PROJECT REPORT

CREDIT CARD SEGMENTATION

RISHABH TEWARI

26 JAN 2020

Chapter 1

Introduction

1.1 Problem Statement

The Bank share it's Dataset of Credit card Holders .We have to develop a customer segmentation to define market Strategy.

we divide the customer data into various segments and suggest the bank with various strategies.

1.2 Data

There are 18 variables in our dataset and all are independent variables . Since there is no target variable or dependent variable in our dataset and all our variable are continuous ,this is a problem of unsupervised learning, the dataset consist of 9000 observations and missing value are present.

Variables Information:

Var. #	Variable Name	Description
1.	CUST_ID	Credit card holder ID
2.	BALANCE	Monthly average balance
3.	BALANCE_FREQUENCY	Ratio of last 12 months with balance
4.	PURCHASES	Total purchase amount spent during last 12 months
5.	ONEOFF_PURCHASES	Total amount of one-off purchases

6.	INSTALLMENTS_PURCHASES	Total amount of installment purchases
7.	CASH_ADVANCE	Total cash-advance amount
8.	PURCHASES_ FREQUENCY	Frequency of purchases (% of months with at least on purchase)
9.	ONEOFF_PURCHASES_FREQUENCY	Frequency of one-off-purchases
10.	PURCHASES_INSTALLMENTS_FREQUENCY	Frequency of installment
11.	CASH_ADVANCE_ FREQUENCY	Cash-Advance frequency
12.	AVERAGE_PURCHASE_TRX	Average amount per purchase transaction
13.	CASH_ADVANCE_TRX	Average amount per cash-advance transaction
14.	PURCHASES_TRX	Average amount per purchase transaction
15.	CREDIT_LIMIT	Credit limit
16.	PAYMENTS-	Total payments (due amount paid by the customer to decrease their statement balance) in the period
17	MINIMUM_PAYMENTS	Total minimum payments due in the period.
18.	PRC_FULL_PAYMENT-	Percentage of months with statement balance
19.	TENURE	Number of months as a customer

1.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics. In the given data set there are 18 variables and data types is numeric. There are 9000 observations and 18 columns in our data set. Missing value is also present in our data.

From EDA we have concluded that there are 18 continuous variable and 2 categorical variable in nature.

Chapter 2

DATA PREPROCESSING

Before feeding the data to the model we need to clean the data and convert it to a proper format. It is the most crucial part of data science project we spend almost 80% of time in it.

2.1 Missing Value Analysis

In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. If a column has more than 30% of data as missing value either we ignore the entire column or we ignore those observations

In our Data set there are missing value in two variables .

	VARIABLE	MISSING PERCENTAGE
16	MINIMUM_PAYMENTS	3.49720670
14	CREDIT_LIMIT	0.01117318

1	CUST_ID	0.00000000
2	BALANCE	0.00000000
3	BALANCE_FREQUENCY	0.00000000
4	PURCHASES	0.00000000
5	ONEOFF_PURCHASES	0.00000000
6	INSTALLMENTS_PURCHASES	0.00000000
7	CASH_ADVANCE	0.00000000
8	PURCHASES_FREQUENCY	0.00000000
9	ONEOFF_PURCHASES_FREQUENCY	0.00000000
10	PURCHASES_INSTALLMENTS_FREQUENCY	0.00000000
11	CASH_ADVANCE_FREQUENCY	0.00000000
12	CASH_ADVANCE_TRX	0.00000000

13	PURCHASES_TRX	0.00000000
15	PAYMENTS	0.00000000
17	PRC_FULL_PAYMENT	0.00000000
18	TENURE	0.00000000

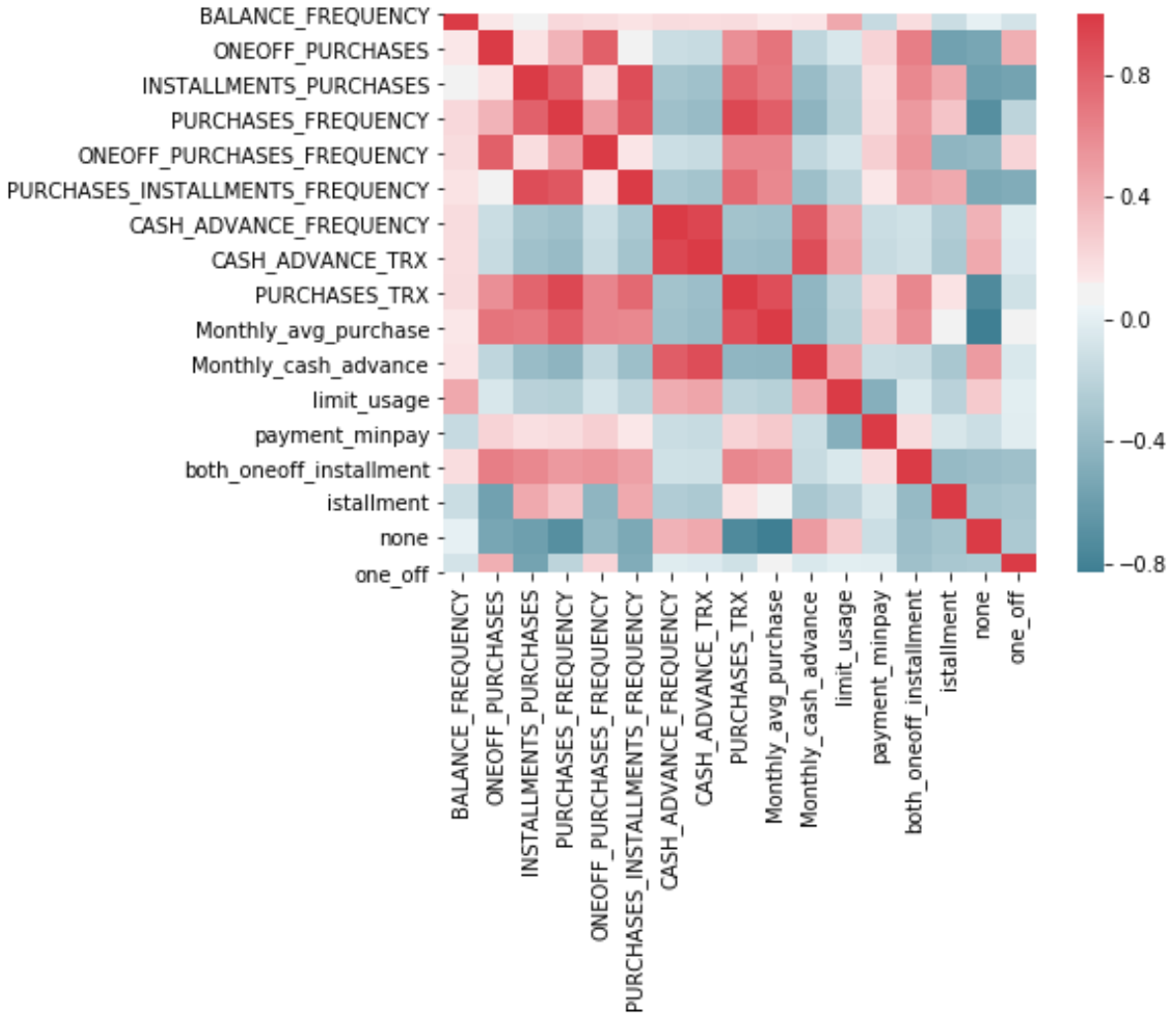
We will use different method like central tendency(mean ,median(and KNN imputation to check which method is best suited.

After applying all the above method median is best suited . so we fill all the values using median imputation method.

2.2Feature Selection

Before performing any type of modeling we need to assess the importance of each predictor variable in our analysis. There is a possibility that many variables in our analysis are not important at all to the problem of class prediction.Selecting subset of relevant columns for the model construction is known as Feature Selection. We cannot use all the features because some features may be carrying the same information or irrelevant information which can increase overhead. To reduce overhead we adopt feature selection technique to extract meaningful features out of data. This in turn helps us to avoid the

problem of multi collinearity. In this project we have selected **Correlation Analysis** for numerical variable and **CHI SQUARE TEST** for categorical variable.



It can be clearly seen that no independent variable is highly correlated with any other independent variable.

As there is no categorical variable in our dataset so no need to apply chi square test.

Chapter 3

NEW KPI

Key Performance Indicator

A **Key Performance Indicator** is a measurable value that demonstrates how effectively a company is achieving key business objectives. Organizations use KPIs at multiple levels to evaluate their success at reaching targets

1. **MONTHLY AVERAGE PURCHASE:**
2. **MONTHLY CASH ADVANCE**
3. **LIMIT USAGE**
4. **PAYMENT TO MIN PAYMENT RATIO.**

MONTHLY AVERAGE PURCHASE can be defined as the ratio of Purchase and tenure.

MONTHLY CASH ADVANCE can be defined as the ratio of CASH ADVANCE and TENURE

LIMIT USAGE can be defined as the ratio of BALANCE and CREDIT LIMIT.

PAYMENT to MIN PAYMENT RATIO can be defined as ratio of PAYMENTS and MINIMUM PAYMENTS.

we found out that there are 4 types of purchase behaviour in the data set. So we need to derive a categorical variable based on their behaviour¶

WE find the relation between ONE OFF PURCHASE and INSTALLMENT PURCHASE

we create a new variable in our data set and named it ***Purchase type***

CASE 1 if ONE off purchase and INSTALLMENTS purchase equals to zero we put it in NONE categorical variable.

CASE 2 if ONE off purchase greater than zero and INSTALLMENTS purchase equals to zero we put it in ONE OFF categorical variable

CASE 3 if ONE off purchase and INSTALLMENTS purchase greater than zero we put it in BOTH categorical variable.

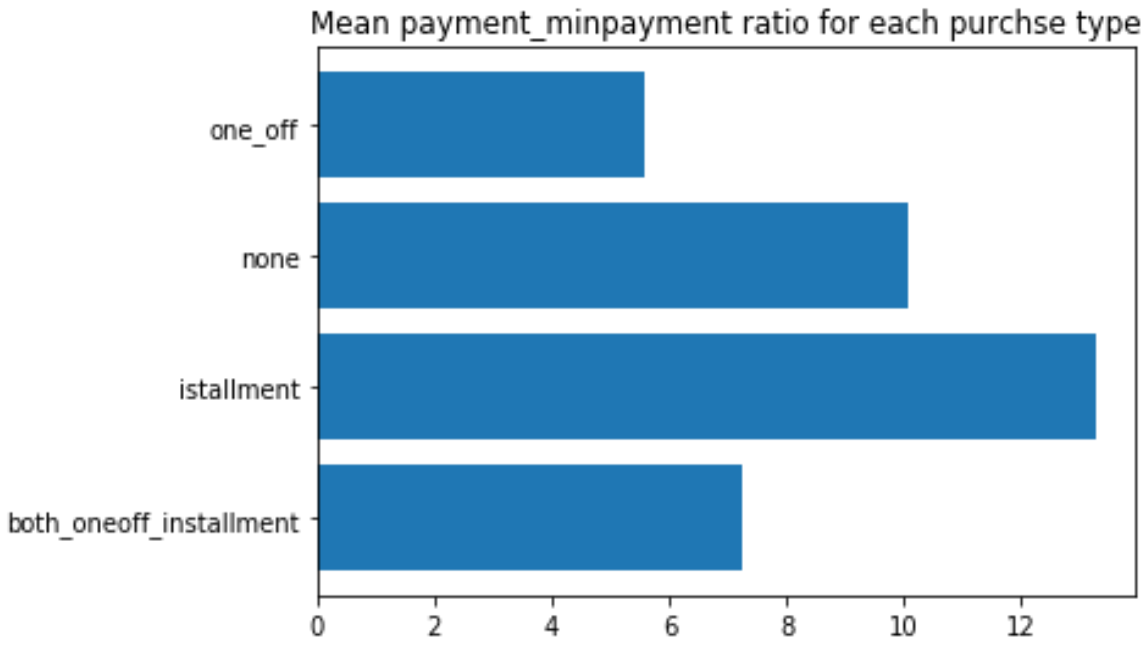
CASE 2 if ONE off purchase greater than zero and INSTALLMENTS purchase equals to zero we put it in ONE OFF categorical variable

CATEGORICAL VARIABLE	COUNT
both_oneoff_installment	2774
installment	2260
none	2042

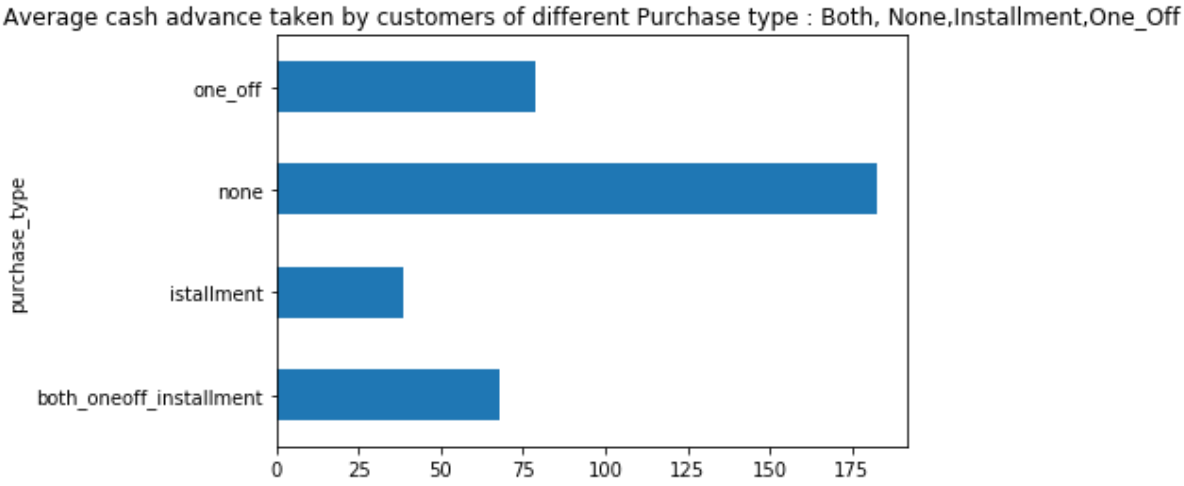
one_off 1874

Insights from KPIs

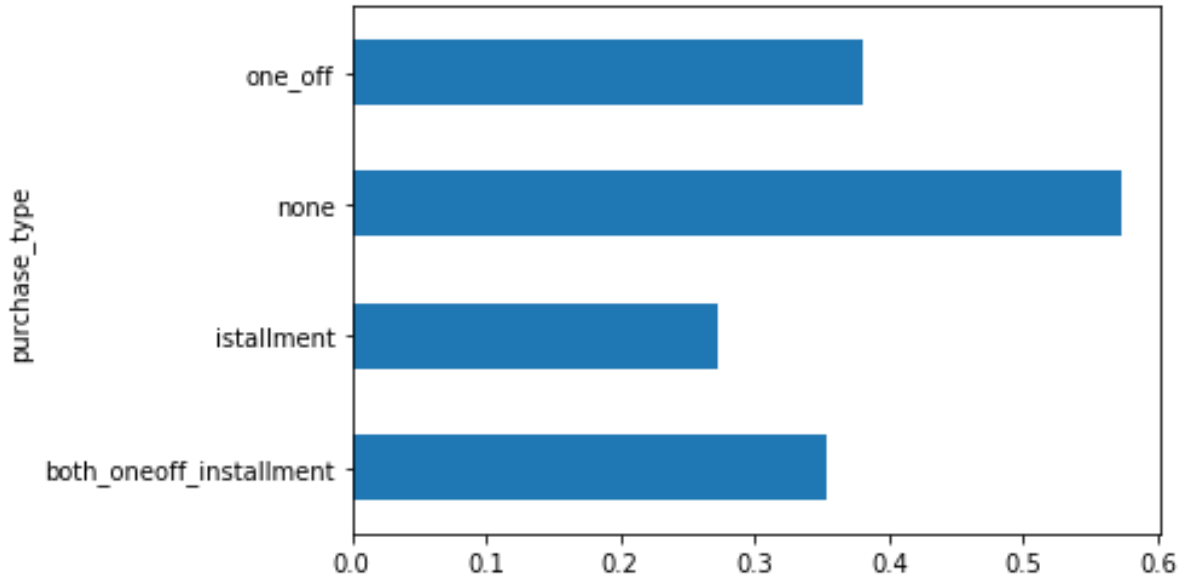
1.Average payment_minpayment ratio for each purchse type.



2. Average cash advance taken by customers of different Purchase type



3. LIMIT USAGE OF CUSTOMER OF DIFFERENT PURCHASE TYPE



PREPARING MACHINE LEARNING ALGORITHM

CLUSTER ANALYSIS

In general terms, Clustering can be termed as the process of breaking down a large population or data-set into smaller groups.

Basically, Clustering in ML allows you to break a population into smaller groups where each observation within every group is more similar to each other than it is to an observation of another group. So, the idea is to group together similar kind of observations into smaller groups and thus break down the large heterogeneous population into smaller homogenous groups.

1. K MEANS CLUSTERING

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.

Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes.

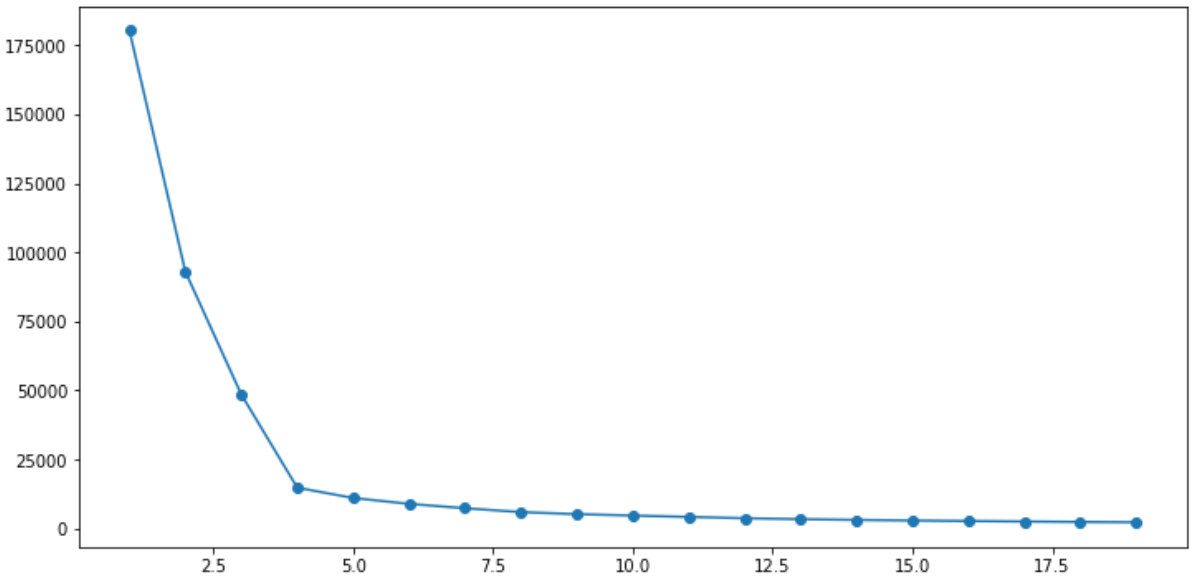
the K-means algorithm identifies *k* number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

ELBOW METHOD

In k means clustering we use Elbow method to define the number of cluster

The **K-Elbow** Visualizer implements the "**elbow**" **method** of selecting the optimal number of **clusters** for **K-means clustering**. ... The **elbow method** runs **k-means**

clustering on the dataset for a range of values for **k** and then for each value of **k** computes an average score for all **clusters**.



In our dataset we use k value=3

because there is a curve at k=3 after which the graph became almost constant

so we apply K clustering algorithm for k =3.

CLUSTER	0	1	2
both_oneoff_installment	2707	3	64
installment	0	49	2211
none	0	2042	0
one_off	1303	571	0

INTERPRETETION:

Cluster 0 is doing maximum one off Transaction and installment transaction. using card for just oneoff transactions (may be for utility bills only). This group seems to be risky group.

CLUSTER 2: This group is performing best among all as cutomers as most of the customers are paying on installments and paying dues on time. -- Giving rewards point will make them perform more purchases.

CLUSTER 1: This cluster is mixed of all types of customers and mostly customers are those which are not doing any type of transactions. We can provide them with different offers