Credit Card Dataset

Dataset:

By segmenting customers, credit card companies can identify their unique needs and preferences, and tailor their marketing efforts accordingly. This can lead to more effective communication, increased customer satisfaction, and ultimately higher profits for the company. The dataset summarises the usage patterns of approximately 9000 active credit card users over six months.

The sample contains 18 behavioural variables at the customer level. This dataset can be used to identify patterns and trends in credit card usage, which can help financial institutions make informed decisions about credit limits, rewards programmes, and marketing strategies. Additionally, the behavioural variables can provide insights into customer preferences and habits, allowing for personalized offerings and improved customer satisfaction.

Attributes:

Following is the Data Dictionary for Credit Card dataset :-

CUST_ID : Identification of Credit Card holder (Categorical)
BALANCE : Balance amount left in their account to make purchases (
BALANCE_FREQUENCY : How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated)
PURCHASES : Amount of purchases made from account
ONEOFF_PURCHASES : Maximum purchase amount done in one-go
INSTALLMENTS_PURCHASES : Amount of purchase done in installment
CASH_ADVANCE : Cash in advance given by the user
PURCHASES_FREQUENCY : How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased)
ONEOFFPURCHASESFREQUENCY : How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)
PURCHASESINSTALLMENTSFREQUENCY : How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done)
CASHADVANCEFREQUENCY : How frequently the cash in advance being paid
CASHADVANCETRX : Number of Transactions made with "Cash in Advanced"
PURCHASES_TRX : Numbe of purchase transactions made
CREDIT_LIMIT : Limit of Credit Card for user
PAYMENTS : Amount of Payment done by user
MINIMUM_PAYMENTS : Minimum amount of payments made by user
PRCFULLPAYMENT : Percent of full payment paid by user
TENURE : Tenure of credit card service for user

Data Pre-processing:

The dataset is first cleaned by removing all the NaN values and replacing them with the median values of that column. Since the model only reads data in numerical form, the

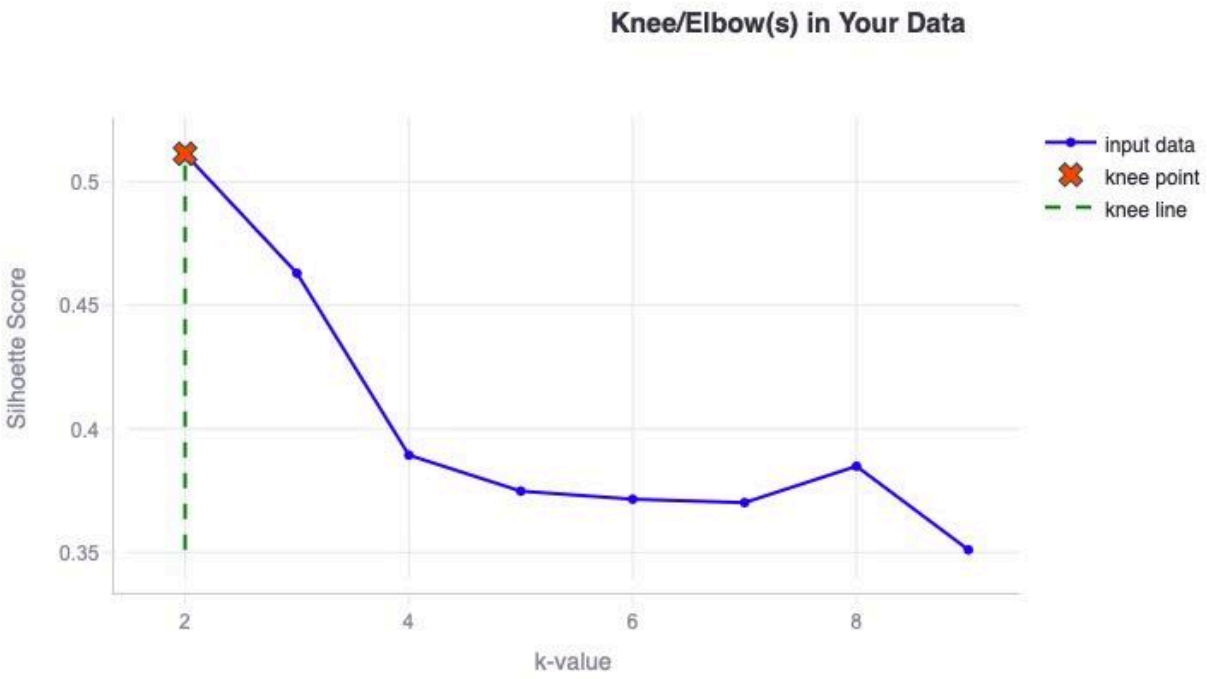
categorical data are transformed into numerical labels for the model's convenience in processing the data.

However, the dataset can still be modified further as needed by the client. For instance, the client may choose to normalise the data or remove outliers to improve the accuracy of the model. It is important to ensure that any modifications made to the dataset are carefully documented and validated to avoid any unintended consequences for the model's performance.

Methodology/Workflow

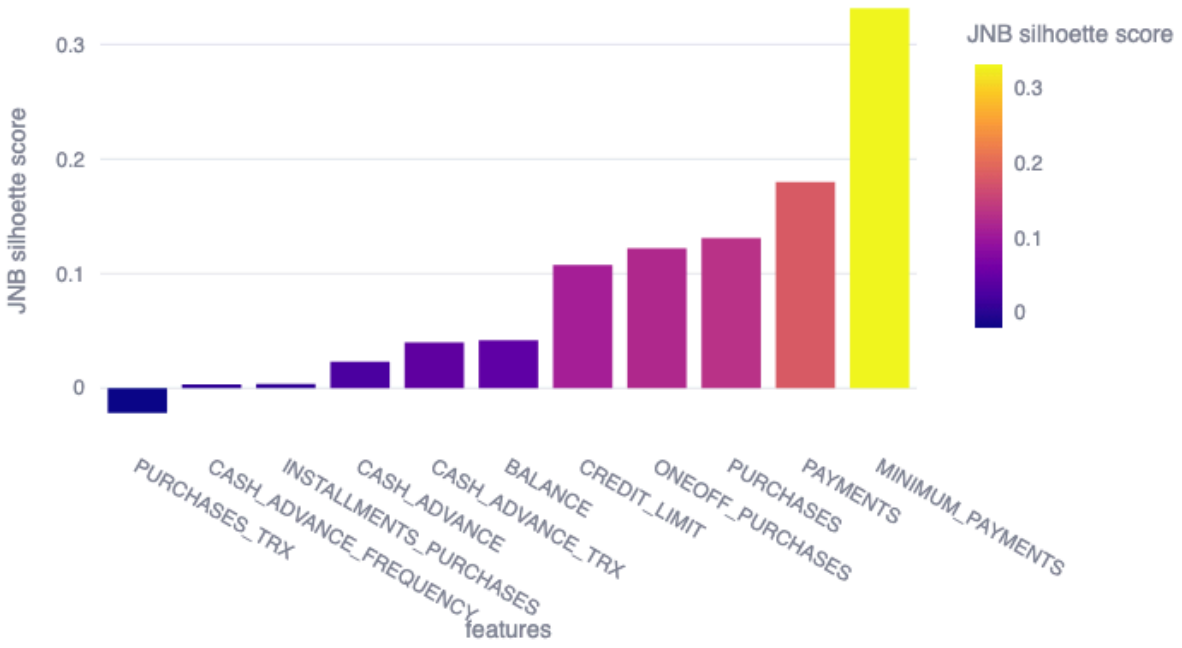
All the attributes are split into either numerical or categorical types based on their clustering threshold,i.e., if the number of unique values is < 25, then the feature is considered categorical. Among all the 18 features, 11 were tagged as numerical features, and the rest 7 as categorical in nature. The numerical features are then clustered into 2 (calculated through the elbow method) groups using JNB clustering [2]. The features are ordered according to their silhouette scores [3] and only the top 40% are selected for further steps. All possible subsets of these feature spaces are plotted into a feature matrix [4]. For this case study, we considered the combination of “ON/OFF PURCHASES “ and “MINIMUM PAYMENTS” with a silhouette score of 0.923. The entropy of all the categorical features and the selected set is compared in Figure [5].

To create each cluster of the chosen numerical features, K-means clustering was employed. This is then the joint entropy calculated against all the categorical features. The maximum entropy is about 0.05 for ‘PURCHASES_FREQUENCY’, and the minimum is 0.01 for “TENURE”. The user can select one or more categorical features from it, depending on the situation.



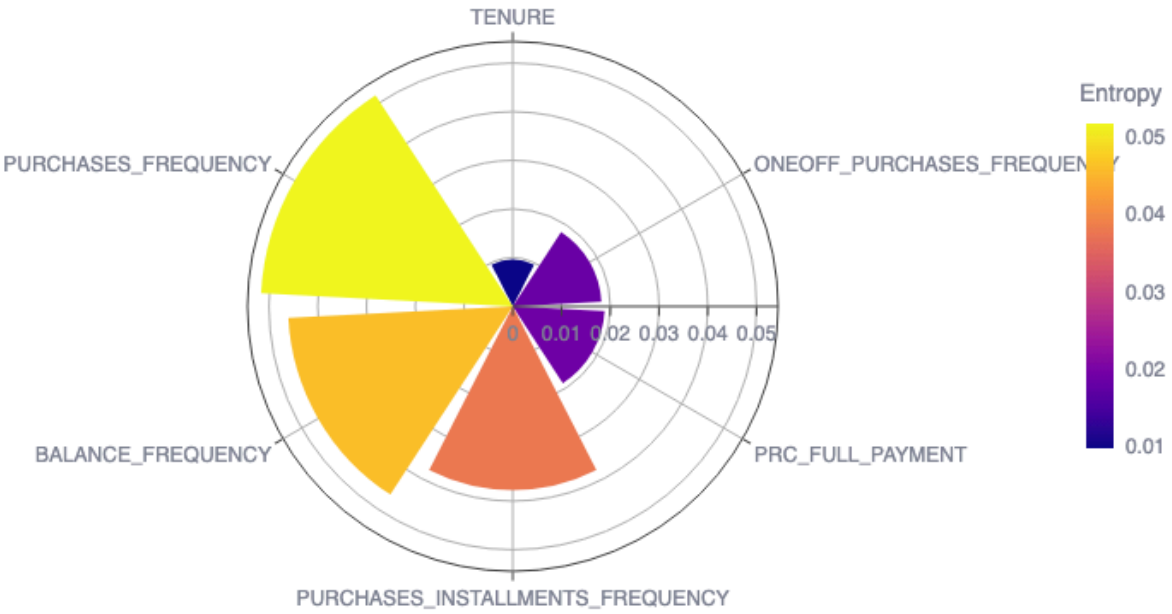
The k-values silhouette score graph, showcasing the elbow value [2]

The Silhouette scores of JNB Clustering are based on the features [3]:



Feature matrix for case study 4 [4]:

The entropy of each categorical feature w.r.t. The selected numeric feature/s[5]



Reference

[1] https://www.kaggle.com/datasets/arjunbhasin2013/ccdata