Submitted By: Manash Bhele (220433)

Analysis and prediction of Customer Clusters based on Various attributes

Using PySpark and Tableau

Softwarica College of IT and E-Commerce

STW7082CEM Big Data Management and Data Visualisation

Contents

[Introduction 4](#_Toc144494165)

[Implementation 5](#_Toc144494166)

[1. PySpark Configuration and Setup 5](#_Toc144494167)

[2. Dataset Description and Analysis Goal 8](#_Toc144494168)

[3. Importing Necessary Libraries and Initializing Spark Context 10](#_Toc144494169)

[4. Data Preprocessing 11](#_Toc144494170)

[a. Dealing With Null Values 14](#_Toc144494171)

[b. Changing Categorical Strings Features to Numerical 15](#_Toc144494172)

[c. Using Vector Assembler to prepare data for creating models 16](#_Toc144494173)

[5. Customer Cluster Analysis Using K-Means Clustering Algorithm 17](#_Toc144494174)

[a. What is K-Means Clustering? 17](#_Toc144494175)

[b. Calculation of Silhouette Score for K-Means Clustering 18](#_Toc144494176)

[c. Performing Clustering using k = 7 20](#_Toc144494177)

[d. Finding out whether customer will accept the final campaign using Random Forest Classifier algorithm 20](#_Toc144494178)

[Discussion of Findings 23](#_Toc144494179)

[1. Findings based on K-Means Clustering 23](#_Toc144494180)

[2. Findings Based on Random Forest Classification 26](#_Toc144494183)

[Conclusion 29](#_Toc144494184)

[References 30](#_Toc144494185)

[Appendix 31](#_Toc144494186)`

# Introduction

Customers are the most important part of any businesses. How a company deals with its customers plays a huge role in determining whether the company grows or falls apart. Analysing the customer base, taking into consideration various customer attributes and patterns play a significant part in helping the company understand their customers well. The more the company understands their customers and their needs, the better the company can work on developing and delivering customer driven products which inturn helps in the growth of the company. (Hannu Kärkkäinen, 2001)

This project focuses on analysing a variety of personal and buying behaviors of customers and find out insights regarding the customers as well as predict how customers react to various campaigns led by the company. For finding out valuable insights related to customers, a Clustering analysis has been performed. Furthermore, a classification analysis has also been implemented to find out whether these attributes can be used to determine whether customer attributes can be used to predict if they accept the campaigns led by the company.

The dataset has been taken from Kaggle and all the related links has been provided below in their respective sections. For the part of analysis and prediction, tools such as Tableau and PySpark has been used. The coding has been done in Python using PySpark, which is a Python API for Apache Spark. Making use of the PySpark Mllib library, different machine learning analysis has been done on the data. Whereas the data exploration and analysis of the results using visualizations have been created in Tableau.

All the details concerning the tasks carried out, the dataset, packages used and the results attained, have been discussed in detail in their respective sections below.

# Implementation

This section breaks down the technical aspects of the project’s implementation part. First, the configuration and installation process for PySpark is discussed.

## PySpark Configuration and Setup

Apache Spark is an analytics engine which is mainly used for large-scale data processing on single node mahcines or in clusters. It is available in many languages. PySpark is a Python API for Apache Spark using Python. The main features of PySpark which is used in the project are:

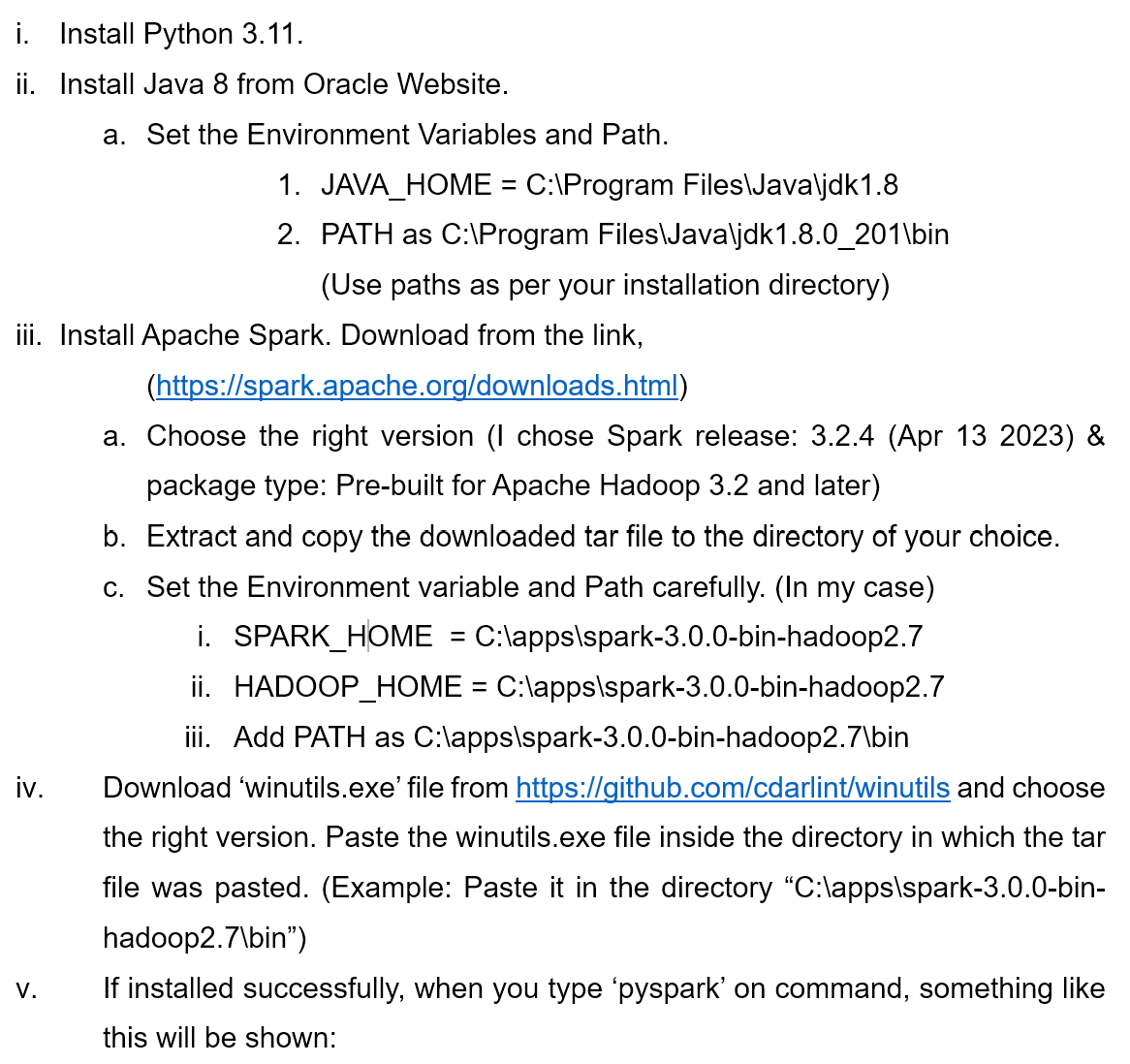
1. PySpark Dataframes

* PySpark Dataframes help to read, write, transform and analyze data efficiently using Python and SQL.

1. Machine Learning (MLlib)

* MLlib is a machine learning library build on top of Spark which allows users to create and train machine learning models.

The following steps were followed in order to install Apache Spark and run pyspark on Windows 10.



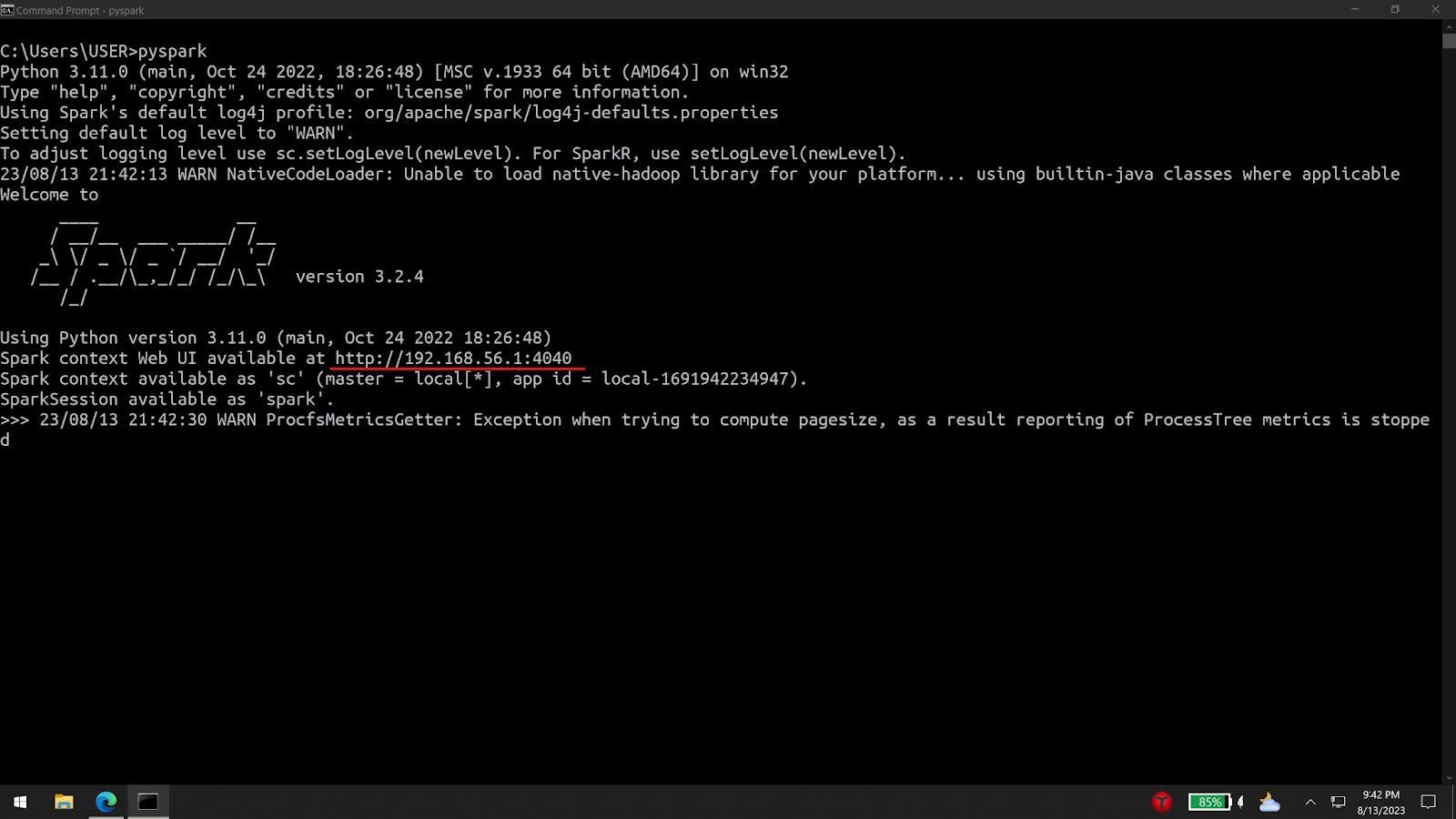


Figure 1 Checking if Spark installed Succesfully

For the coding task, Jupyter Notebook was used. To use pyspark in jupyter notebook, pip (a package installed for python) was used to install pyspark. To make sure no issue arises, the Apache Spark installation version and pyspark installation version must be the same. (3.2.4 in this case). When using pip to install pyspark, the specific version can be specified and the pyspark version can be checked as shown in the picture.



Figure 2 Installing and Checking Version of PySpark

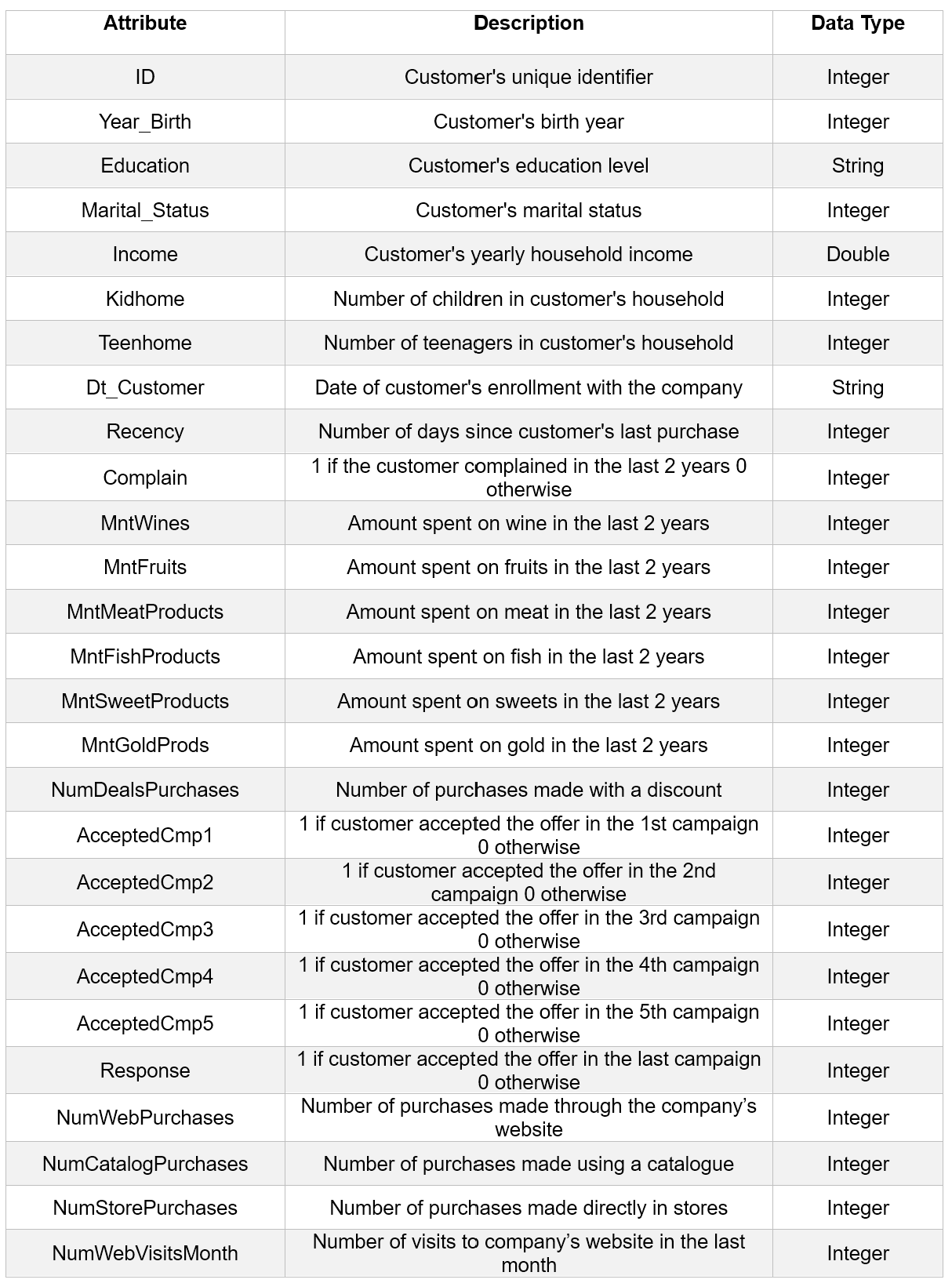
## Dataset Description and Analysis Goal

The dataset was download from Kaggle. It can be found in Kaggle’s website with name “Customer Personality Analysis” or can be downloaded using the link <https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis>. The dataset is a collection of various attributes of different customers of a company. Initially it contains 2240 rows and 29 columns or features.

The dataset consists of integers, doubles and strings. Some features have been assigned data types different from what it really is. For example, the ‘Dt\_Customer’ which should originally be a Date datatype has been assigned as string. Also, the ‘Year\_Birth’ feature has been assigned a string datatype. These kinds of incosistencies will be addressed when preprocessing the data which will be discussed and explained in the latter sections.

The analysis of this dataset can provide many valuable insights on the customers as well as spread light on the parts which the company needs to focus on. The features consist of both personal descriptions of the customers as well as their behaviours regarding how and when they buy goods. Using these features, this project aims on performing a cluster analysis of the customers which might help unfold different clusters of customers based on different attributes which would help the company in mnay ways. Also, another goal of this analysis is to create a predictive model which would help in predicting if a customer would accept the promotional or other campaigns led by the company. Following sections will provide the findings of exploratory analysis of the dataset.

The folowing table below provides detail about the various features with their data types.



## Importing Necessary Libraries and Initializing Spark Context

This project uses PySpark and its functions solely for all data preprocessing, analysis and model creating tasks. The following image provides a code snippet showing the libraries that needs to be imported.

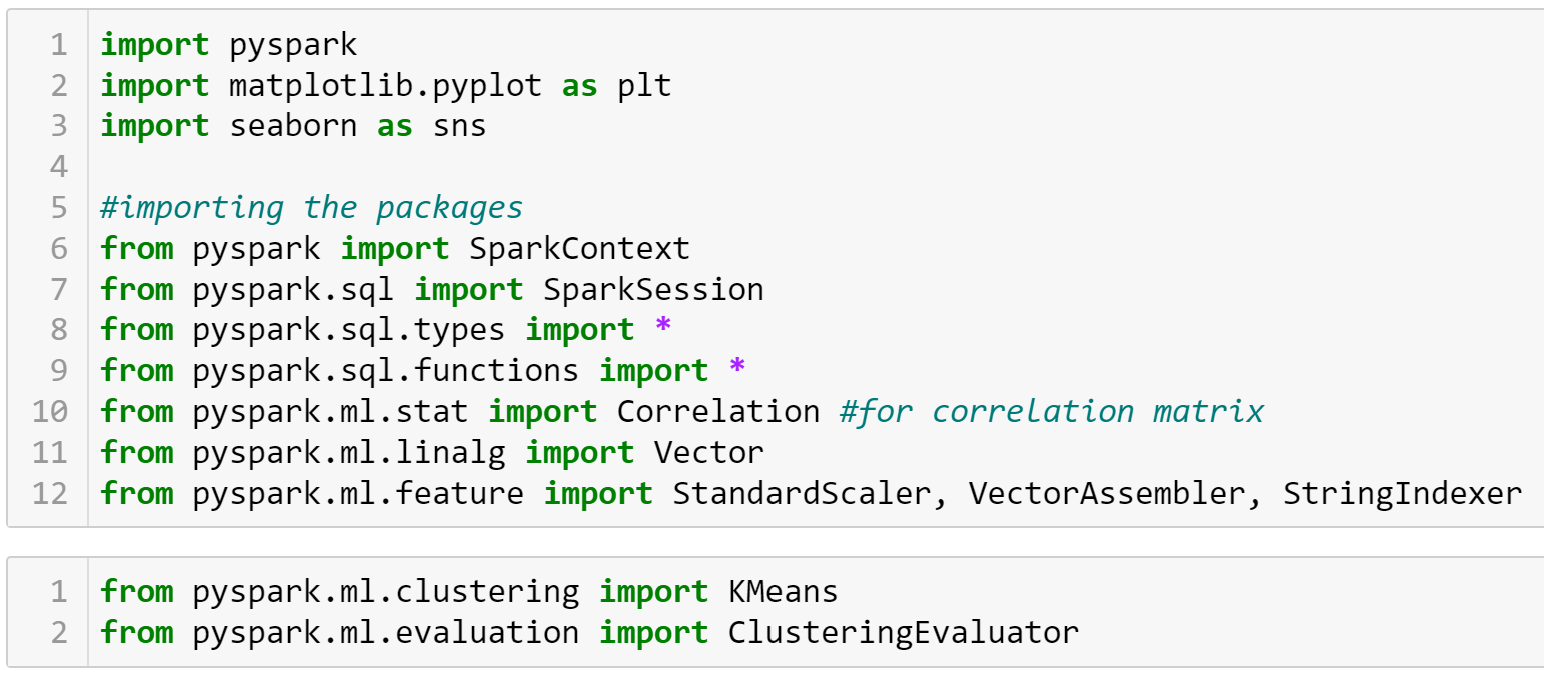


Figure 3 Importing required libraries and functions

To interact with Apache Spark, either a Spark Context or a Spark Session can be created as it is the entry point to connect with Spark. Here, the spark session has been used to interact with Apache Spark. ((NNK), 2023)

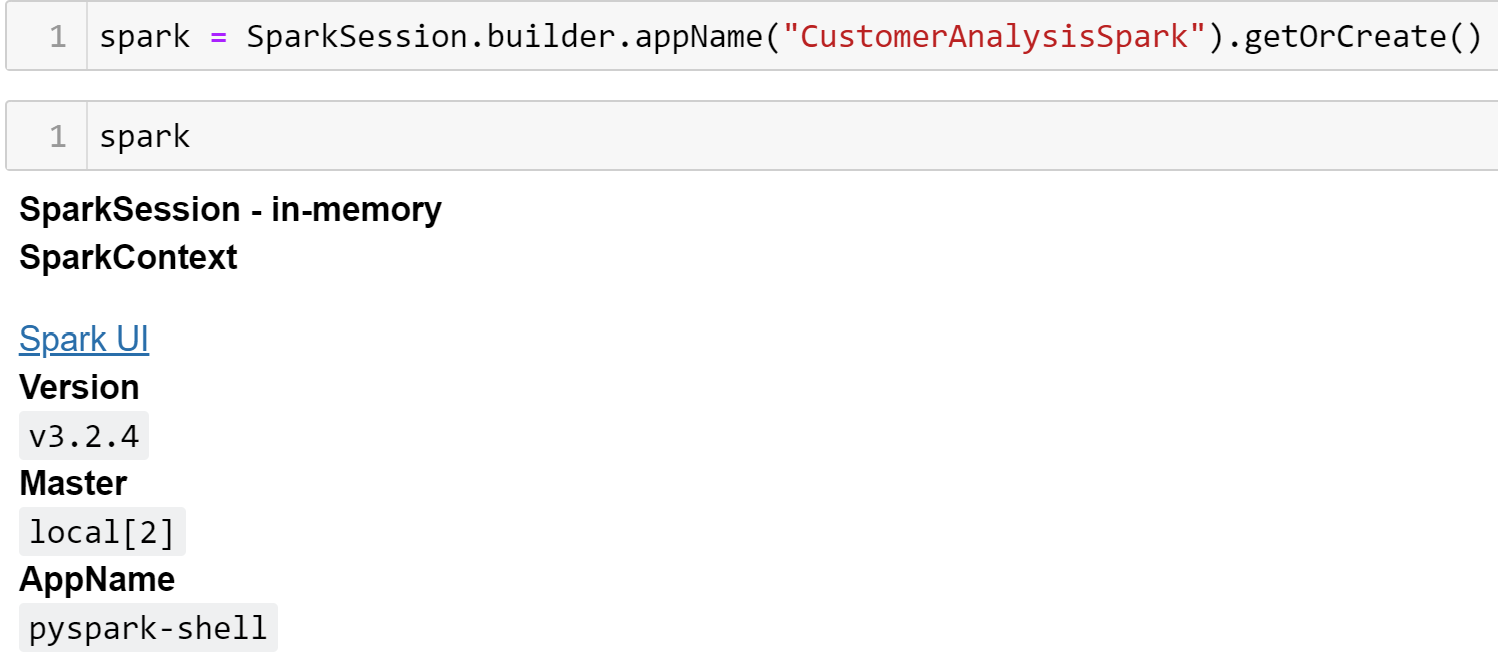


Figure 4 Spark Session Creation

After creating the spark session, the UI of the spark session can be accessed using the ‘Spark UI’ link. The created session is then used to load the dataset as follows:



Figure 5 Loading the Dataset using spark

## Exploratory Data Analysis

## Data Preprocessing

For any data related task, data preprocessing is the most crucial step. Data Preprocessing is essential before using it. The data is preprocessed in order to check for any missing values, noisy data, and other inconsistencies before executing it to any kind of algorithms. (Pushpa Singh, 2021) Computational intelligence needs clean, numeric, homogeneous, well organized, and normalized data. Data quality is a major concern in big data processing and knowledge management systems. (Khalid K. Al-jabery, 2020)

Upon looking at the dataset, it is seen that it has 30 columns, where one column is an additional index column added by the dataframe. It is dropped for the sake of simplicity. The number of rows in the dataset is 2240. The following image shows the results of printing the schema of the DataFrame.

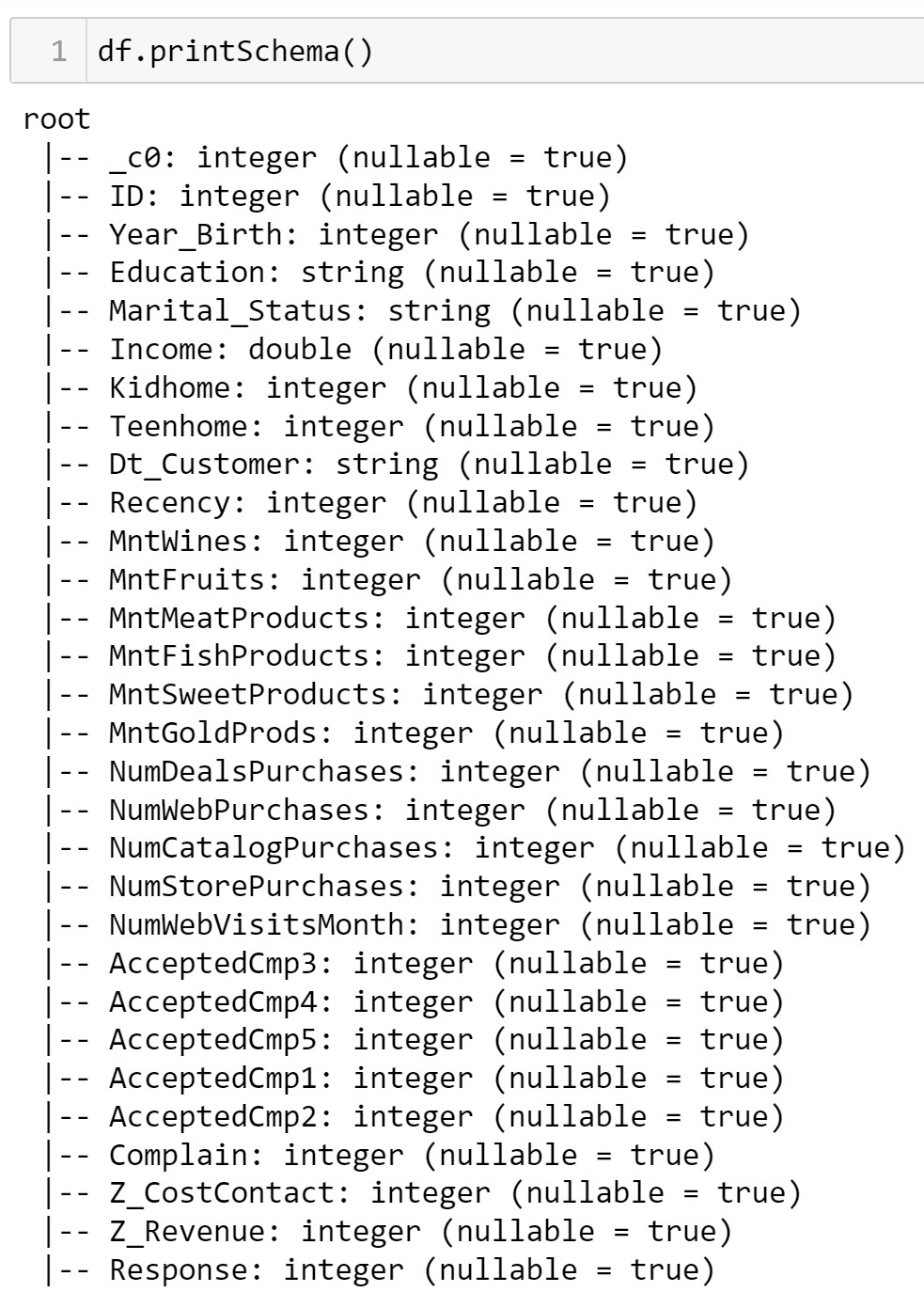


Figure 6 Schema of the DataFrame

As discussed earlier, there are many inconsistencies in the data types of the dataset which needs to be addressed. And also, there are columns which needs to changed from one type to another which would be helpful for later part to create the machine learning models.

The ‘Dt\_Customer’ column has the dates when the customer got engaged with the company. As it is a date column but has been assigned as a string type, it needs to be changed to date type. It can be done by using the “df.withColumn()” method.

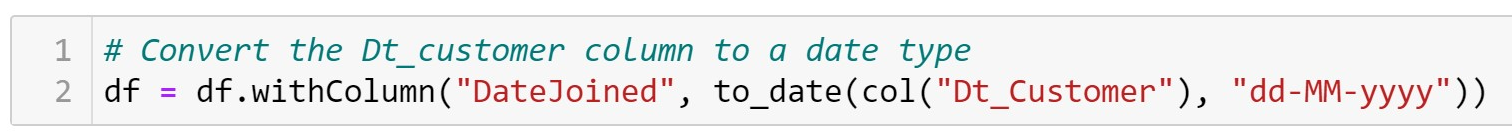


Figure 7 Changing Column Dt\_Customer to Date type and changing name to DateJoined

Along with changing the date type, the name of the column has also been changed to ‘DateJoined’. Furthermore, another column ‘YearJoined’ and ‘DaysSinceJoined’ has also been derived from the feature which would help in visualisation and model building aspects.

One main point to be noted with this dataset is that the data was collected back in 2015 and many columns have values that would only make sense if the reference date for present is taken as 2015. A feature which would benefit from this assumption is the ‘Year\_Birth’ feature which is the year of birth of the customers. The ages of the customers have been calcuated for the year 2015 as shown in the image below.

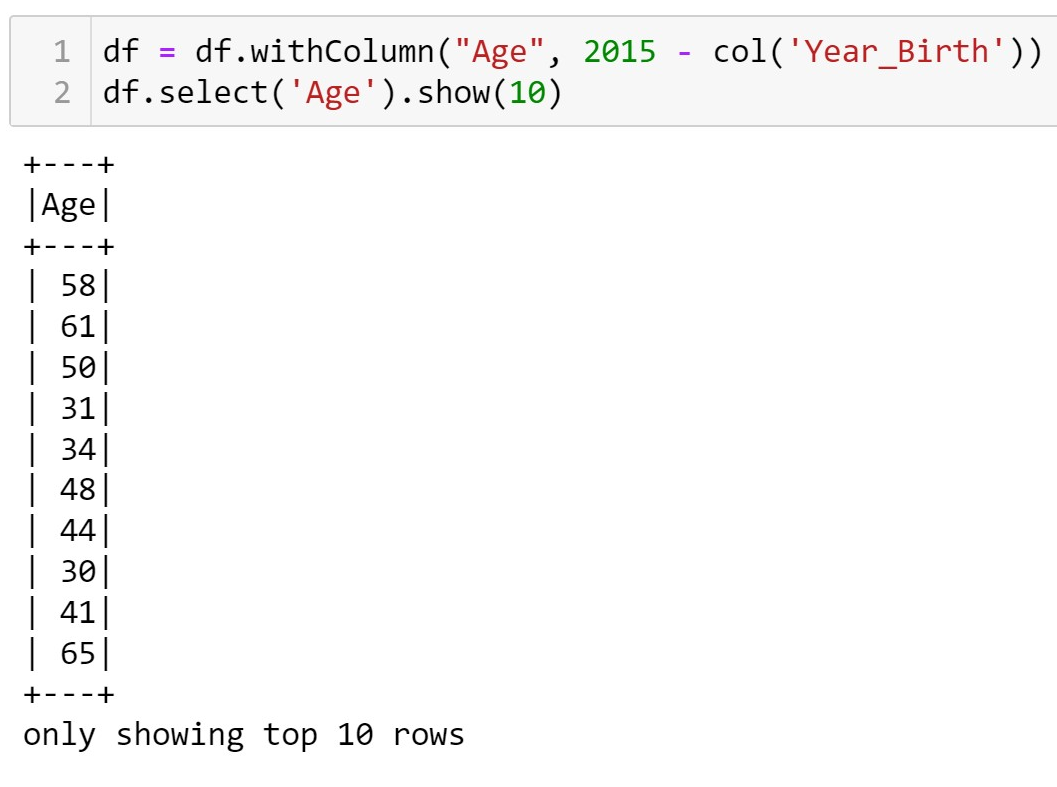


Figure 8 Creating Age attribute form Year\_Birth

Features like ‘ID’ which would not provide much value to the dataset was also dropped from the dataset. To deal with the repetitive columns that were created earlier and for simplicity of model building the created columns ‘Year\_Birth’, ‘YearJoined’, and ‘DateJoined’ were also removed. The resulting dataframe then had 28 columns with 2 non numerical columns and 26 numerical columns.

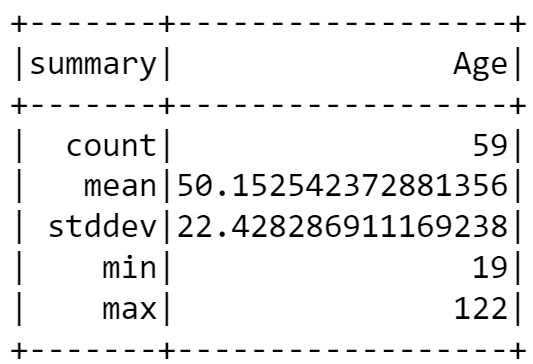
### Dealing With Null Values

When checked for null values, the only column with null values was ‘Income’ with 24 null values. They were dropped using the “df.na.drop()” function which brough the number of rows down to 2216.



Figure 9 Null values in Income

The ‘Age’ feature of the dataset provides the age of the customers. Upon looking at the counts of unique ages of the customers it was found that there are customers of age more than 100. These values seemed unreal so they were also dropped.



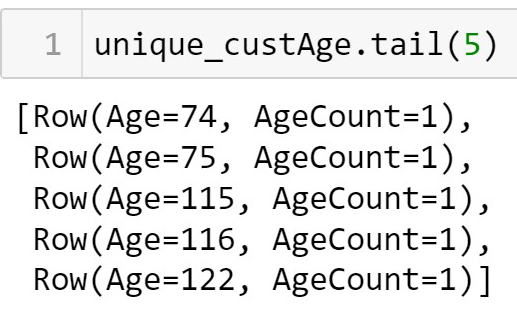


Figure 10 Ages more than 100 Figure 11 Summary of Age

‘Education’ and ‘Marital\_Status’ are two string categorical features in the dataset. For create ML models, such string categories needs to be changed to numerical values.

### Changing Categorical Strings Features to Numerical

String Indexer is a PySpark feature that helps change the string categorical values into numerical indices. Using the string indexer feature, the two columns were changed into numerical features.

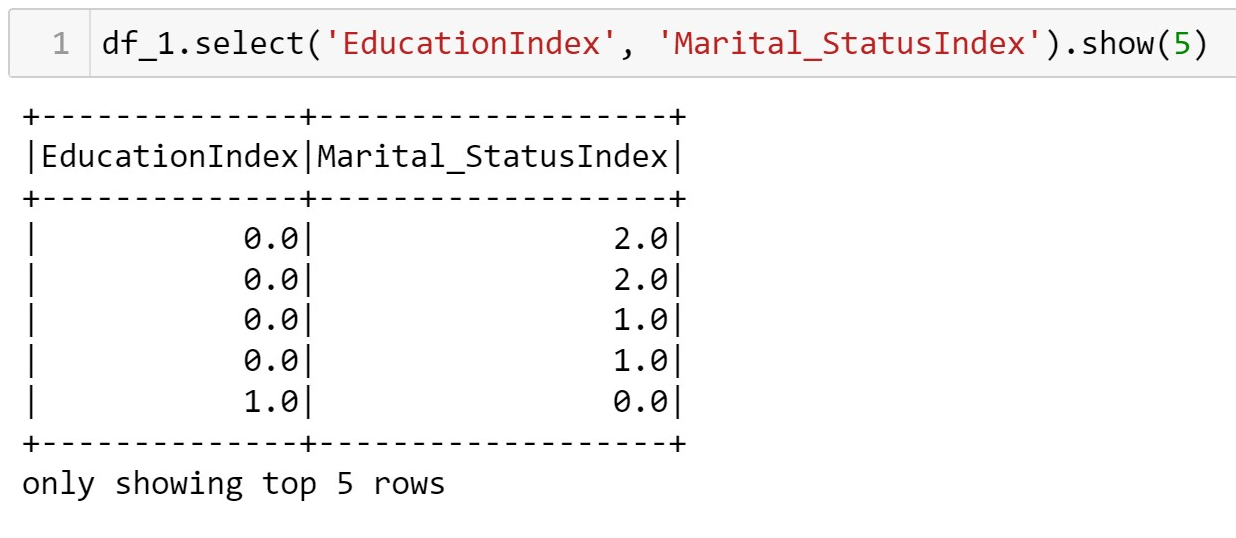


Figure 12 Changing Categorical String columns to Numerical Indices

### Using Vector Assembler to prepare data for creating models

In PySpark, all the columns need to be converted in to a single vector column for it to be passed on to the model. After the string indexing of the categorical string column was done, all the features were merged into one dataframe. The dataframe was then used to create another dataframe which had the vector with all the features assembled.

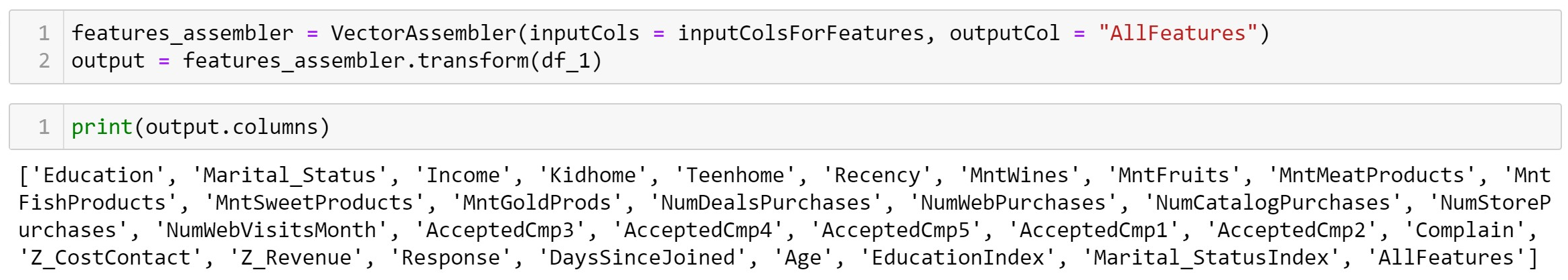


Figure 13 Vector Assembler

A new column named ‘AllFeatures’ can be seen in the columns of the ‘output’ dataframe. It is the vector column created using the vector assembler.

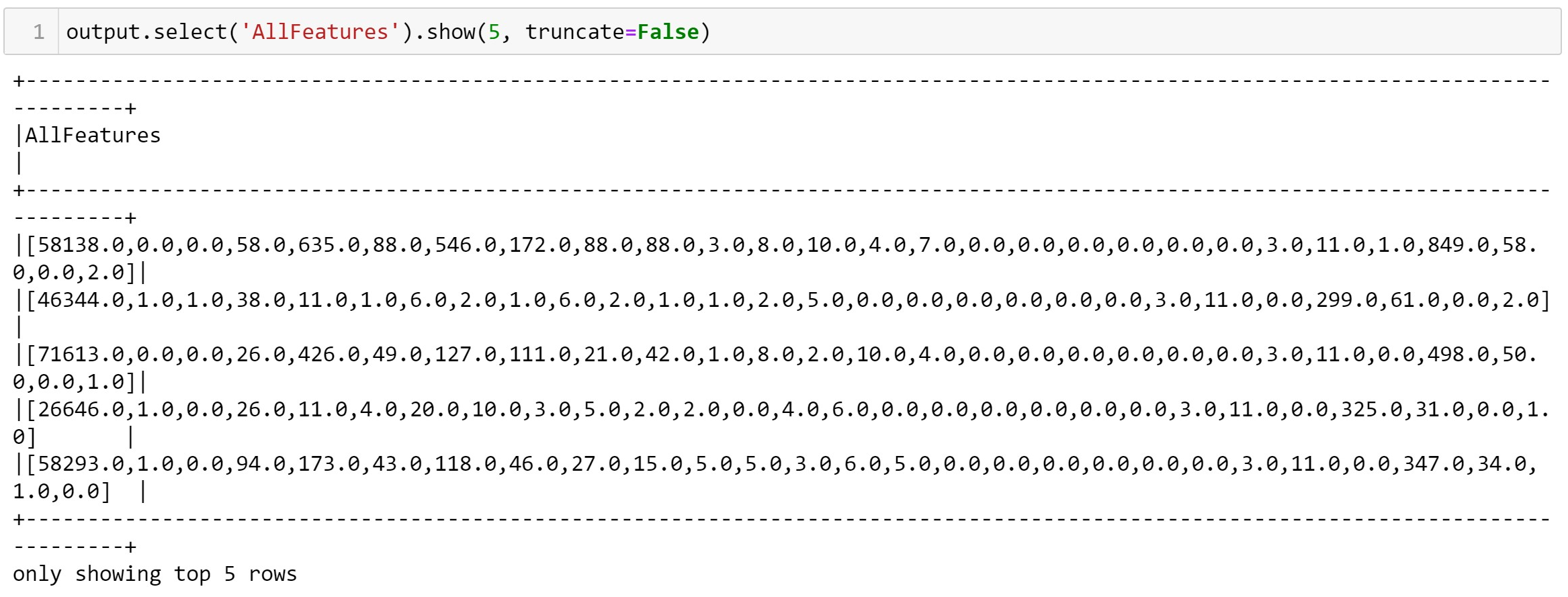


Figure 14 Vector Column created using the Assembler

The ‘AllFeatures’ column will be used as input for the K Means Clustering algorithm which will be discussed next.

## Customer Cluster Analysis Using K-Means Clustering Algorithm

Clustering is an Unsupervised Machine Learning task which means that, clustering algorithms can learn patterns from unlabeled data. These algorithms take the input data and find natural categories in which the data lies. Kmeans clustering is an example of such clustering algorithm.

### What is K-Means Clustering?

The K in K-Means Clustering refers to the numbers of clusters that will be created by the algorithm, where k=3 means there will be 3 clusters created. It is a centroid based algorithm which clusters the data based on the distance of the data point from the centroid. The main aim of this algorithm is to minimize the sum of distances between the centroid and their respective clusters. (javatpoint, n.d.)

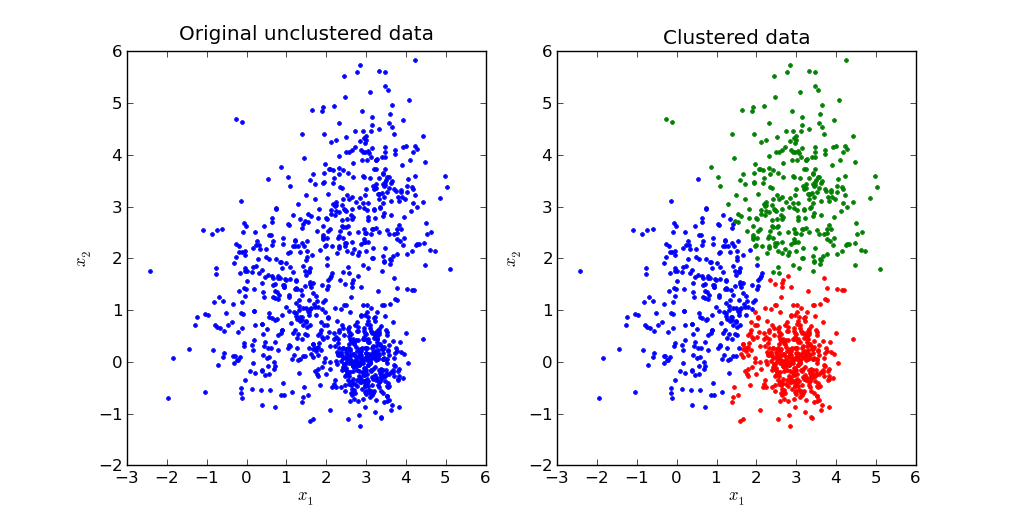


Figure 15 K-Means Clustering (Sybernix, 2017)

For this algorithm to produce meaningful clusters, the number of clusters to be created is really important. One of the ways to find the right value for K is the ‘Silhouette Score’ method. It lies in the range of [-1, 1].

* 1 means that the clusters are well separated and identifiable
* 0 means the clusters are overlapping, and
* Score less than 0 means that the clusters are wrongly assigned.

The Silhouette Score is calculated by using the formula,

Where,

* a = the average distance between each point within a cluster,
* b= the average distance between all clusters.



Figure 16 Silhouette Score calculation (Bhardwaj, 2020)

### Calculation of Silhouette Score for K-Means Clustering

In PySpark, the silhouette score can be calculated using the ‘ClusteringEvaluator’ function. The scores were calcualted making use of the ‘AllFeatures’ vector column calculated earlier.

The scores were as follows:

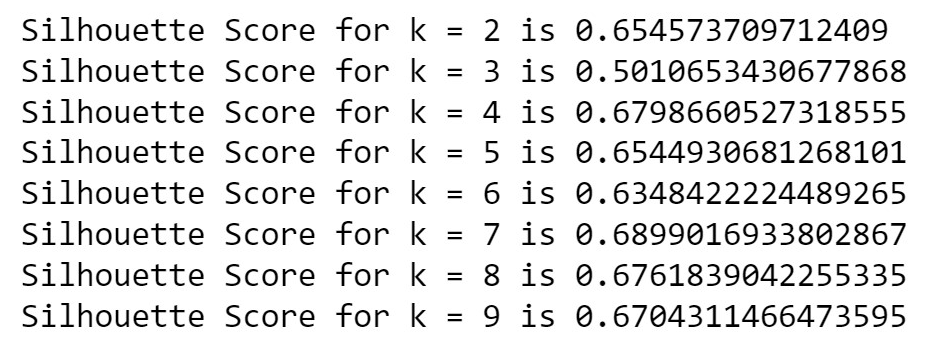


Figure 17 Silhouette Scores

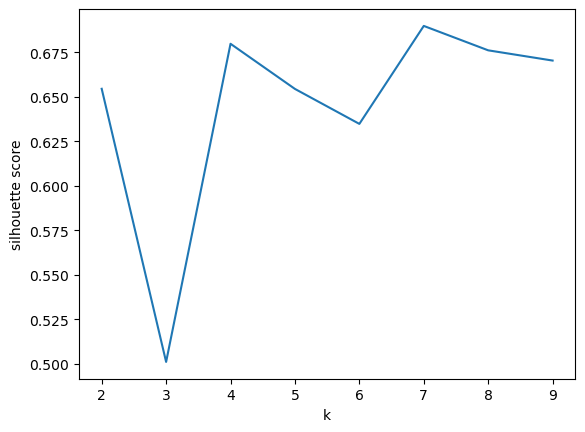


Figure 18 Plot of the silhouette scores

The highest score was 0.6899016933802867 for k = 7. The scores signifies that the clusters are somewhat finely separated. So, k = 7 was finalized for using in the K-Means clustering algorithm.

### Performing Clustering using k = 7

Using the ‘AllFeatures’ vector column created earlier using Vector Assembler, the model was fit using the dataframe, with k = 7 and the predictions were made.

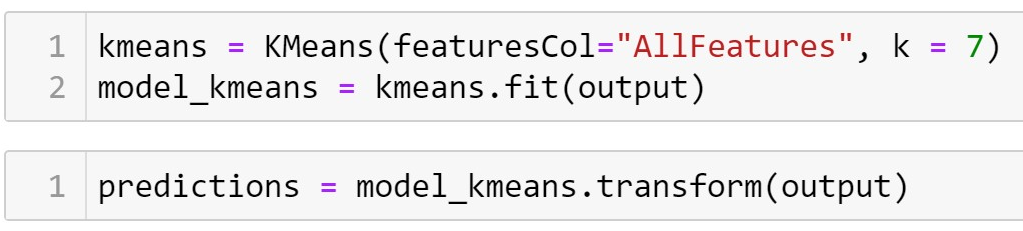


Figure 19 Fitting the K-means model

In total, 7 clusters were created. The following image shows the count of the predictions of the data based on the clusters.

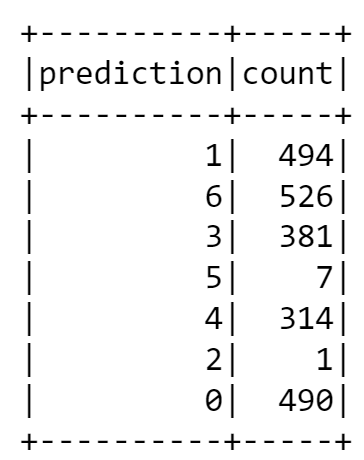


Figure 20 Count of clusters prediction

The details about the clusters and their findings are discussed in the upcoming sections.

### Finding out whether customer will accept the final campaign using Random Forest Classifier algorithm

The aim of this classification was to find out whether a customer would accept the final campaign led by the company based on acceptances of previous campaigns and the customer’s other attributes. For this purpose, the Random Forest Classifier was used.

Random Forest Classifer is one of the most used classifier algorithms. It works on the ensemble learning technique which combines the output of multiple decision trees to provide more accurate results.

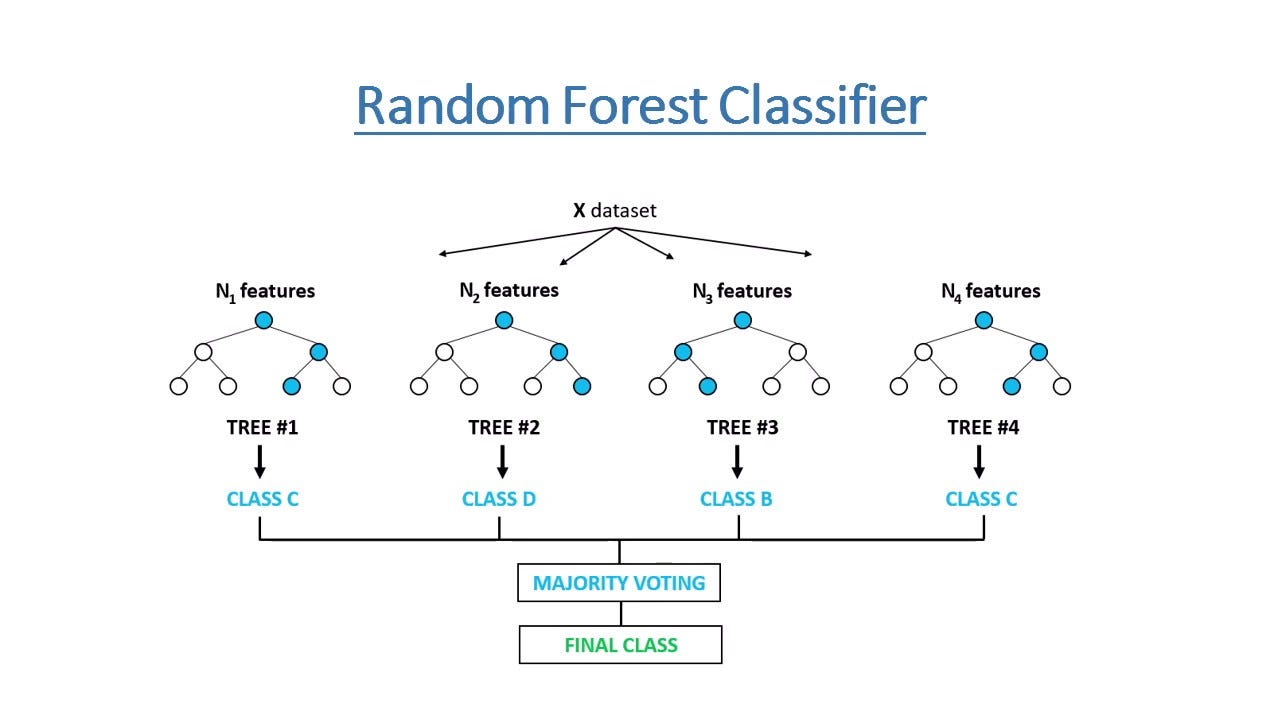


Figure 21 Working of a Random Forest Classifier (Chauhan, 2021)

One of the most important hyperparameter of the Random Forest Classifier model is ‘numTrees’. There is no set rule on what should the number of trees be, but according to the official resource, as random forests don’t overfit, a generous number of tree would work. Here I have chosen it to be 100. (Leo Breiman, n.d.)

Similar to the previous process, a vector containing the features of the dataset was created. But as this is a supervised learning task, the ‘Response’ column was used as the target variable and all remaining columns were transformed to a vector column called ‘Features’ using the Vector Assembler as shown in the figure below.



Figure 22 Creating Features vector

The ‘Feature’ column and the target variable, ‘Response’ column were both then put in to another dataframe.

As it is a Supervise Learning problem, the dataset was then split into training and testing sets in the ratio of 70-30.

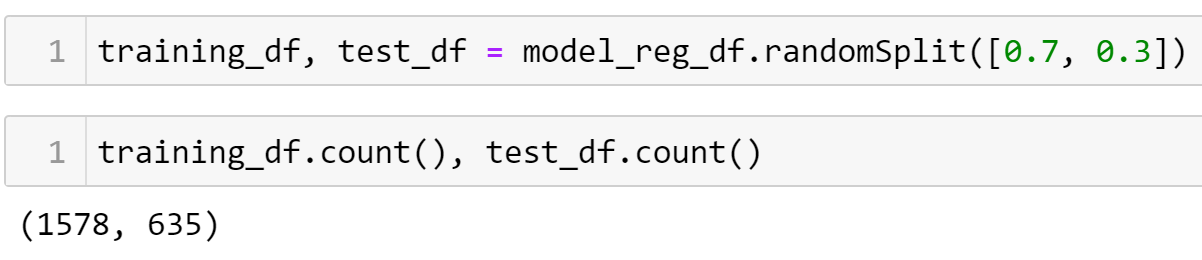


Figure 23 Train - Test Split

The training set was then fit on the random forest classifer model and predictions were made on the testing set.

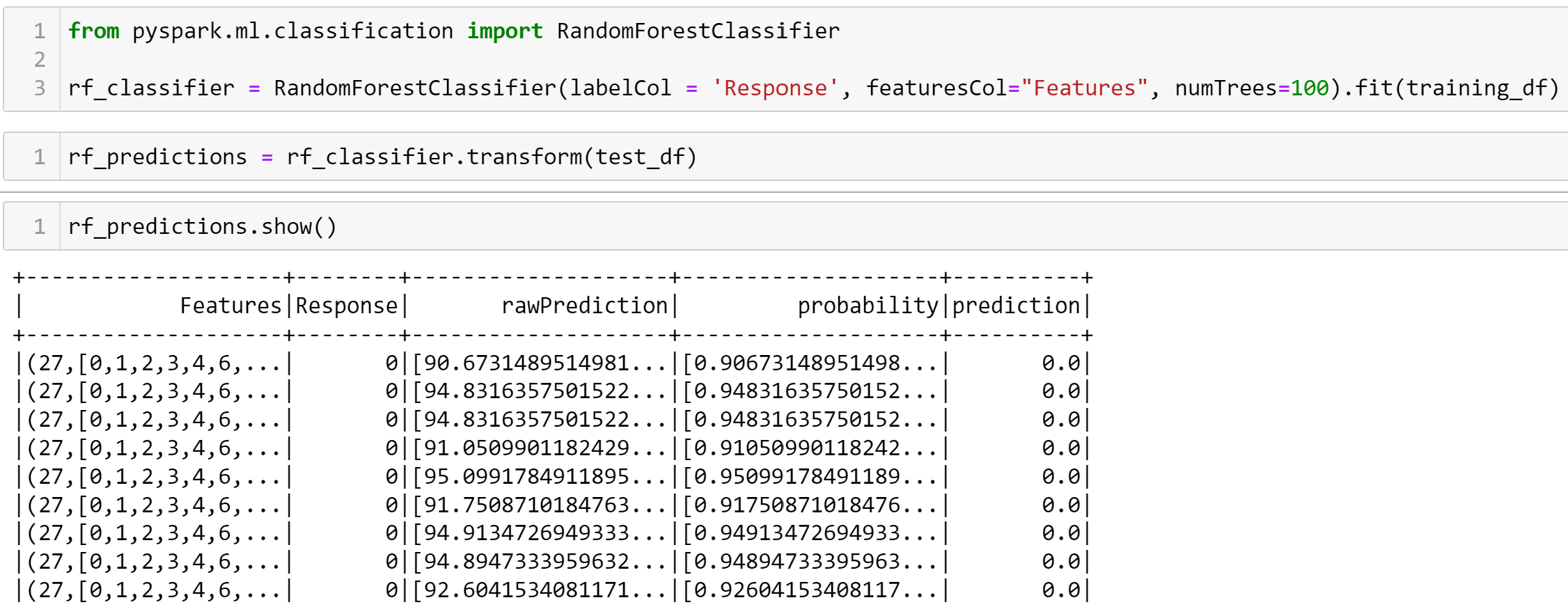


Figure 24 Fitting training set and prediction on test set

The details on the findings of the result is discussed in the next section.

# Discussion of Findings

The figures below have been created using Tableau.

## Findings based on K-Means Clustering

## 

Figure 25 Income vs Total Spendings

In the above figure, the X axis represents the Total Spendings and the Y axis represents the Income of the customers. The colors in the scater plot represents the predicted cluster of the customers. Based on the image it can be concluded that,

* Cluster 3 has income range of 71K – 100K.
* Cluster 4 has the lowest income range of 1K – 28K.
* Cluster 5 has the highest income range of more than 150K. (very few)
* It can also be concluded that most of the times the customers with high income range tend to spend a lot as well. (Cluster 3 - Yellow)
* Customers with low income tend to spend less. (Cluster 4 – Red)

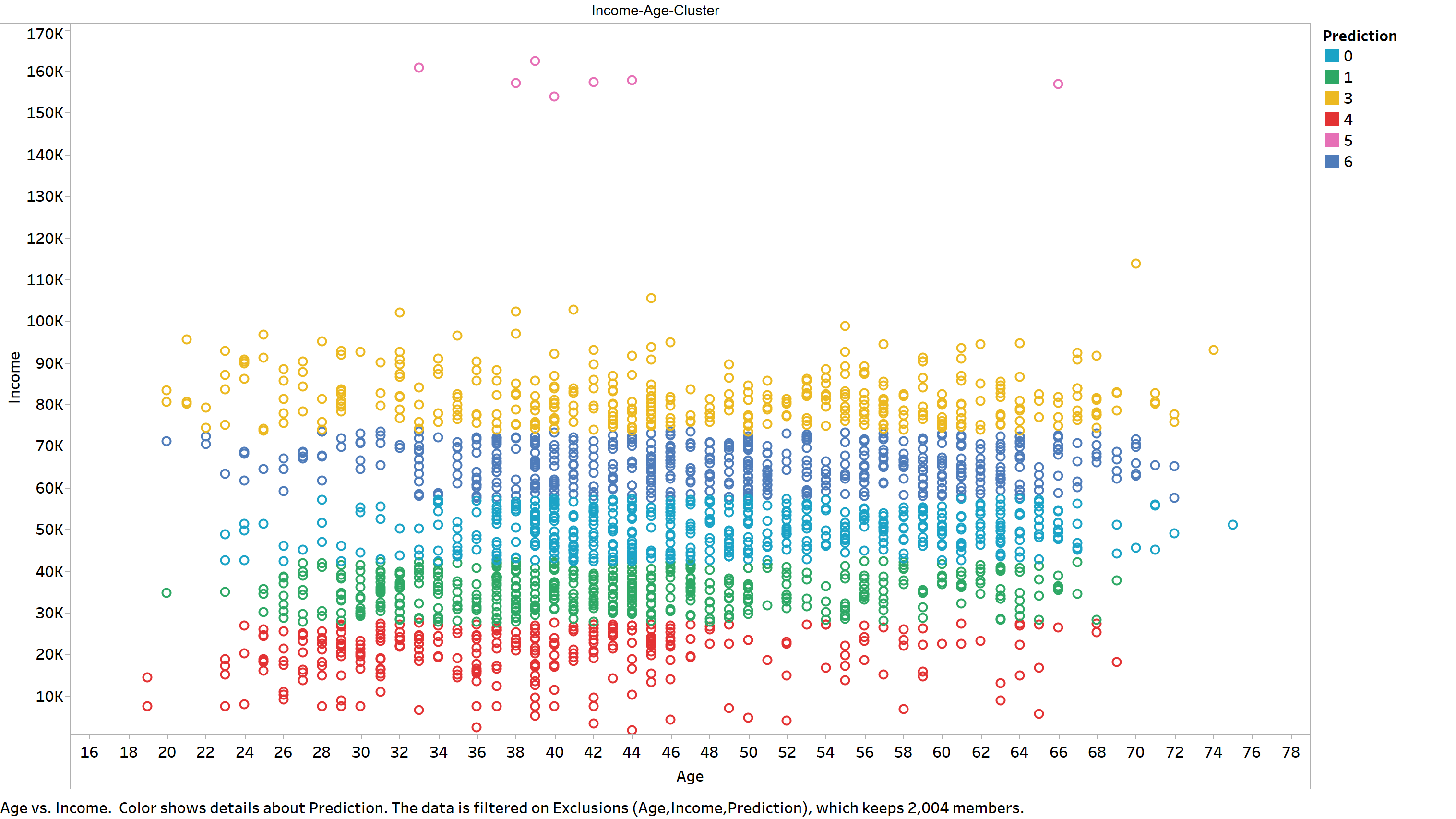


Figure 26 Age vs Income plot with Clusters

The following conclusions can be drawn looking at the above chart:

* Cluster 4 (red) has people with the age 26 – 40.
* Cluster 0 and 6 (sky blue and dark blue) has people of ages 36 – 64.
* Cluster 3 (yellow) mostly has people of age more than 40.

## 

Figure 27 Cluster vs Deals Purchased

* Cluster 0 has the most no. of deals purchased.
* Cluster 2 has the least no. of deals purchased.
* Cluster 6 and 1 has almost similar no. of deals purchased.

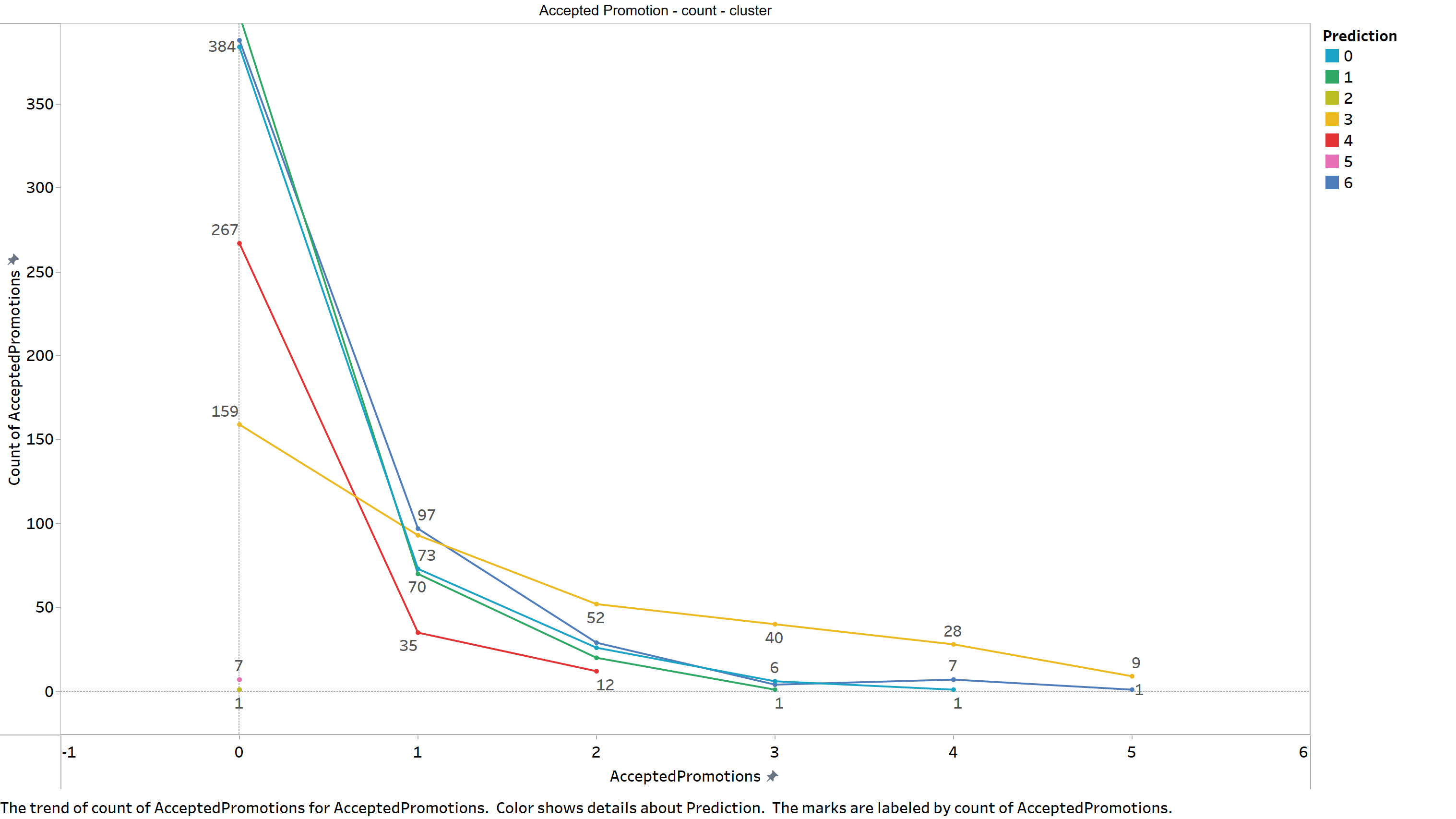


Figure 28 Promotion Acception with Cluster

The above chart shows the count of the promotions accepted based on the clusters. It can be seen that,

* Cluster 1, 6 and 0 were the ones to accept most of the first promotion which drastically dropped down.
* Cluster 4 customers accepted campaigns for the first 3 promotions.
* A general trend can be seen. Most customers tend to accept the promotional offers at first but slowly the number decreases.

Based on all these findings, the following conclusions can be drawn:

1. Cluster 3 (yellow) customers are the richest, more than 40 years old and have high tendencies of accepting all promotional offers.
2. Cluster 6 and 0 customers have high income and aged between 36-64. They have tendencies to accept primary promotional offers and purchase the highest number of deals.
3. Cluster 4 customers are mainly young have lower income range.
4. People who are older mostly have high income and spend more than young customers.

## Findings Based on Random Forest Classification

The accuracy of the predictions made by the Random Forest Classifier can be evaluated using the AUC (Area under the ROC curve) metric. ROC (Receiver operating characteristics) curve is a graph which shows the performace of classification model at all thresholds.

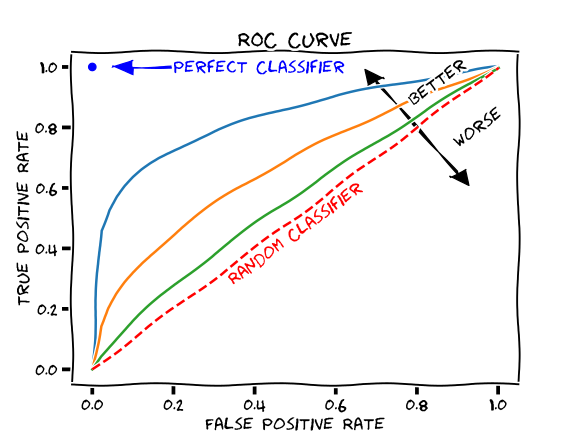


Figure 29 ROC curve (Thoma, 2018)

AUC takes values from 0 to 1. Generally, 0 indicates a perfectly inaccurate case and 1 indicates a perfectly accurate case. A value of 0.8 – 0.9 is considered excellent and more than 0.9 is considered outstanding. (Jayawant N. Mandrekar, 2010)

The **AUC score** of the predictions were calculated using the ‘BinaryClassificationEvaluator’ function, which yielded a value of **0.8545046536438236** which is an excellent score, meaning our model was able to predict with high accuracy.

The feature importance was also calculated for different features. The most important top 3 features were:

1. AcceptedCmp5 (Accepted 5th campaign) – 15% importance
2. AcceptedCmp1 (Accepted 1st Camapign) – 10% importance
3. AcceptedCmp3 (Accepted 3rd Campaign) – 9.5% importance

From the above findings it can be concluded that, to predict whether a customer accpets the final campaign can be predicted using the data related to whether the customer has accepted the previous campaigns.

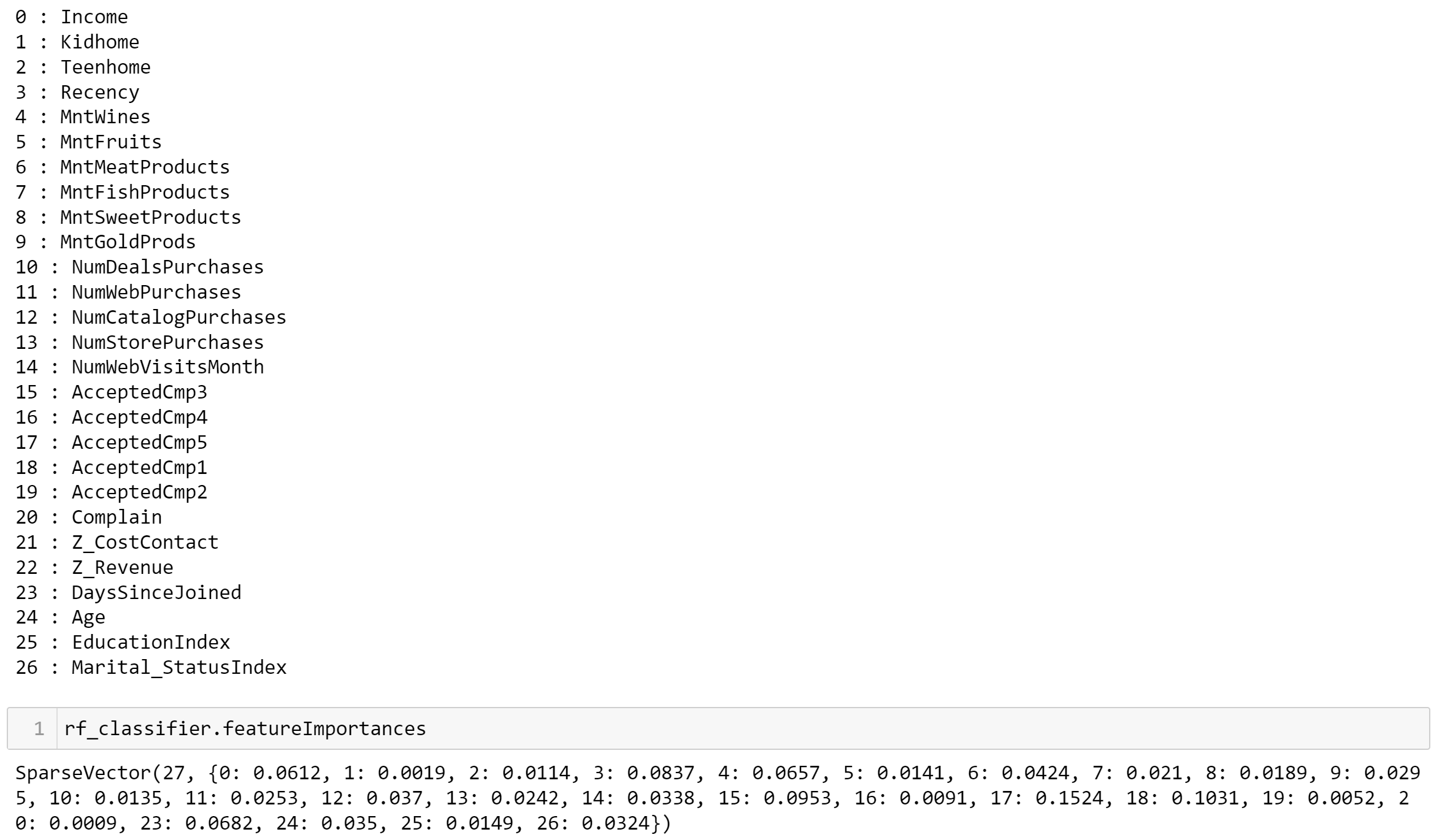


Figure 30 Feature Importances

# Conclusion

Customer analysis is a major part of any business. From the above it can be said that various insights can be gathered about the customers using various machine learning algorithms. In this case the clustering of the customers allowed to:

1. Narrow down the most active and profitable customer cluster,
2. Gain insights on how the customers were reacting to the various iterations of company’s campaigns.

Also, with the help of the classification algorithm model, it could be concluded that whether a customer accepts the final campaign of the company offers. It means that, a company can make campaigns more attractive in order to have more customers accepting the starting offers and the companies campaign offers are actually working. The company can also devise plans to retain the number of customers to accept the final campaign offers.

# References

# Appendix