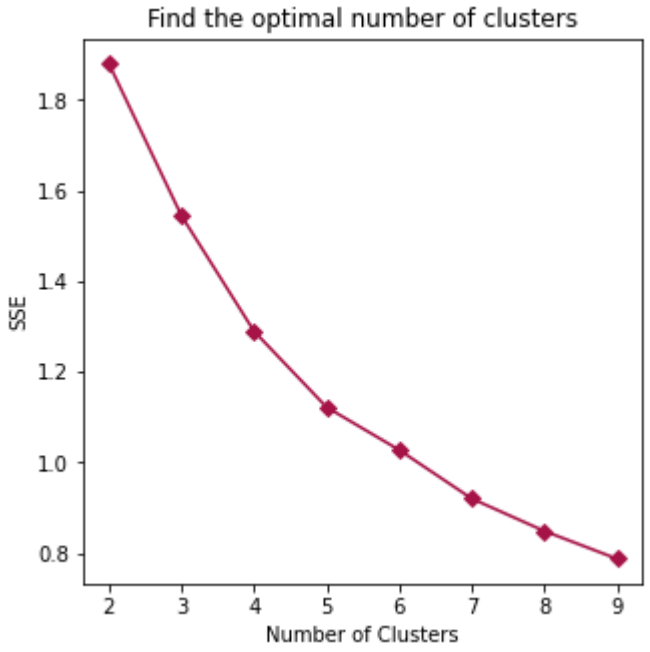
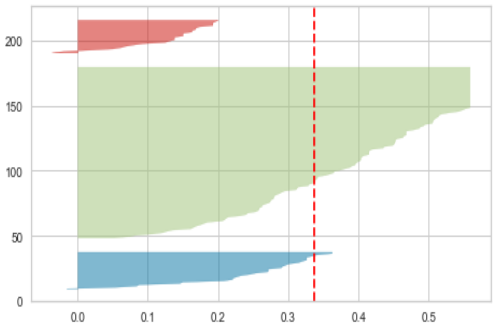


Figure No. 7.3: Elbow Method

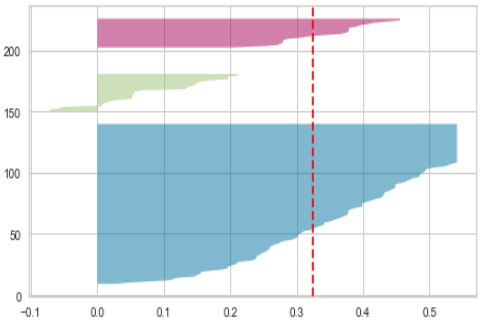
From the above Figure, the change in the curve is called the ‘Elbow’ which represents the optimal number of clusters to be formed which is 6 in this case.

K-Means - Silhouette Method:

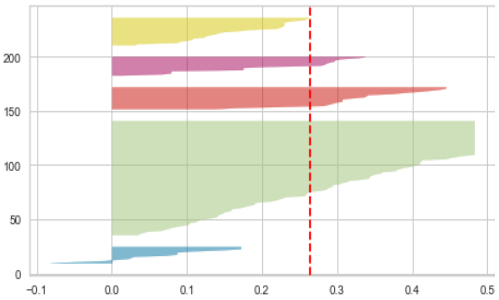
Further, the Silhouette method helps in evaluating the quality of clusters where the clusters are grouped well and no overlapping of samples with the other clusters. The Silhouette score is calculated for a different number of clusters to identify the correct number of clusters.



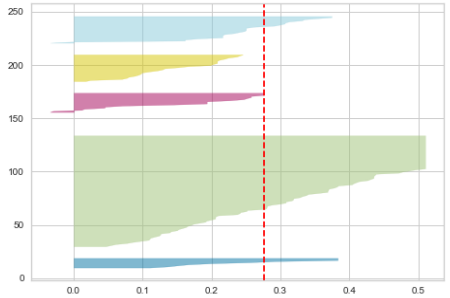
Clusters = 3



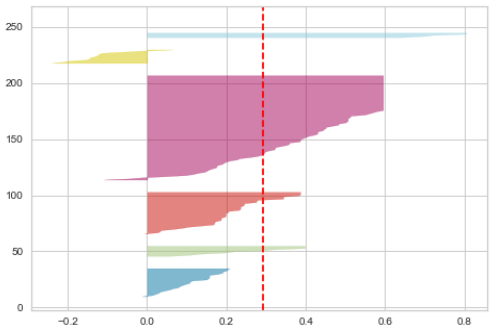
Clusters =4



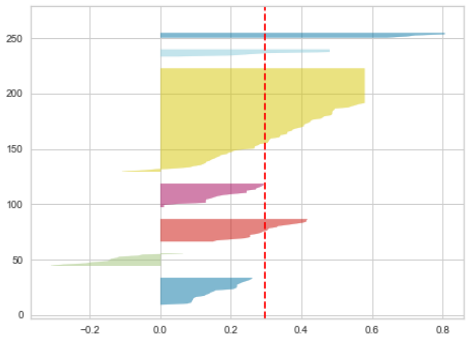
Clusters = 5



Clusters =6



Clusters =7



Clusters = 8

Figure No. 7.4: Silhouette K-means cluster representation

Table No. 1 represents the Silhouette Scores for each cluster

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
Cluster 9	0.880

Table No. 7.1: Silhouette Scores

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.

Clustering:

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. The Below Figure No. 7.5 represents the different Clusters.

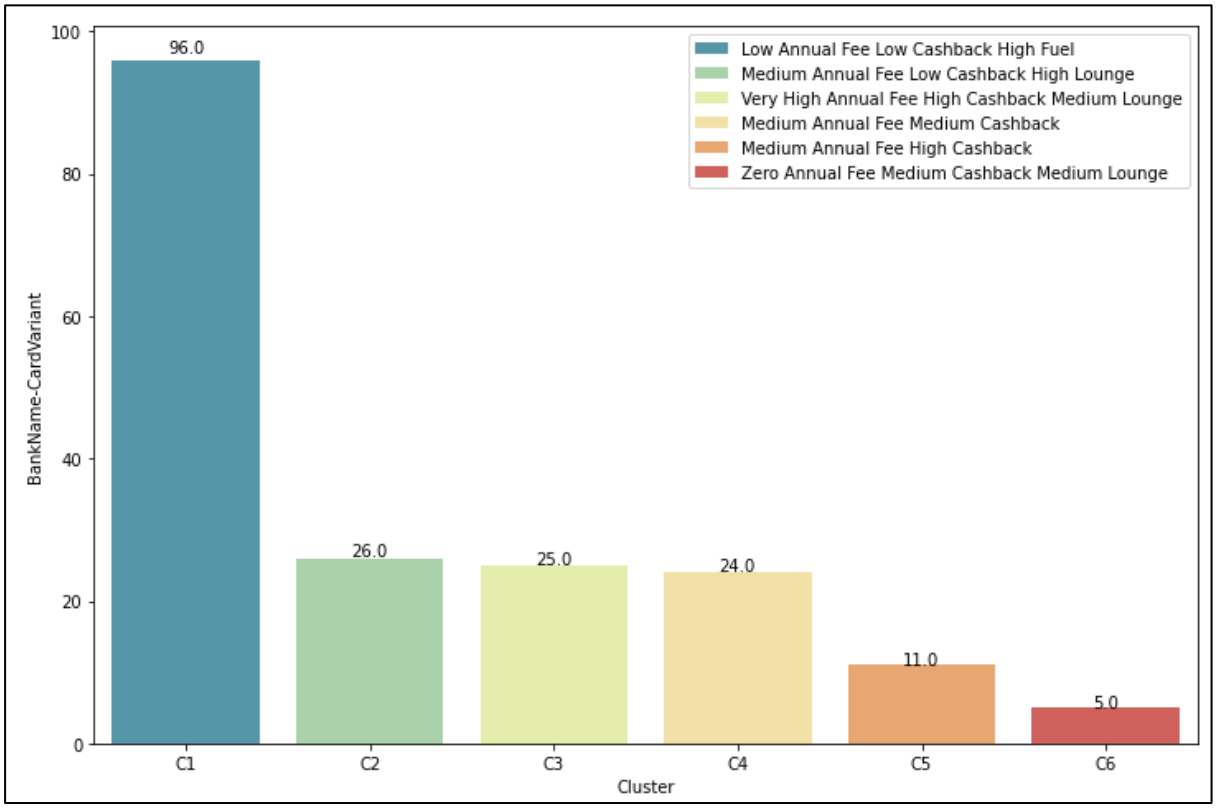


Figure No. 7.5: Credit card Clusters

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.

$$EuclideanDistance(x, xi) = sqrt(sum((xj - xij)^2))$$

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 (i.e. when k = 1), that is

$$Cn^{1}nn(x) = Y(1)$$

For our initial analysis, we have considered k=1. If k = 1, then the test dataset sample will be grouped to the cluster of its single nearest neighbor.

Below Figure No. 7.6 represents the 1-nearest neighbor for each card using KNN.

	idx	Bank-card	pred_id	pred_Bank_Card
0	0	HDFC-6E Rewards XL-IndiGo HDFC Bank	24	SBI-BPCL SBI Card
1	1	HDFC-6E Rewards-IndiGo HDFC Bank	55	SBI-FABINDIA SBI CARD SELECT
2	2	ICICI-Accelero ICICI Bank Credit Card	59	ICICI-ICICI Bank Coral American Express Credit...
3	3	SBI-Air India SBI Platinum Card	64	ICICI-ICICI Bank Rubyx Credit Card
4	4	SBI-Air India SBI Signature Card	40	SBI-Club Vistara SBI Card PRIME
...	...	...	...	...
182	182	HDFC - Regalia ForexPlus Card	167	HDFC - MakeMyTrip Hdfc Bank ForexPlus Card
183	183	HDFC - Reward Card	167	HDFC - MakeMyTrip Hdfc Bank ForexPlus Card
184	184	HDFC - Titanium Edge Credit Card	176	HDFC - Multicurrency Platinum ForexPlus Chip Card
185	185	HDFC - Visa Signature Credit Card	179	HDFC - Platinum Plus Credit Card
186	186	HDFC - World MasterCard Credit Card	176	HDFC - Multicurrency Platinum ForexPlus Chip Card

Figure No. 7.6: 1-nearest neighbor for each card using KNN

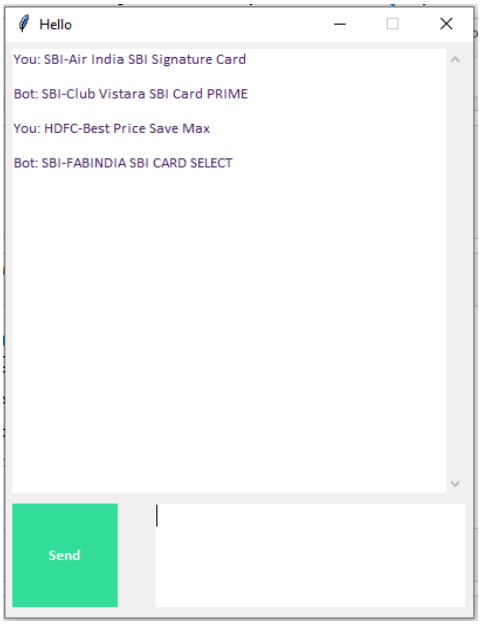
From Figure No. 7.6, it is observed that there are cards from SBI that provides almost similar offers of the HDFC card. Similarly, there are ICICI card that provides similar offers to SBI card. This tells us the customers can have a substitute card from another bank if they are not satisfied with the card they hold.

This also interprets that there are cards from HDFC that provide similar offers from its own fleet of cards. So, this helps for banks to either discontinue the low-earning cards or promote the cards that get high profits.

Chat-bot for recommending credit card:

Below figure represents a simple chatbot created to recommend a similar card based on the user input. It takes the card name as input and recommends the most similar card from the group of cards from each cluster.

This recommendation is based on the 1-nearest neighbor for each card using KNN.



Chapter 8: Conclusions and Recommendations for future work

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. Thus, The Clusters are formed as Low Annual Fee Low Cashback, High Annual Fee High Cashback Medium Lounge, Medium Annual Fee Medium Cashback, Medium Annual Fee High Cashback, Zero Annual Fee Medium Cashback Medium Lounge, and Medium Annual Fee Low Cashback High Lounge.

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.

Bibliography

Aggarwal, N. (2022). *Digital Payments Tracker Technology.* Mumbai: Motilal Oswal.

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.

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

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

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

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.

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.

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

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

Sarween Zaza. (2015). Mining and Exploration of Credit Cards Data in UAE . *Fifth International Conference on eLearning, 5*, 6.

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

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

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.

Zaza, S. (2015). Mining and Exploration of Credit Cards Data in UAE. *2015 Fifth International Conference on e-Learning*, 5.

Appendix

Plagiarism Report[1]

Credit Card Segmentation and Recommendation System

by Surendra Tanniru

Submission date: 15-Aug-2022 09:38PM (UTC+0530)
Submission ID: 1882841404
File name: Credit_Card_Segmentation_and_Recommendation_System-Surendra.docx (619.55K)
Word count: 3886
Character count: 20235

[1] Turnitn report to be attached from the University.

Credit Card Segmentation and Recommendation System

ORIGINALITY REPORT

8%	6%	2%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	global.oup.com Internet Source	2%
2	www.financialexpress.com Internet Source	1%
3	eprints.maynoothuniversity.ie Internet Source	1%
4	Lorran Santos Rodrigues, Marcos dos Santos, Claudio de Souza Rocha Junior. "Application of DEA and Group Analysis using K-means; compliance in the context of the performance evaluation of school networks", Procedia Computer Science, 2022 Publication	1%
5	Submitted to University College London Student Paper	1%
6	Submitted to University of Nottingham Student Paper	<1%
7	www.coursehero.com Internet Source	<1%

8	Submitted to Georgia Institute of Technology Main Campus Student Paper	<1 %
9	www.communitylawcentre.org.za Internet Source	<1 %
10	rickreagan.com Internet Source	<1 %
11	en.wikipedia.org Internet Source	<1 %
12	onlineasdma.assam.gov.in Internet Source	<1 %
13	link.springer.com Internet Source	<1 %
14	iiespace.iie.ac.za Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		

Publications in a Journal/Conference Presented/White Paper[2]

Paper submitted:

Surendra Tanniru, *Ratnakar Pandey, Shinu Abhi*, "Credit Card Segmentation and Recommendation System" 9th International Conference on Business Analytics and Intelligence, IIMB.

Paper id: 2672

Submission Date: 21st October 2022.

[2] URL of the white paper/Paper published in a Journal/Paper presented in a Conference/Certificates to be provided.

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.

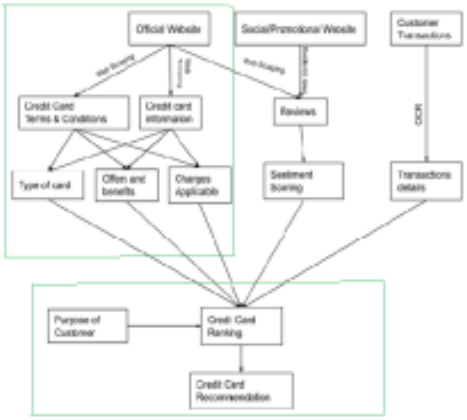


Fig. 1 System Design

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.

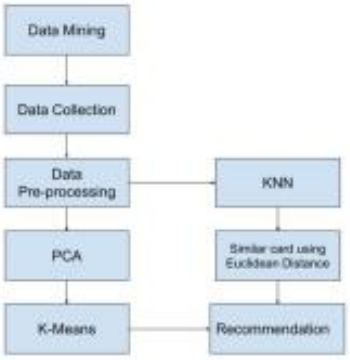


Fig. 2 Data Flow chart

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.

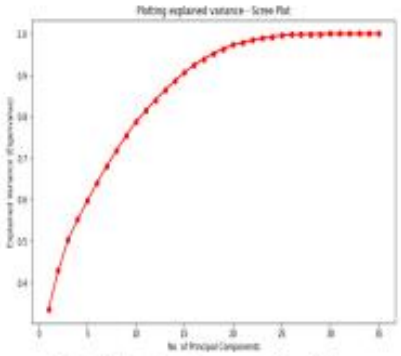


Fig. 3 Principle Component Analysis

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

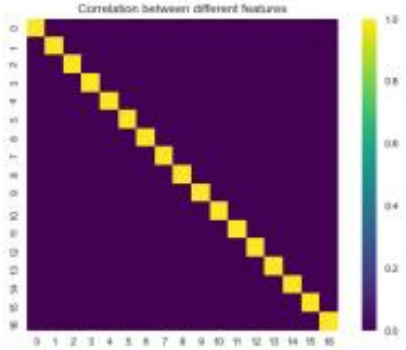


Fig. 4 Correlation of variables after PCA

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.

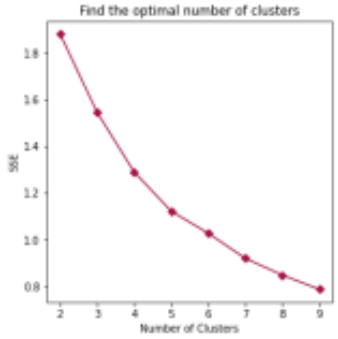


Fig. 5 Elbow Method

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.

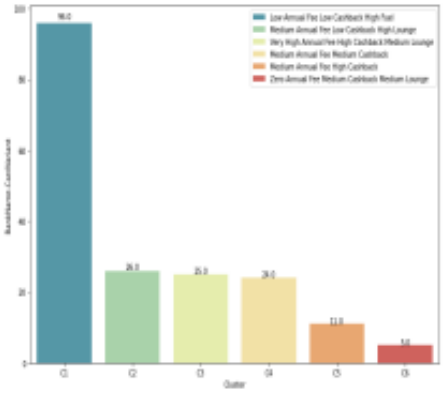


Fig. 6 Credit card Clusters

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.

Any Additional Details