CREDIT CARD CUSTOMER SEGMENTATION

Contents

Abstract 2

Problem Statement 3

Objective 4

Preparation of Data: 6

Exploratory Data Analysis 6

Column wise analysis of data 7

Feature Engineering Selection/ Creation 10

ELBOW METHOD 11

INSIGHTS AND RECOMMENDATIONS 13

K- Means Algorithm 15

Elbow Method 15

INSIGHTS 17

DBSCAN 18

Hierarchical clustering 19

Auto Encoder 19

UI Development 20

Deployment & UI Testing 20

Technology Stack 21

Conclusion 22

Future Scope 23

Abstract

Credit cards are a very important source of revenues for Banks all over the world! Issuers find it difficult to differentiate between their customers based on their behaviours and payment or spending patterns.

Differentiating or segmenting these customers by their needs, behaviours and attitudes by determining the reason why customers use their credit cards by taking into account a small number of financial behaviours, helps us in solving this problem.

Credit card issuers have traditionally targeted consumers by using information about their behaviours and demographics. Behaviours are often based on credit reports on how a person spends and pays over time.

Customer segmentation is the practice of dividing a customer base into groups of individuals that are similar in specific ways. A customer segmentation model allows for the effective differentiation of customers and gives the issuers an insight on how they can attract them to use their cards more or attract new customers with new schemes.

Armed with enhanced segmentation, card issuers can not only craft better value propositions but also identify groups that are not well served by current offers. These groups can be studied on base of their behavioural pattern and given better options and schemes so as to tend to their needs.

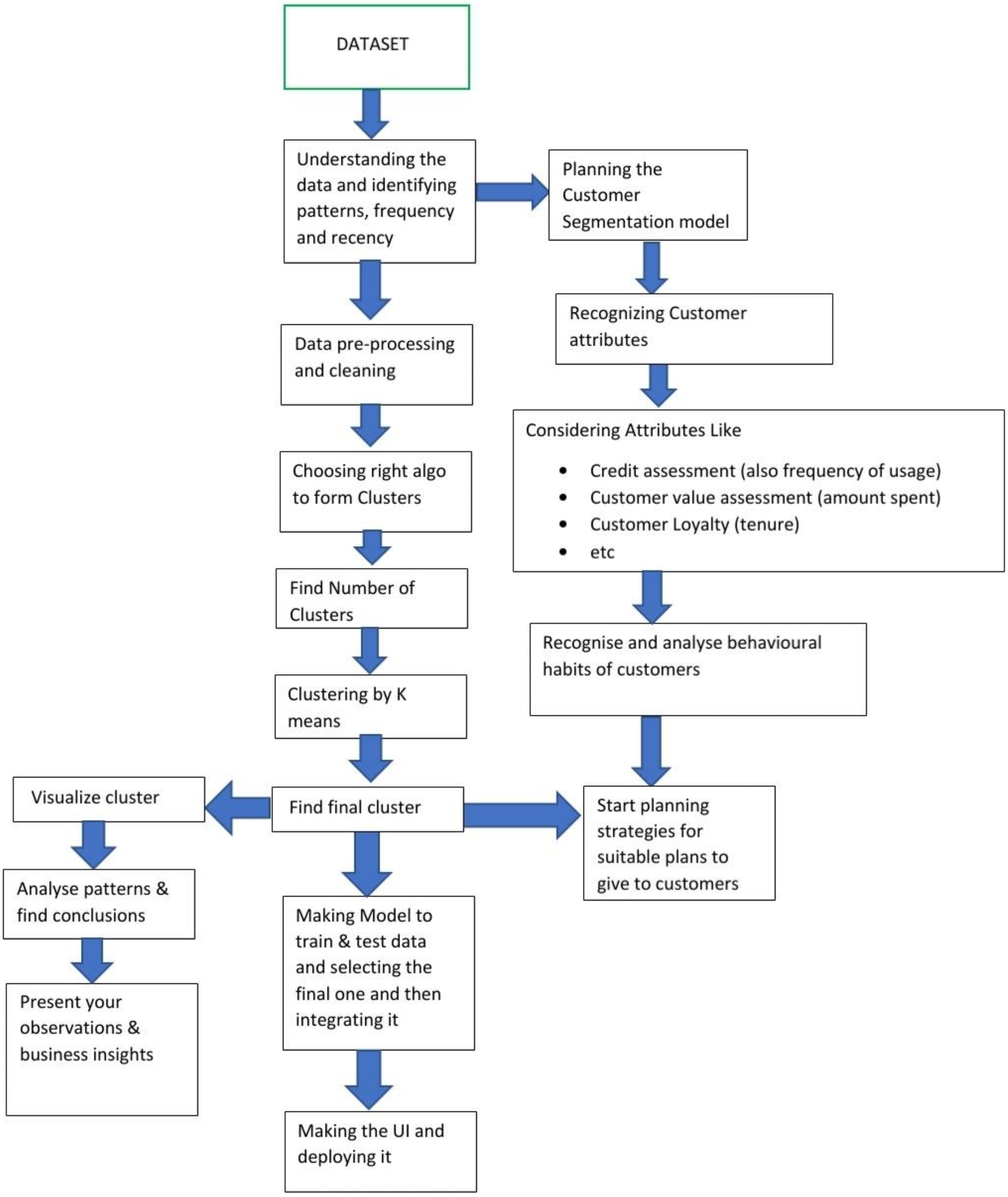
Problem Statement

RBL’s marketing department collects various customer specific data of the credit card holders. They need a mechanism to segment the customer based on underlying characteristics and form market clusters which will be easy for them to target and provide product ideas to the management. Currently these tasks are performed manually by trusting the judgment of experts in the field. These can lead to human error, biased decision and other factors which may not be helpful to create a customer cluster that actually exist. They want to design a system that would automate this process and help the different stakeholders to make informed business decision.

Objective

- To develop a solution focused on developing a clustering algorithm based on unsupervised learning.
- Try various clustering algorithm, and identify the best model for the business scenario and build and fine tune them based on characteristics.
- Deploy the machine learning model using Flask API and pickle files.
- Design a UI for entering model inputs and display results accordingly.

Flow diagram of the project:


DATASET
Understanding the data and identifying patterns, frequency and recency
Planning the Customer Segmentation model
Recognizing Customer attributes
Data pre-processing and cleaning
Considering Attributes Like
• Credit assessment (also frequency of usage)
• Customer value assessment (amount spent)
• Customer Loyalty (tenure)
• etc
Choosing right algo to form Clusters
Find Number of Clusters
Recognise and analyse behavioural habits of customers
Clustering by K means
Visualize cluster
Find final cluster
Start planning strategies for suitable plans to give to customers
Analyse patterns & find conclusions
Making Model to train & test data and selecting the final one and then integrating it
Present your observations & business insights
Making the UI and deploying it

Procedure

Preparation of Data:

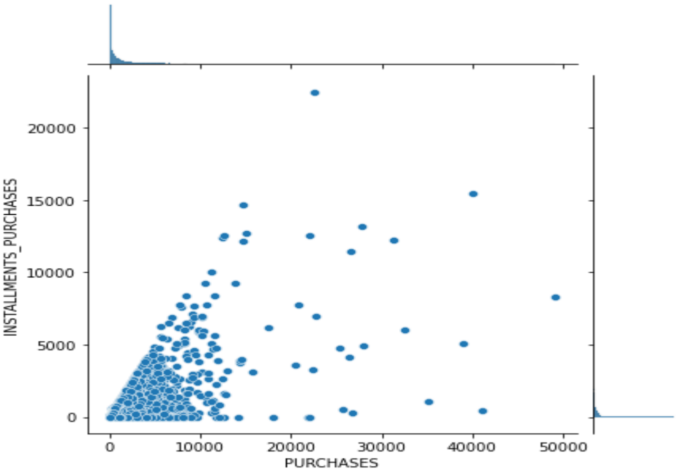
- Load the dataset onto Jupyter IDE and import all necessary libraries that will be needed for analysis.
- Performed initial analysis of the provided dataset, checking the format and type of the various numerical and categorical columns.
- Reduced the functional/business sense of each of those columns, their relevance in predicting the business solution and form initial hypothesis of the most critical features which could be crucial in designing the model.
- Checked the columns for missing values and treated those columns with feasible imputation methods.
- Column CREDIT_LIMIT with only one row having null value was dropped from the dataset.
- Column MINIMUM_PAYMENTS had around 313 missing rows. For some rows it seemed that for null value of MINIMUM_PAYMENTS, PAYMENTS column is also 0 which might suggest that the customer has not made any payment at all. For rows where MINIMUM_PAYMENTS column is Null, we imputed the value 0 considering that the customer has not made any MINIMUM_PAYMENTS.
- In a different approach we also can impute the median of that column to fill missing values.
- Column CUST_ID was dropped from the dataset, since no business sense can be derived from it and does not add value to our insights.
- The datatype of the columns was also checked and no discrepancies were found.

Exploratory Data Analysis

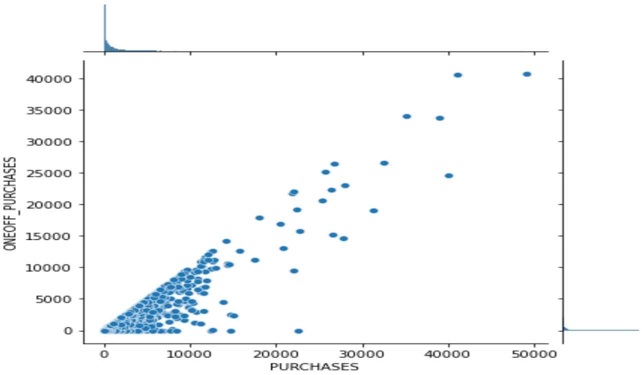
- For Exploratory Data analysis, we use the heat map for further correlation between the columns.
- There is a correlation between the columns PURCHASE_INSTALLMENT_FREQ & PURCHASES_FREQ which is 0.86 and we can see that tenure has no much relations with other columns. So, we may remove any one of these columns in the further processing.

Column wise analysis of data

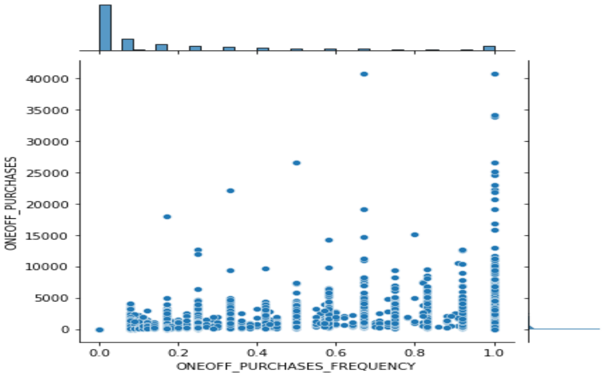
Purchases vs Installment purchases

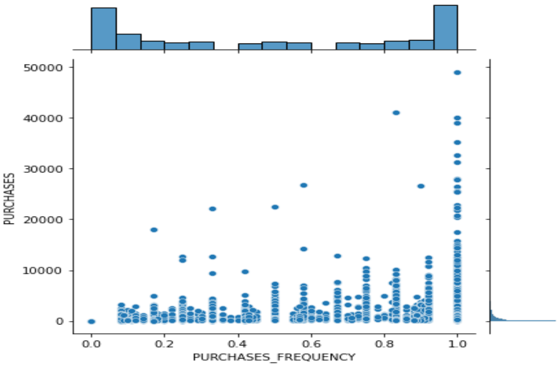


Purchases vs one-off purchases



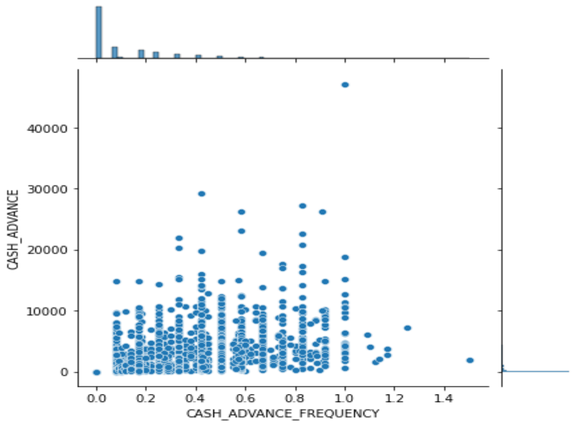
- This shows that there is a linear corelation between both the graphs which also because of the fact that one off purchase is a result of subtraction of installment purchases from purchases so we can discard one of these columns.

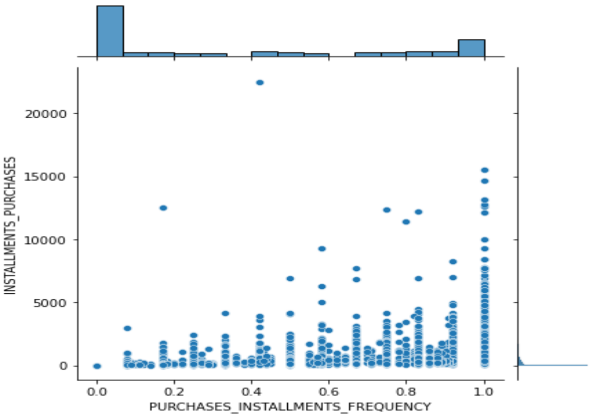




one-off purchases vs one-off purchase frequency

purchases vs purchases frequency



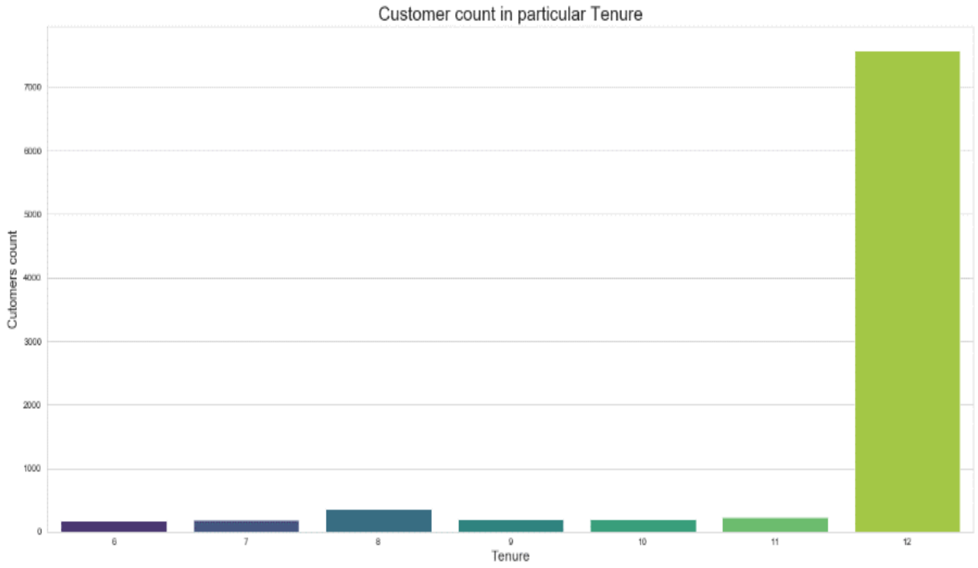


cash advanced vs cash advance frequency

installment purchases vs installment purchase-frequency

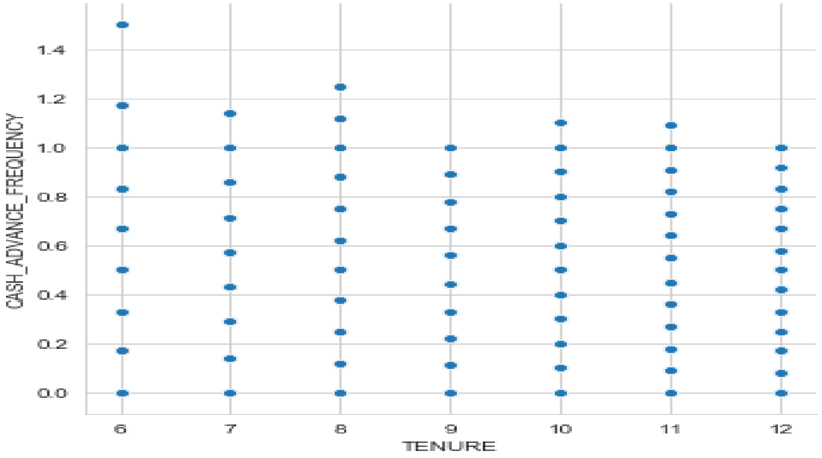
- We can discard one of the columns from one-off purchases and its frequency as it shows frequency increasing with purchase that shows an increasing trend and the same for purchase and its frequency and for purchase installments and its frequency but the same cannot be said about cash advance and its frequency since it shows no trend.

- We Further carry on our analysis with respect to tenure



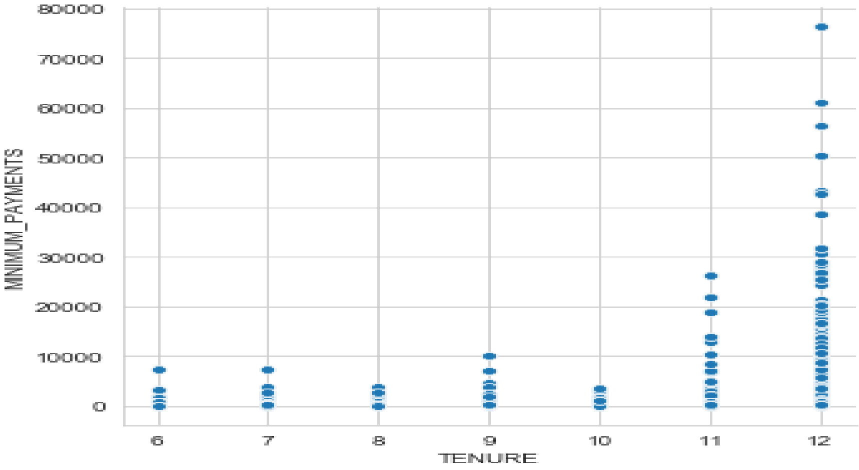
- Most of the customers have 12 months as their Tenure. So further analysis of the data set.

Tenure vs Cash advance frequency



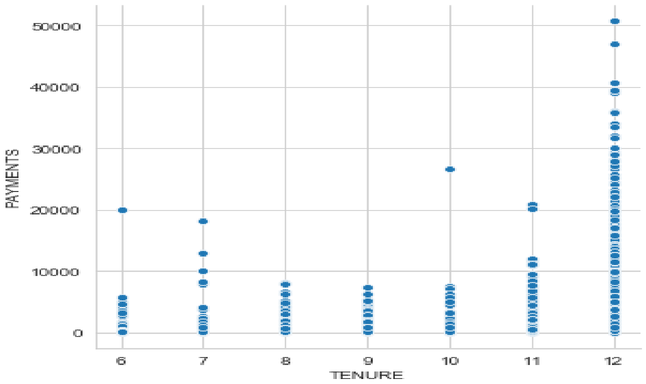
- Old customer with 12 years has cash frequency less than one the new customer who takes more cash advance.

Tenure vs minimum payment



- More the tenure more minimum payment.

Tenure vs Payments

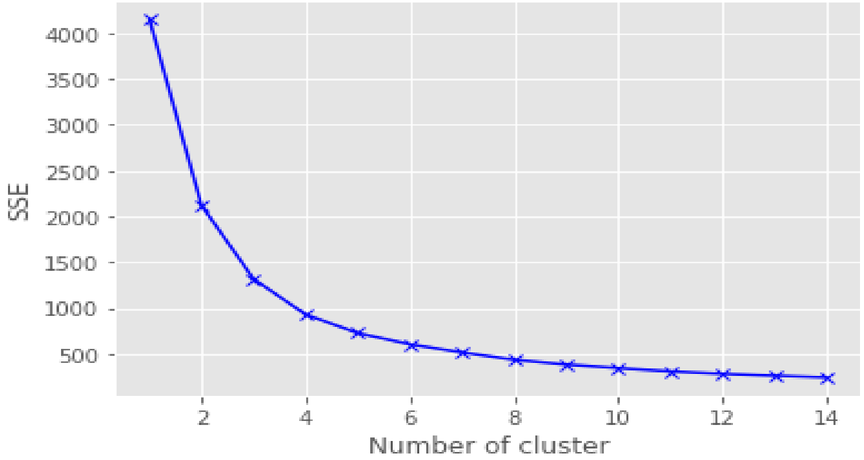


- More payments mean customer is old customer because they are more familiar with the payment system.
- Insights: Customers with a tenure range of 12 months have Higher spending power, they also pay the credit amount, and take the least cash advance.

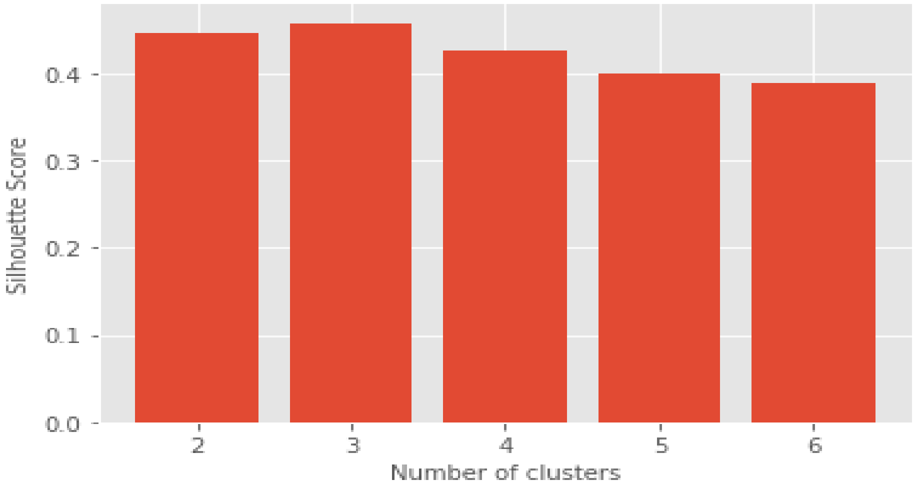
Feature Engineering Selection/ Creation

- Feature Engineering during EDA.
- The feature engineering and the new features that were created are.
- TOT_TRX column was created by adding the PURCHASES AND CASH ADVANCE.
- Average cash advance was created by dividing cash advance by its tenure.
- Average purchases were created by adding dividing purchases by its tenure.

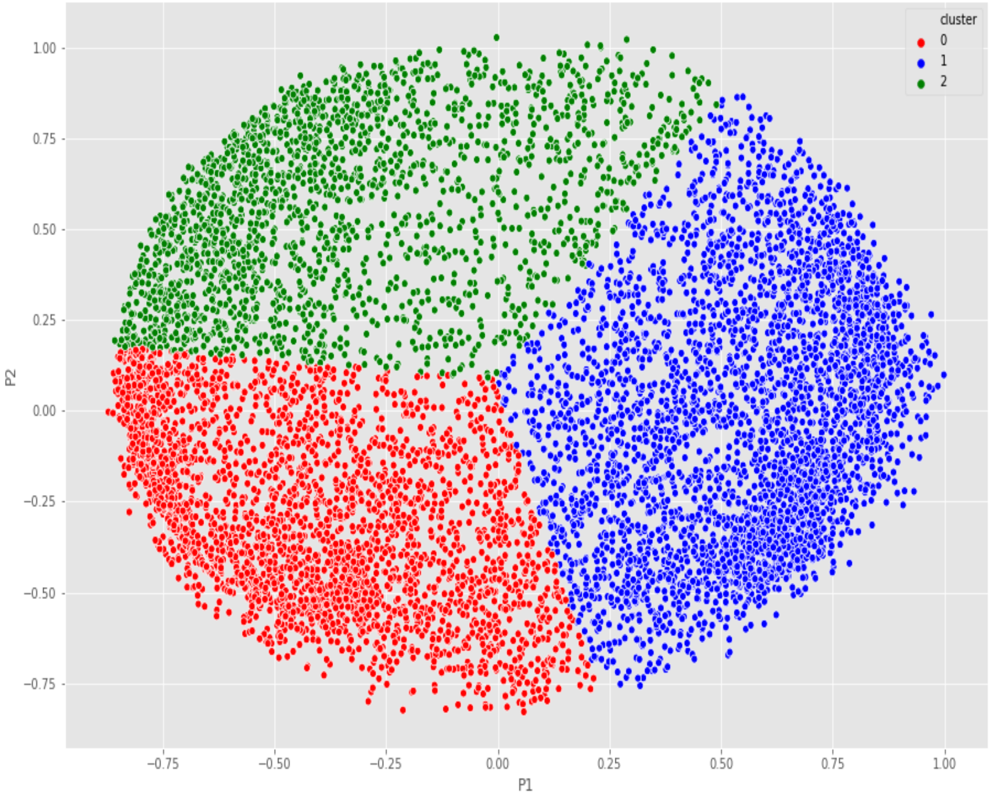
ELBOW METHOD



- We see that in the in the elbow curve, the slope of the curve starts changing after 3 or 4 hence we can consider any of these two numbers as optimal number of clusters.

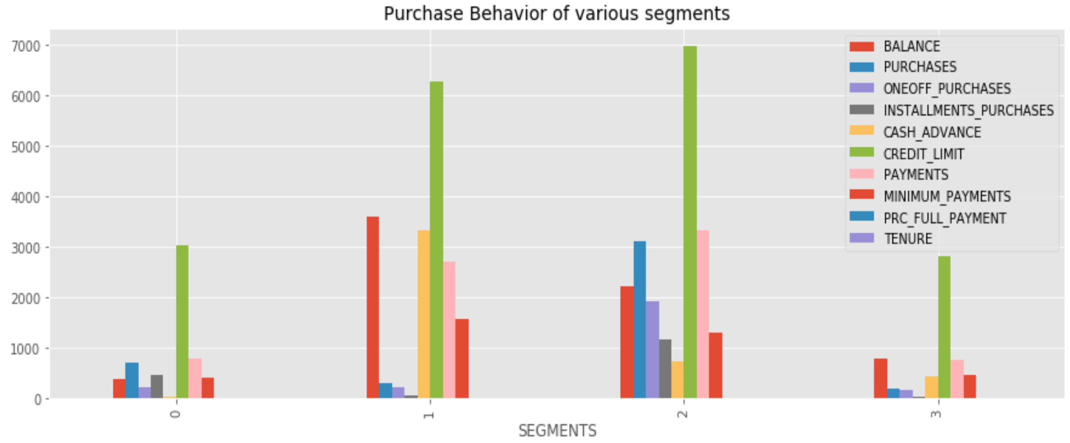


- We see that we get the best silhouette score at cluster 3.



- The approach we take to choose the cluster is through silhouette score and Elbow curve, even though the silhouette score is highest for 2, to get better insights, we can go for 4 clusters.

INSIGHTS AND RECOMMENDATIONS



GROUP 0(Middle Class but Good Buyers) = Their Balance is very less but still they purchase without taking much cash advance and prefer installments to do the purchases. They make sure to repay the installments on time.

Business Insights: These are good customers and even with less balance are managing to purchase So we should give them schemes where they can have discounts on some expensive products so that they are attracted to buy more stuff since they are frequent buyers. They are quite a number of people in this group so the company should definitely focus more on them.

Group 1(Rich but Purchase less): They are rich people with a lot of balance but not frequent buyers. Even with a lot of balance they still take a lot of cash advance to purchase suggesting that they buy expensive items. They are new customers and a few in quantity.

Business Insights: In order to make them shell their money we need to provide them schemes for expensive items as well as the ordinary ones which they are not buying using the credit card. If through the schemes they can spend money on that as well then, they have the capacity to bring a lot of profit.

Group2(Cream Customers): These are the cream customers who have been for a long time now and even do a lot of purchases. They take minimal cash advance and installments to do purchases. They even have a good balance and can afford their buy.

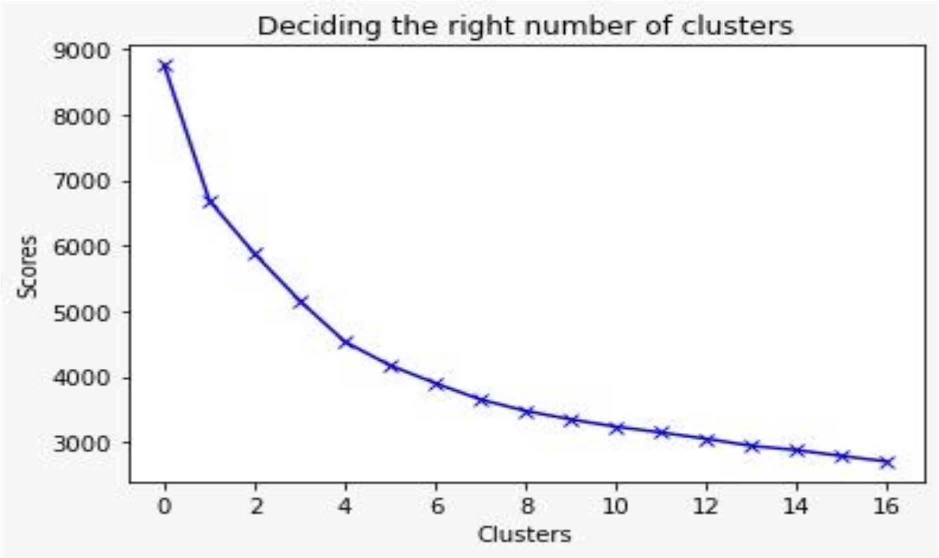
Business Insights: Make sure not to lose them.

Group3(Poor and Inert): Their balance is less and don’t do purchases. They only buy the necessities and don’t take the cash advance and installments to do so. they are a lot in number and have also been there for a good time.

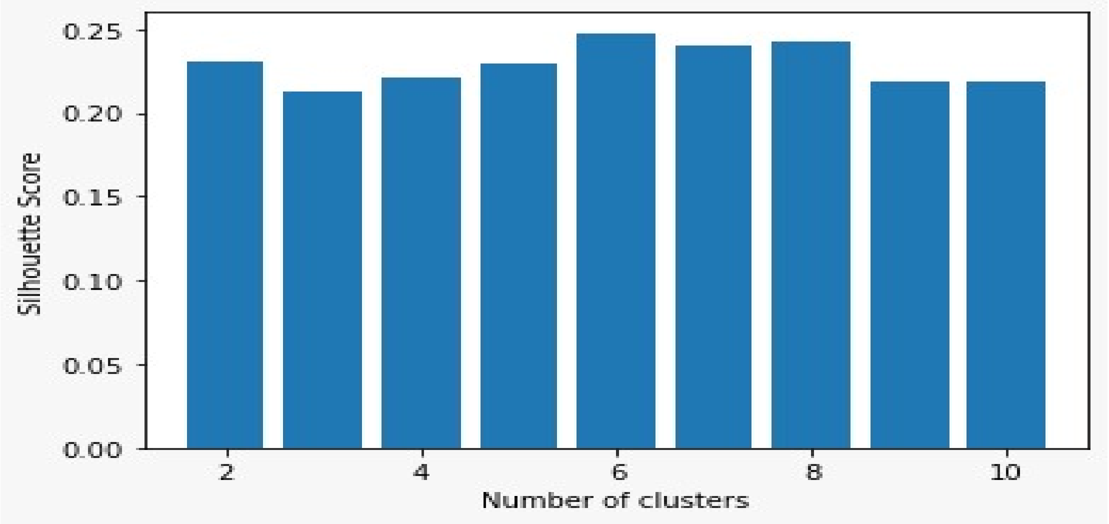
Business Insights: Focus a lot on them as they are a lot in number and aren't giving much profits. Come up with the schemes that make them comfortable to make purchases in installments and by taking cash advances.

K- Means Algorithm

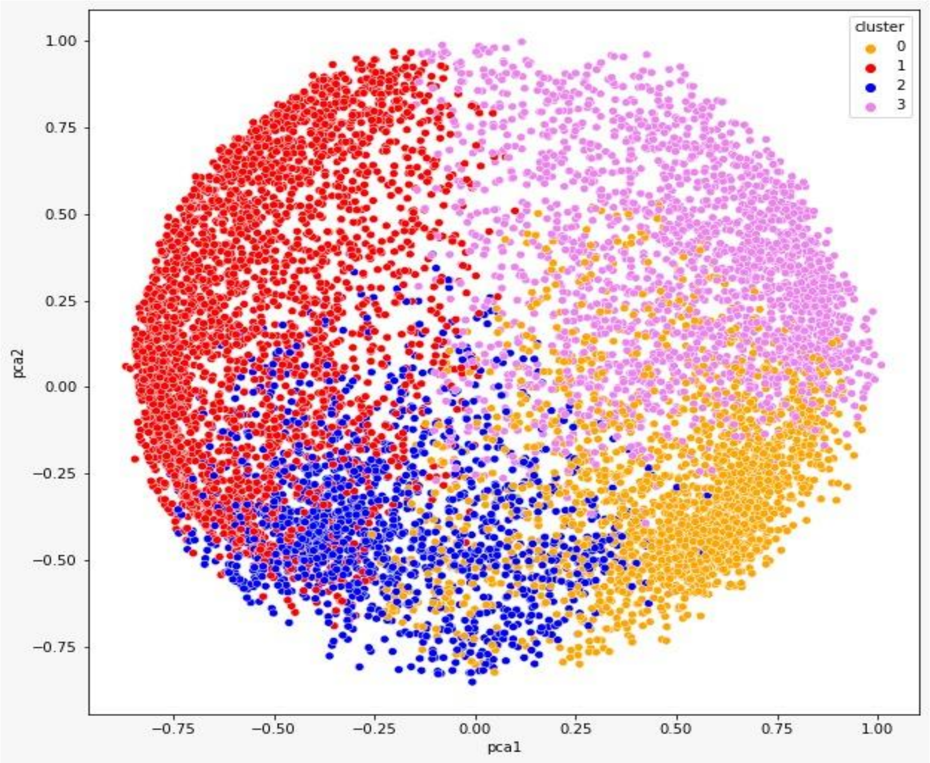
Elbow Method



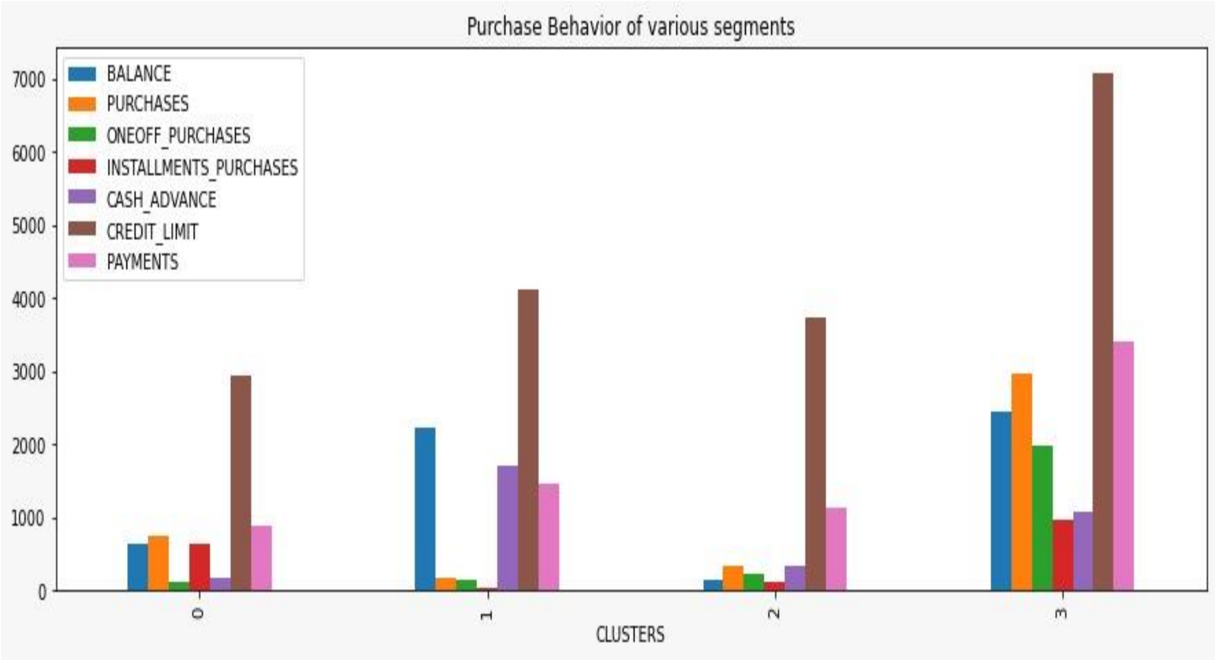
- We See that in the in the elbow curve, we the slope of the curve starts changing after 3 or 4 hence we can consider any of these two numbers as optimal number of clusters 4.



- We see that we get the best silhouette score at cluster 6.



- The approach we take to choose the cluster is through silhouette score and Elbow curve, even though the silhouette score is highest for 6, to get better insights, we can go for 4 or 5 clusters.



INSIGHTS

- **Cluster 0 (Gold)**: Customers of cluster 0 have a pretty good credit limit and balance but these customers have a very low one-off purchase and purchases in comparison to cash advance. This means they take more cash advance and spend less through purchasing. These customers should be given schemes accordingly to increase their purchase which will indirectly increase one-off purchase and installment purchases.

- **Cluster 1 (Silver)**: Credit limit of customers of cluster 1 is lowest amongst all. These customers have a low balance but they have quite high purchase and installment purchases. These customers have a very low one-off purchase and cash advance.

- **Cluster 2 (Bronze)**: These customers have a good credit limit but their balance, one-off purchase, purchases, installment purchases is very low. So, these customers should be encouraged by giving schemes to spend more and to increase the balance by keeping a minimum balance

- **Cluster 3 (Platinum)**: These customers have the highest credit limit and one-off purchase followed by purchases and installment purchases. They have highest payment of all clusters. These customers should be encouraged to keep decent balance in their account by keeping a minimum balance rule

Note: We went through DBSCAN, Autoencoders but we ended up using ***K Means*** clustering for our model

DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a popular learning method utilized in model building and machine learning algorithms. This is a clustering method that is used in machine learning to separate clusters of high density from clusters of low density. Clustering analysis or simply Clustering is basically an unsupervised learning method that divides the data points into a number of specific batches or groups, such that the data points in the same groups have similar properties and data points in different groups have different properties in some sense.

Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing. Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns. Density-based spatial clustering of applications with noise (DBSCAN) is a well-known data clustering algorithm that is commonly used in data mining and machine learning. The easier-to-set parameter of DBSCAN is the mints parameter. Its purpose is to smooth the density estimate, and for many datasets it can be kept at the default value of mints = 4 (for two-dimensional data). The advantage of this is great at separating clusters of high density versus clusters of low density within a given dataset and is great with handling outliers within the dataset.

Hierarchical clustering

Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

Hierarchical clustering can be performed with either a distance matrix or raw data. When raw data is provided, the software will automatically compute a distance matrix in the background.

Hierarchical clustering starts by treating each observation as a separate cluster. Then, it repeatedly executes the following two steps: (1) identify the two clusters that are closest together, and (2) merge the two most similar clusters. This iterative process continues until all the clusters are merged together.

Auto Encoder

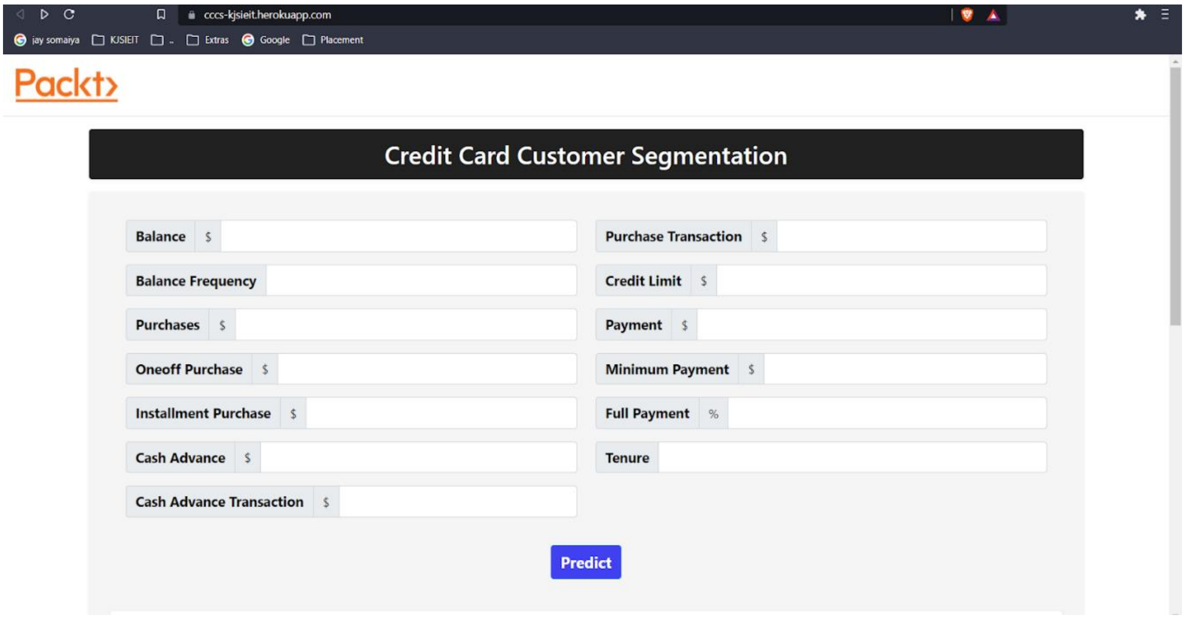
Autoencoders are a type of artificial neural network that is used to learn feature representation in an unsupervised manner. It uses the same data for input and output. By adding a bottleneck in the network, it forces the network to create a compressed version of the input data, which is how the encoder works. Meanwhile, the decoder reconstructs the encoded features to its original input.

There are different types of autoencoder models. In a stacked autoencoder model, encoder and decoder have multiple hidden layers for encoding and decoding.

Clustering is difficult to do in high dimensions because the distance between most pairs of points is similar. Using an autoencoder lets you re-represent high dimensional points in a lower-dimensional space. It doesn't do clustering per se - but it is a useful preprocessing step for a secondary clustering step. You would map each input vector xixi to a vector zizi (not a matrix...) with a smaller dimensionality, say 2 or 3. You'd then use some other clustering algorithm on all the zizi values

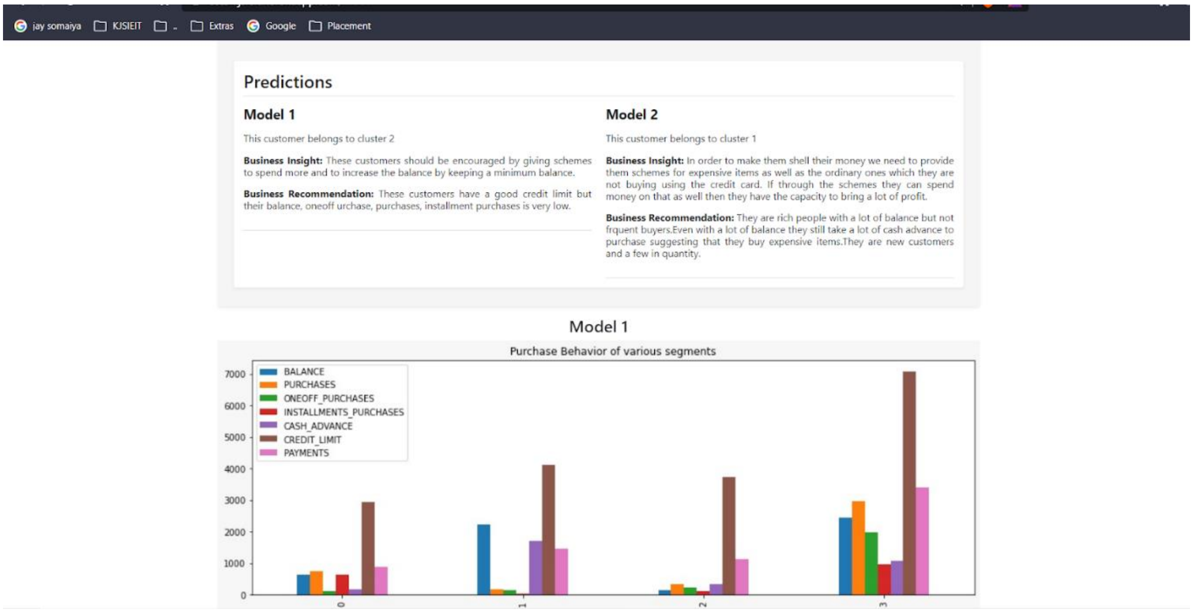
UI Development

- In order to develop an interactive and smooth User Interface (UI) to enable the end user for input his/her own inputs and receive model insights, we have used web development frameworks such as HTML, CSS.
- We incorporated three model along with user inputs and predict buttons for each of these models that are integrated with the backend using Flask Framework.
- Flask is a micro web framework written in python. It is classified as microframework because it doesn’t require any particular tools and libraries.



Deployment & UI Testing

- Flask API was used to deploy the machine learning model in backend. It is used to manage HTTP requests and uses API function to get the data and display the result to the end user in front end UI.
- User enters the inputs for all three models, post routing of those values, our machine learning model makes predictions and returns the same.
- As a final result, on inputting the user inputs, a backend call routed via Flask framework helps us in providing relevant insights and recommendations to the end user.



Predictions

Model 1

This customer belongs to cluster 2

Business Insight: These customers should be encouraged by giving schemes to spend more and to increase the balance by keeping a minimum balance.

Business Recommendation: These customers have a good credit limit but their balance, oneoff urchase, purchases, installment purchases is very low.

Model 2

This customer belongs to cluster 1

Business Insight: In order to make them shell their money we need to provide them schemes for expensive items as well as the ordinary ones which they are not buying using the credit card. If through the schemes they can spend money on that as well then they have the capacity to bring a lot of profit.

Business Recommendation: They are rich people with a lot of balance but not frquent buyers.Even with a lot of balance they still take a lot of cash advance to purchase suggesting that they buy expensive items.They are new customers and a few in quantity.

Model 1

Technology Stack

- Jupyter: Open-source IDE used for development
- Python: Programming language
- Numpy: Python library for creating arrays
- Pandas: python library for data manipulation and analysis
- EDA Libraries: matplotlib, sea-born visualization libraries
- Flask API: Web Framework for Python for making web apps & managing HTTP requests
- HTML, CSS: website designing frameworks

Conclusion

- We were able to load the dataset onto Jupyter IDE and perform relevant analysis and data manipulation for deriving further insights.
- Initial Data Quality measures were performed for making the data useful for further analysis.
- Exploratory Data Analysis was performed in order to obtain visual information on how the data was impacting certain business requirements. Feature correlations were visualized and their distributions were plotted.
- Post EDA, further data quality improvement measures such as Standardization, Normalization and dimensionality reduction techniques were implemented.
- Multiple trials of various cluster formation techniques such as DBSCAN clustering, Hierarchical clustering, K means clustering (with/without encoders, with/without PCA) were implemented in order to finalize best possible clustering algorithm to implement the model upon.
- Cluster Evaluation techniques such as Elbow Score, silhouette method was used to determine the quality of each of these clusters.
- Cluster visualization helped us determine the critical features and their insights to conclude on business recommendations.
- UI development using HTML, FLASK API was done in order to create a suitable and interactive UI for determining the clusters based on end user’s inputs

Future Scope

- The UI web page can be used to introduce more user interactive and automatic features and make it easier for end user to enter inputs and get predictions.
- The web page can be useful for clients belonging to the banking and finance industries wherein this interface could be embedded into their systems permanently.
- More feature engineering in terms of interest rate.
- More clustering techniques like DBSCAN, Auto encoder can be used further.
- Displaying or predicting more in graphical format.