Data Analytics and Machine Learning Model for Building E-commerce Customer Strategies

Course: MIS 637 Spring 2021

Under The Guidance Of: Prof. Mahmoud Daneshmand

Student Name: Ruchi Hiralal Mendhegiri

Index

Introduction

Retail is one of the fastest-growing and highly competitive sectors across globe. In view of the rapid economy digitalization, e-commerce is becoming one of the most important areas in retail activity.

There are thousands of players in this segment, and it becomes more and more difficult to win the loyalty of new customers and retain the loyalty of regular customers. Therefore, the ability to offer customer centric strategies is key for successful business activity.

Market segmentation can help to define and better understand the target audiences and ideal customers. Market segmentation is the process of splitting buyers into distinct, measurable groups that share similar wants and needs.

Thus, modeling consumer behavior is an actual problem, which solution will not only improve the efficiency of e-commerce but also contribute to the development of the whole economy and better fulfilling of consumers' needs.

In particular, the important task of e-marketing is to classify online store consumers by the level of their purchasing activity. The peculiarities of this task are a large amount of available data and their constant updating and accumulation, which requires the use of Data Science techniques, including machine learning methods.

The goal of this work is to classify online store customers by the level of their purchasing activity based on Data Science techniques, including machine learning methods.

To perform the analysis and all calculations we will use python programming language and Tableau visualization software.

Standard Process : CRISP- DM

1. Business Understanding:

- What is the data mining project we are trying to accomplish?

2. Data Understanding:

- What data can we use?
- Where is the data?
- Familiarize with the data.

3. Data Preparation:

- Perform Transformation on data.
- Clean data from missing values and outliers.
- Select specific data for analysis.

Standard Process : CRISP- DM

4. Modelling :

- Select appropriate data modelling technique.
- Optimize model.
- Define alternative models for project.

5. Evaluation:

- Evaluate the model.
- Determine success rate of model.
- Does model meet project objectives?

6. Deployment:

- Use the model for production data.
- Generate statistics report on data to ensure success.
- Provide business with the deployment instructions for the model so that it can be reused.

Business Understanding

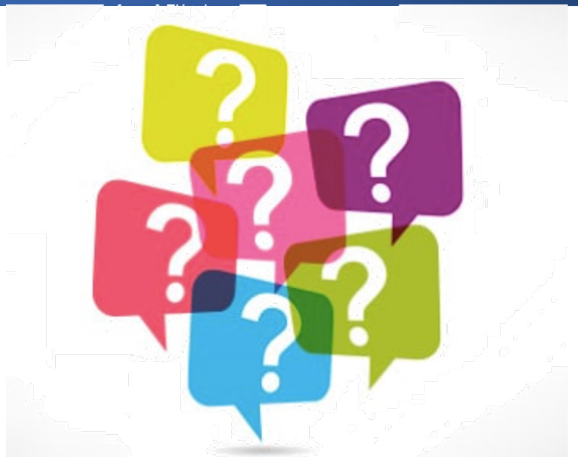
- **What is the Profound Question?**

  Can online purchase history of retail customers be used to categorize and predict customer behaviour

- **Goals:**
  1. Customer segmentation based on recency, frequency and monetary value for customer purchases
  2. Predict customer behaviour and categorising into different segments using classification techniques

- **What will be the Accomplishments?**

  Customer segmentation will help in better understanding of customer behaviour and develop effective customer centric sales and marketing strategies



Data Understanding

Data Source:

https://archive.ics.uci.edu/ml/datasets/Online+Retail+II#

Data Set Details:

- This data set contains all the transactions occurring for a UK-based and registered, non-store online retail between 01/12/2009 and 09/12/2011. The company mainly sells unique all-occasion gift-ware. Many customers of the company are wholesalers.

- Number of Instances: 1067371
- Number of Attributes: 8

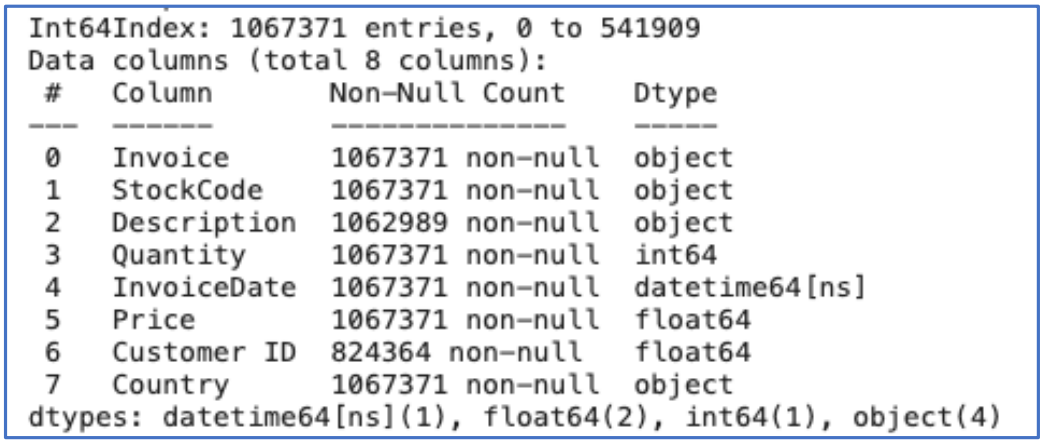
Data Understanding : Sample Data

Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	01/12/09 7:45	6.95	13085	United Kingdom
489434	79323P	PINK CHERRY LIGHTS	12	01/12/09 7:45	6.75	13085	United Kingdom
489434	79323W	WHITE CHERRY LIGHTS	12	01/12/09 7:45	6.75	13085	United Kingdom
489434	22041	RECORD FRAME 7" SINGLE SIZE	48	01/12/09 7:45	2.1	13085	United Kingdom
489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	01/12/09 7:45	1.25	13085	United Kingdom
489434	22064	PINK DOUGHNUT TRINKET POT	24	01/12/09 7:45	1.65	13085	United Kingdom
489434	21871	SAVE THE PLANET MUG	24	01/12/09 7:45	1.25	13085	United Kingdom
489434	21523	FANCY FONT HOME SWEET HOME DOORMAT	10	01/12/09 7:45	5.95	13085	United Kingdom
489435	22350	CAT BOWL	12	01/12/09 7:46	2.55	13085	United Kingdom
489435	22349	DOG BOWL , CHASING BALL DESIGN	12	01/12/09 7:46	3.75	13085	United Kingdom
489435	22195	HEART MEASURING SPOONS LARGE	24	01/12/09 7:46	1.65	13085	United Kingdom
489435	22353	LUNCHBOX WITH CUTLERY FAIRY CAKES	12	01/12/09 7:46	2.55	13085	United Kingdom
489436	48173C	DOOR MAT BLACK FLOCK	10	01/12/09 9:06	5.95	13078	United Kingdom
489436	21755	LOVE BUILDING BLOCK WORD	18	01/12/09 9:06	5.45	13078	United Kingdom
489436	21754	HOME BUILDING BLOCK WORD	3	01/12/09 9:06	5.95	13078	United Kingdom
489436	84879	ASSORTED COLOUR BIRD ORNAMENT	16	01/12/09 9:06	1.69	13078	United Kingdom
489436	22119	PEACE WOODEN BLOCK LETTERS	3	01/12/09 9:06	6.95	13078	United Kingdom
489436	22142	CHRISTMAS CRAFT WHITE FAIRY	12	01/12/09 9:06	1.45	13078	United Kingdom
489436	22296	HEART IVORY TRELLIS LARGE	12	01/12/09 9:06	1.65	13078	United Kingdom
489436	22295	HEART FILIGREE DOVE LARGE	12	01/12/09 9:06	1.65	13078	United Kingdom
489436	22109	FULL ENGLISH BREAKFAST PLATE	16	01/12/09 9:06	3.39	13078	United Kingdom
489436	22107	PIZZA PLATE IN BOX	4	01/12/09 9:06	3.75	13078	United Kingdom
489436	22194	BLACK DINER WALL CLOCK	2	01/12/09 9:06	8.5	13078	United Kingdom
489436	35004B	SET OF 3 BLACK FLYING DUCKS	12	01/12/09 9:06	4.65	13078	United Kingdom
489436	82582	AREA PATROLLED METAL SIGN	12	01/12/09 9:06	2.1	13078	United Kingdom
489436	21181	PLEASE ONE PERSON METAL SIGN	12	01/12/09 9:06	2.1	13078	United Kingdom

Data Understanding : Attribute Details

Sr. No	Attribute Name	Type	Description
1	Invoice	Nominal	• A 6-digit integral number uniquely assigned to each transaction. • If this code starts with the letter 'c', it indicates a cancellation.
2	StockCode	Nominal	• A 5-digit integral number uniquely assigned to each distinct product.
3	Description	Nominal	• The name of the product
4	Quantity	Numeric	• The quantity of purchased/returned products
5	InvoiceDate	Numeric	• Invoice date and time. • The day and time when a transaction was generated.
6	Price	Numeric	• The price of the goods (per unit)
7	Customer ID	Nominal	• A 5-digit integral number uniquely assigned to each customer.
8	Country	Nominal	• The name of the country where a customer resides.

Data Understanding

Available Data	
Unique Number of Items	4629
Unique Customer ID	5873
Unique Countries	41
Price Range	$0 - $10953.5

Data Preparation : Data Cleaning

Handle Missing Values & Identify Outliers:	1. Customer Description and Customer ID had many null values 2. Records with null values have been removed 3. No significant outliers detected in the dataset
Look For Attributes That Might Need Transformation:	1. The Invoice Number converted into Integer format 2. Min Max normalization is used for Feature scaling
Select Attributes Appropriate For Analysis:	1. No irrelevant attributes 2. Returned records has been removed

Total Number of Missing Values:

```
Invoice             0
StockCode           0
Description      4382
Quantity            0
InvoiceDate         0
Price               0
Customer ID    243007
Country             0
dtype: int64
```

Data Preparation : Cleaned Data

	Quantity	Price	Customer ID
count	791007.000000	791007.000000	791007.000000
mean	13.306012	3.199388	15332.479132
std	145.380324	29.037699	1696.986733
min	1.000000	0.000000	12346.000000
25%	2.000000	1.250000	13982.000000
50%	5.000000	1.950000	15271.000000
75%	12.000000	3.750000	16805.000000
max	80995.000000	10953.500000	18287.000000

```
Int64Index: 791007 entries, 0 to 541909
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   Invoice      791007 non-null  object
 1   StockCode    791007 non-null  object
 2   Description  791007 non-null  object
 3   Quantity     791007 non-null  int64
 4   InvoiceDate  791007 non-null  datetime64[ns]
 5   Price        791007 non-null  float64
 6   Customer ID  791007 non-null  int64
 7   Country      791007 non-null  object
dtypes: datetime64[ns](1), float64(1), int64(2), object(4)
```

Modelling

Modelling : Goal 1

- The first part will be focused on the e-commerce customer segmentation by cluster analysis methods (k-means clustering) on different criteria.

- On the basis of the Recency, Frequency, and Monetary model, customers will be segmented into various meaningful groups using the *k*-means clustering algorithm .

- Using Elbow method, the optimal k for clustering algorithm can be automatically calculated.

RMF Analysis

The first step in the customer segmentation is the selection of criteria for evaluating the level of customer purchasing activity:

- We will take a classic RFM approach to measure purchasing (Recency - Frequency - Monetary).
- **Recency** for each individual customer is calculated as the difference between the actual date in the database and the date of the customer's last purchase.

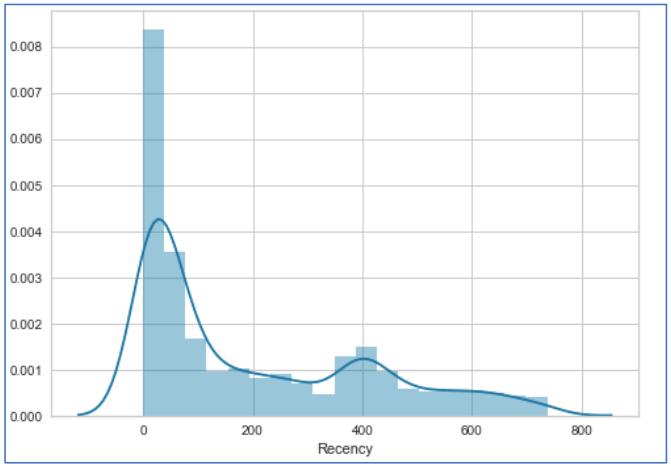
    - In our case, this metric is measured in days.

- **Frequency** of purchases for each individual customer is determined by the number of transactions performed by the customer during his customer life(2009-2011).
- **Monetary** for each individual customer is defined as the total return on all customer transactions during his or her customer life.

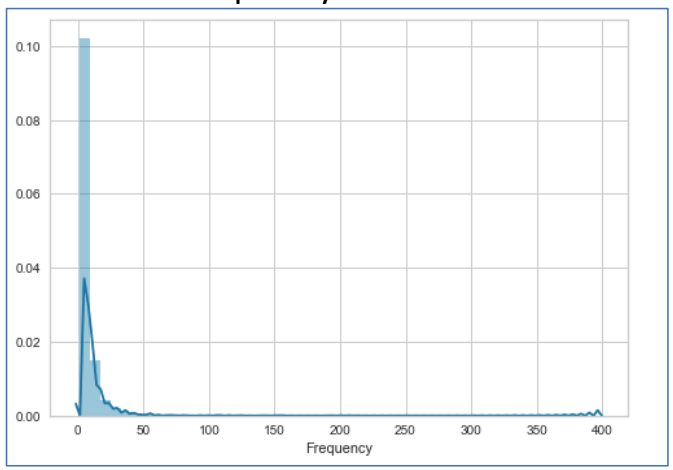
    - In our case, this metric is measured in dollars.

Distribution Plot

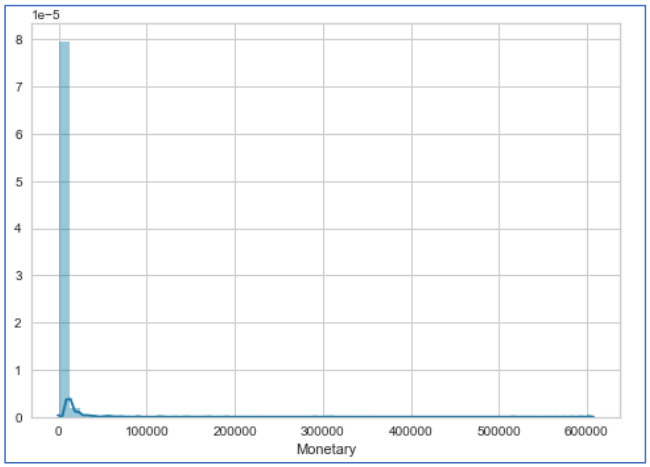
Recency Distribution



Frequency Distribution



Monetary Distribution



- The Distribution plot for Recency, frequency and monetary values **shows** the frequency of values in data by grouping it into equal-sized intervals (bins) .
- It gives idea about the approximate probability **distribution** of our 3 basic quantitative parameters of RFM model

RFM Analysis Results

These RFM - customer activity metrics were calculated for each customer in the sample.

	Recency	Frequency	Monetary
count	5874.000000	5874.000000	5874.000000
mean	200.165475	6.267790	2965.031792
std	209.632592	12.946219	14539.420660
min	-1.000000	1.000000	0.000000
25%	24.000000	1.000000	342.865000
50%	94.000000	3.000000	880.020000
75%	379.000000	7.000000	2267.542500
max	737.000000	397.000000	606599.490000

Distribution Of Purchasing Activity Metrics

- Average customer of our online store had the last purchase 200 days ago.
- On average the customers buy 6 goods during the client's life while spending $ 2965.
- The highest value customer has spent the total amount of $606599 at our online store.
- The highest frequency that customer has purchased at online store is 397 .

K-means Methodology

- The process of clustering an online store's customer base relates to Unsupervised Learning algorithms .
- The main goal of the implementation of these algorithms is to find certain patterns in the available data and to characterize their structure.
- The most efficient and simple algorithm for cluster analysis is k-means.
- **Procedure of K Means Clustering:**
  - It involves multiple executions of an algorithm with different random assignment of initial centroids was applied. An iteration with a minimum value of W_{total} is selected as the final clustering option.
  - Within Cluster Sum of Squares (W_{total}) measures the squared average distance of all the points within a cluster to the cluster centroid

$$W_{total} = \sum_{i=1}^{m} \frac{\varrho\,(c^i)}{n^i}$$

where (ϱc^i) is the sum of Euclidean distances between points within the cluster l

n^i number of points in cluster l; m is the total number of clusters.

  - The sum of Euclidean distances between points within cluster l is calculated by the formula:
    - $\varrho c^i = \sum_{i=1}^{n} \varrho\,(x^i, c^i)$

where n is the number of points in cluster l; is the cent of the weight of cluster l

Data Preparation : Min- Max Normalization

- The three variables are not on comparable scales, and the value ranges are quite different: Recency [0 to 737]; Frequency [0,397] and Monetary [0 to 606599], respectively. As such, these variables should be normalized before the clustering analysis.
- **Min–max normalization:**
  - It performs a linear transformation on the original data. This technique gets all the scaled data in the range [0,1]. The formula to achieve this is the following:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

  - Min-max normalization preserves the relationships among the original data values. The cost of having this bounded range is that we will end up with smaller standard deviations, which can suppress the effect of outliers.

K-means Methodology:

The process of clustering an online store's customer base relates to Unsupervised Learning algorithms.

The main goal of the implementation of these algorithms is to find certain patterns in the available data and to characterize their structure.

The most efficient and simple algorithm for cluster analysis is k-means.

Procedure of K Means Clustering:

- It involves multiple executions of an algorithm with different random assignment of initial centroids was applied.
- An iteration with a minimum value of Wtotal is selected as the final clustering option.
- Within Cluster Sum of Squares (W_{total}) measures the squared average distance of all the points within a cluster to the cluster centroid :

$$W_{total} = \sum_{i=1}^{m} \frac{\varrho\,(c^i)}{n^i}$$

- where (ϱc^i) is the sum of Euclidean distances between points within the cluster l

- n^i number of points in cluster l; m is the total number of clusters.

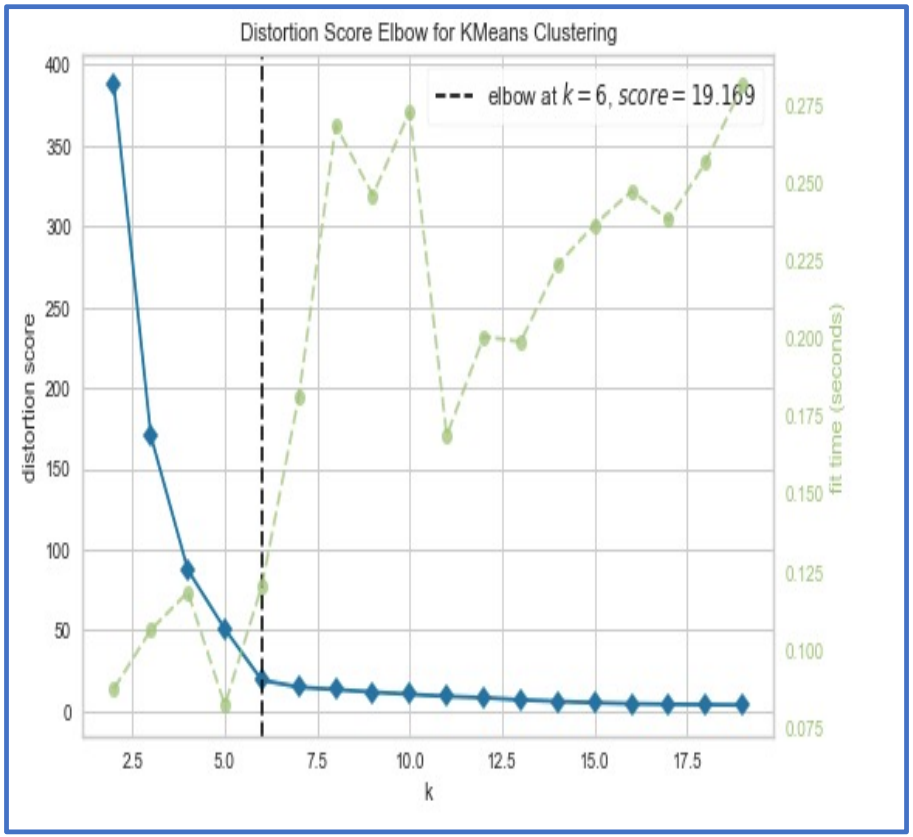
- The sum of Euclidean distances between points within cluster l is calculated by the formula:

$$\varrho c^i = \sum_{i=1}^{n} \varrho\,(x^i\,, c^i)$$

where n is the number of points in cluster l; is the cent of the weight of cluster l

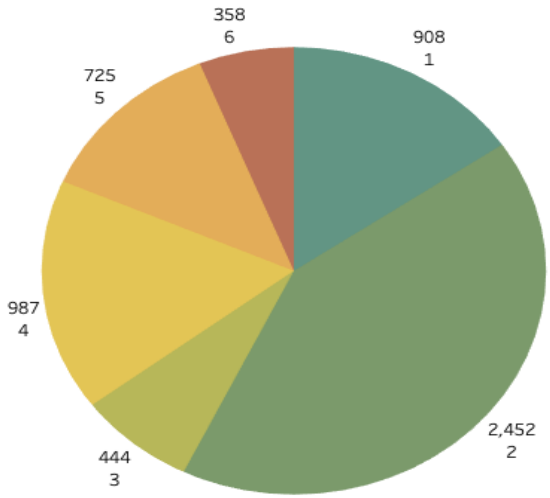
K-means Methodology: Finding Optimal K

- The second problem is the need to prioritize a fixed number of clusters for partitioning, which is certainly not always chosen to be optimal.
- The elbow method explores the nature of the (W_{total}) variation spread with an increasing number of groups k. Combining all n observations into one group, we have the largest intra-cluster variance, which will decrease to 0 as k → n .
- **The Optimal Number of Clusters obtained : K=6**



Customer Segmentation Results:

cluster_no	Number of Customers	(Monetary, mean)	(Frequency, mean)	(Recency, mean)
1	908	1038.747358	2.737885	400.961454
2	2452	5380.648254	10.493883	22.577080
3	444	706.084212	2.063063	534.943694
4	987	1861.448455	4.831814	104.389058
5	725	1273.588126	3.364138	243.462069
6	358	575.313469	1.329609	668.388268



Cluster 1 : Retail buyer
High average monetary value and above average activity, these customers buys over a long period.

Cluster 2 : Active Loyal Customers
High activity, buys for high monetary value. The most valuable and loyal type of customers for the business.

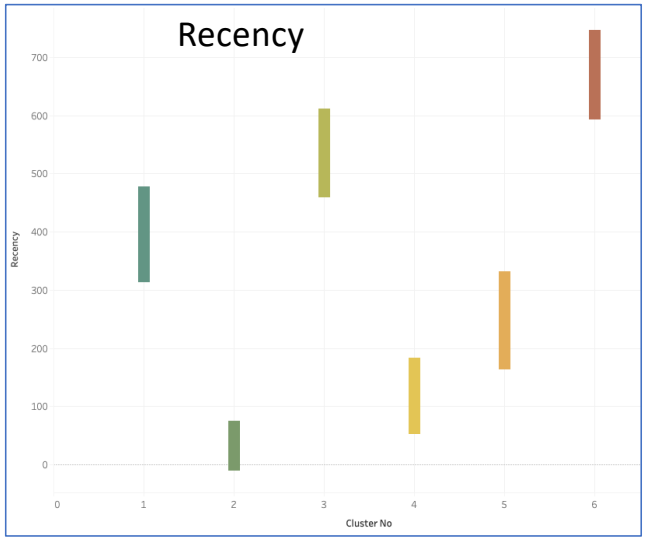
Cluster 3 : Threshold Buyer
Moderate average check but its more than one and half year since the last the purchase. Efforts must be made to increase customer loyalty to the business.

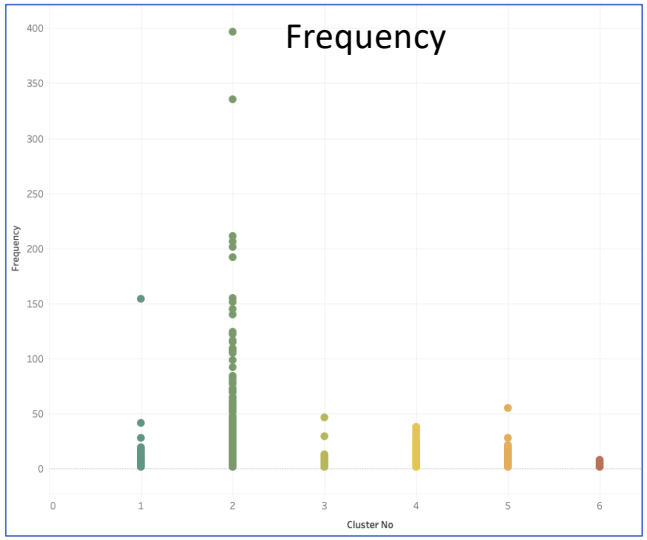
Cluster 4 : New buyers
Recently joined customers with above average frequency and above average monetary value. Need to work on these customers to convert them to loyal customers.

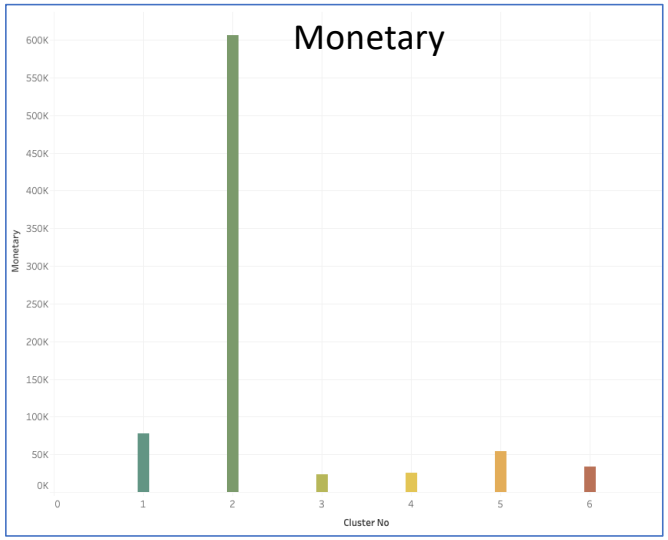
Customer Segmentation Results:

Cluster 5 : Occasional Buyer
Once a while purchases higher value items. Need to build strategies in order to increase purchase frequencies of these buyers.

Cluster 6 : Lost Cheap Clients
Has made less than 2 purchases, the last of which was more than two year ago. Special efforts needs to be made in order to win back these customers.







Modelling : Goal 2

- The second part will be developing a customer classification algorithm which can used to update current customer segment as well as assign a segment to the future customers.

- For classification task, multiple models will be developed using different algorithms (CART, KNN and Random forest) and based on accuracy and relevancy the best fitted algorithm will be selected for model implementation.

Data Preparation : Data Division

- After data clean, the data set consisting of 5,874 clients which is divided into 2 sets.
- **Training Data Set:**

  70%of the data (4,112 customers) is used to develop the model.
- **Testing Data:**

  30% of the data (1,762 customers) is used to evaluate the model.

KNN Classification

Principle :

- The KNN algorithm uses '**feature similarity**' to predict the values of any new data points.
- This means that the new point is assigned a value based on how closely it resembles the points in the training set.

Algorithm Steps :

- The distance between the objects may be a measure of similarity. The **first step** is to calculate the distance between the new point and each training point.
- There are various methods for calculating this distance, of which the most commonly known methods are : Euclidian, Manhattan (for continuous) and Hamming distance (for categorical).
- The **second step** is to select the k value. This determines the number of neighbours we look at when we assign a value to any new observation.

KNN Classification : Finding Distances

Euclidean Distance:
Euclidean distance is calculated as the square root of the sum of the squared differences between a new point (x) and an existing point (y).

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Manhattan Distance:
This is the distance between real vectors using the sum of their absolute difference.

$$\sum_{i=1}^{k}|x_i - y_i|$$

Hamming Distance:
It is used for categorical variables. If the value (x) and the value (y) are the same, the distance D will be equal to 0 . Otherwise D=1.

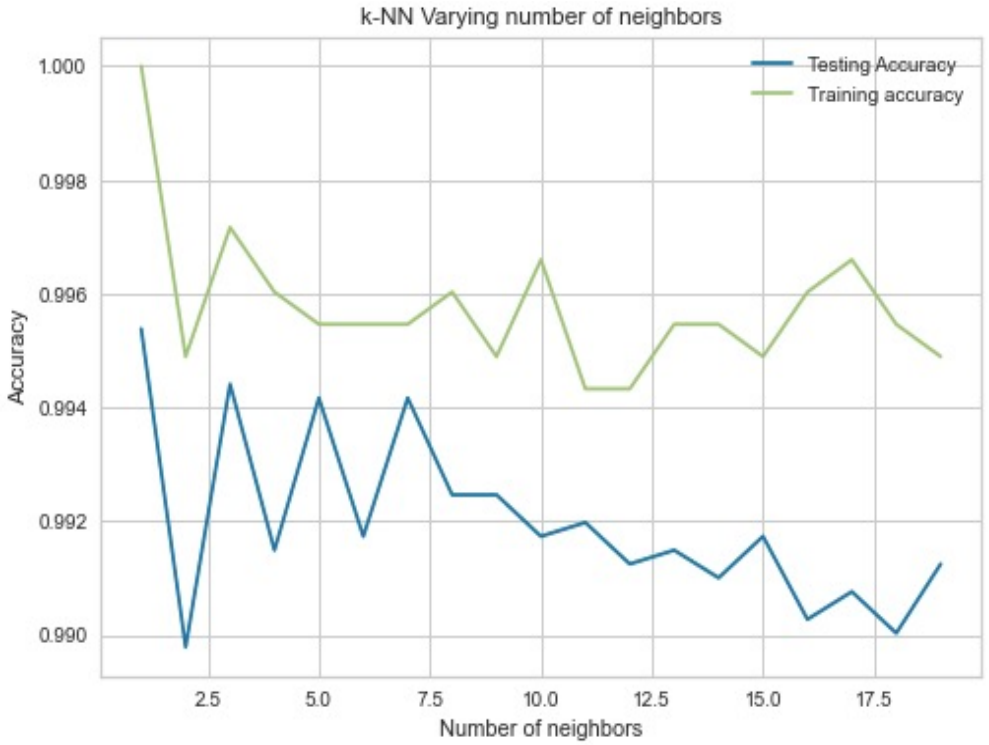
$$D_H = \sum_{i=1}^{k}|x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

Finding Optimal K for KNN Algorithm

- For a very low value of k (suppose k=1), the model overfits on the training data, which leads to a high error rate on the validation set.
- For a high value of k, the model performs poorly on both train and validation set.
- We will decide value of k based on the error calculation for our train and validation set as minimizing the error is our final goal.
- the validation error curve reaches a minima at a value of **k = 7**. This value of k is the optimum value of the model.
- This curve is known as an '**elbow curve**' (because it has a shape like an elbow) and is usually used to determine the k value.



Decision Tree Classification (CART)

The CART algorithm is structured as a sequence of questions, the answers to which determine what the next question, if any should be. The result of these questions is a tree like structure where the ends are terminal nodes at which point there are no more questions.

The main elements of CART are:

Rules for splitting data at a node based on the value of one variable;	Stopping rules for deciding when a branch is terminal and can be split no more; and	Finally, a prediction for the target variable in each terminal node.

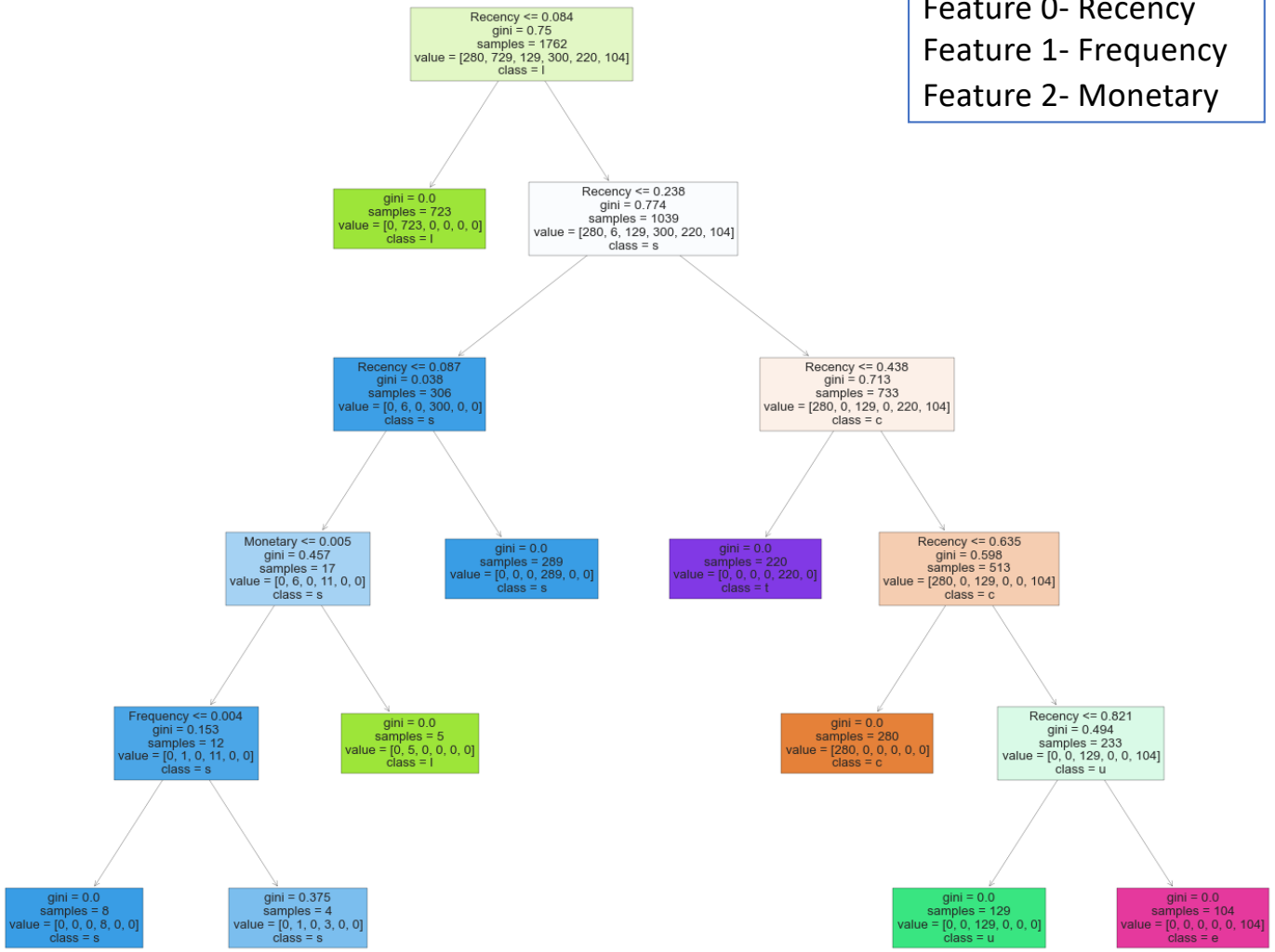
We will carry out hyperparameter tunning for:

- To find out the maximum depth allowed for the decision tree.
- What should be the minimum number of samples required at a leaf node in my decision tree

CART Decision Rules:

Condition	Class
If Recency <=0.08	Class 2
If Recency > 0.08 If Recency<= 0.24 If Recency<= 0.09 If Monetary <= 0.01 If Frequency <= 0.00	Class 4
If Recency > 0.08 If Recency<= 0.24 If Recency<= 0.09 If Monetary >0.01	Class 2
If Recency > 0.08 If Recency<= 0.24 If Recency<= 0.09 If Monetary <= 0.01 If Frequency > 0.00	Class 4
If Recency > 0.08 If Recency<= 0.24 If Recency > 0.09	Class 4
If Recency > 0.08 If Recency > 0.24 If Recency <= 0.44	Class 5
If Recency > 0.08 If Recency > 0.24 If Recency >0.44 Recency <= 0.63	Class 1
If Recency > 0.08 If Recency > 0.24 If Recency >0.44 Recency > 0.63 Recency <= 0.82	Class 3
If Recency > 0.08 If Recency > 0.24 If Recency >0.44 Recency > 0.63 Recency > 0.82	Class 6

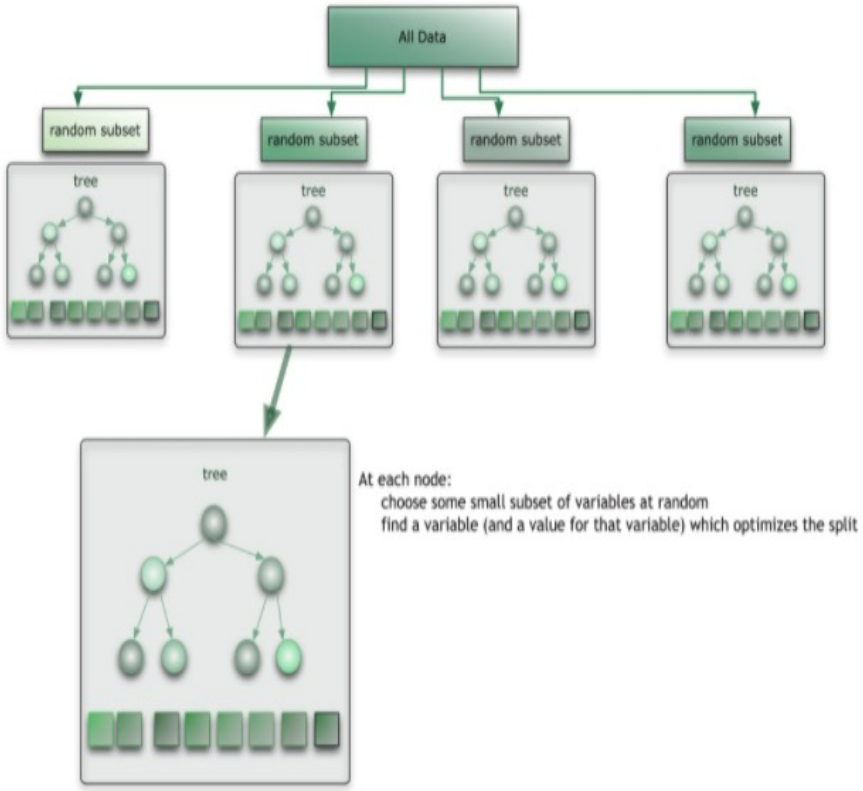
CART Results



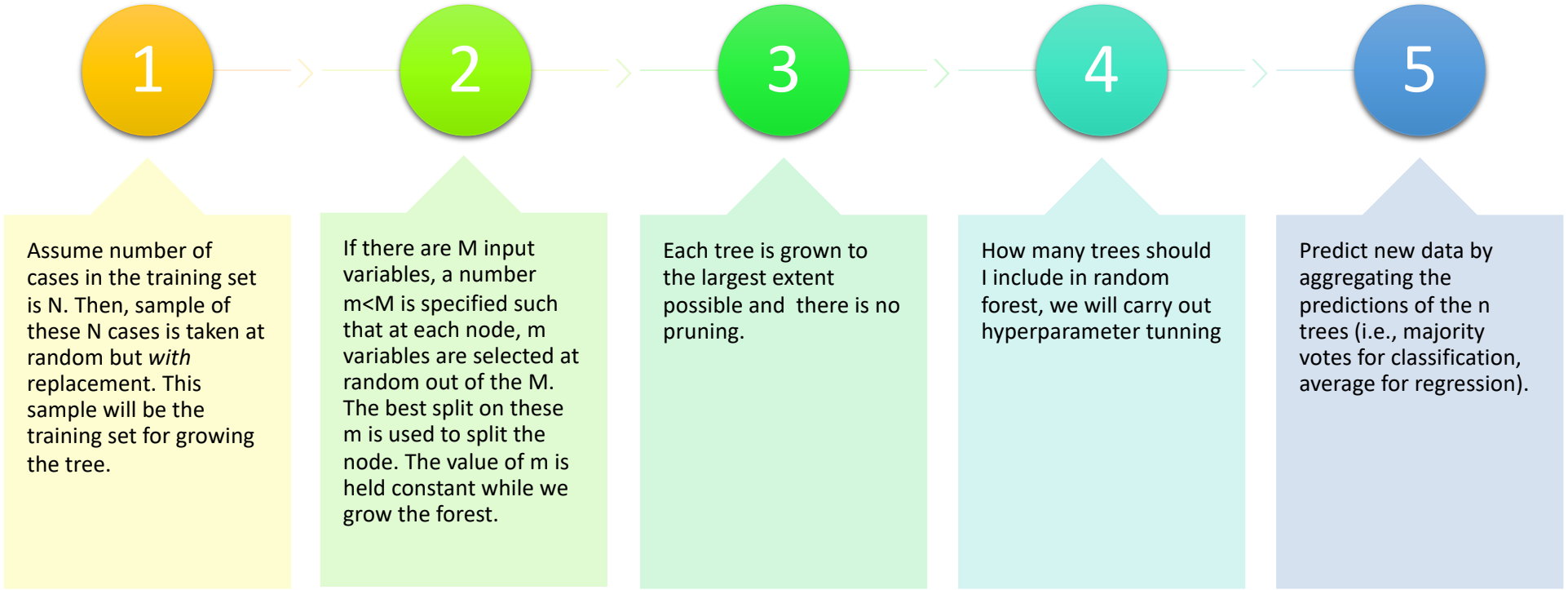
Feature 0- Recency
Feature 1- Frequency
Feature 2- Monetary

Random Forest Classification

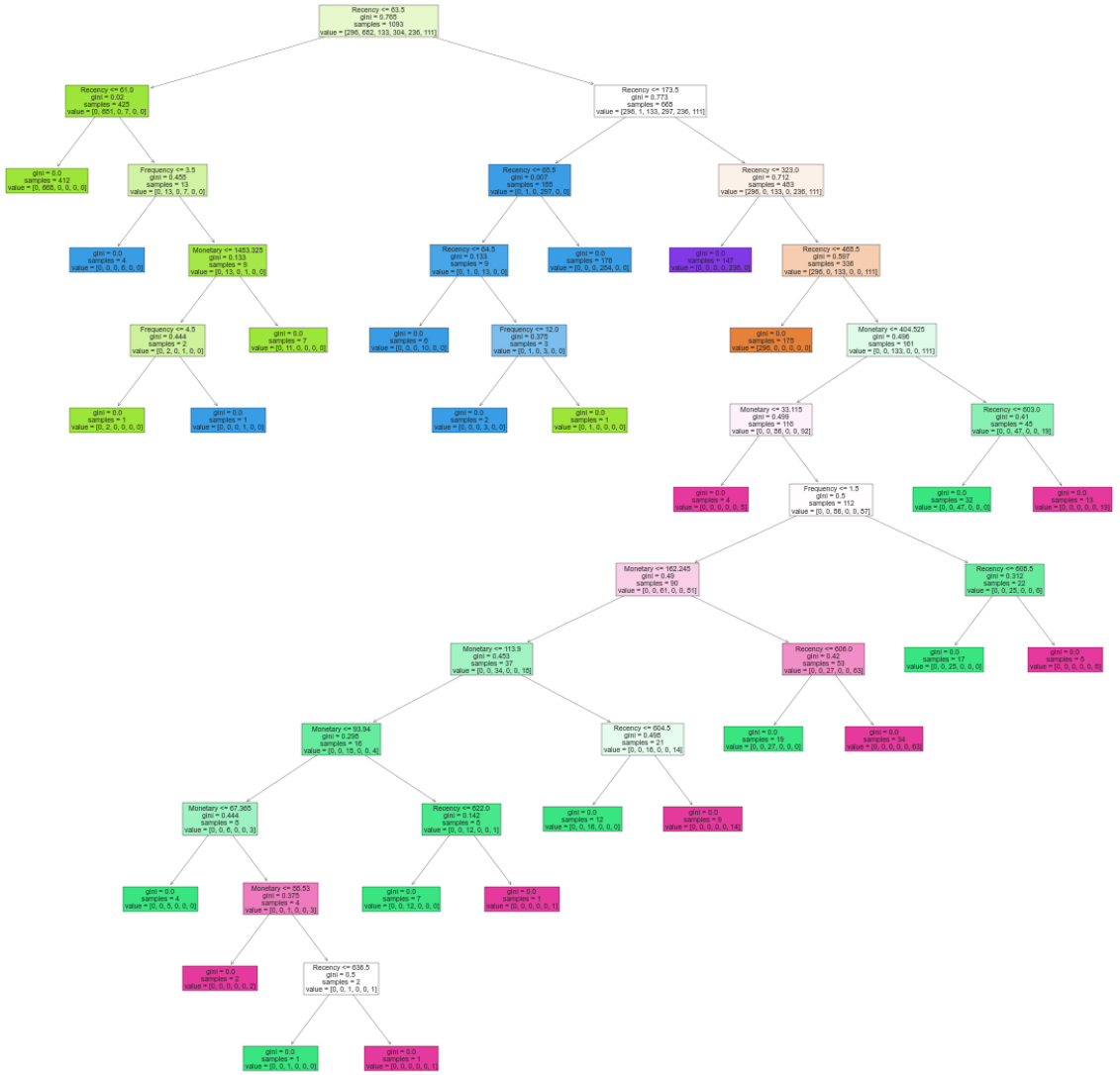
- Random Forest (RF) is an ensemble model that builds several trees and classifies objects on a "vote" basis.
- The object belongs to the class that has the majority of votes from all the trees.
- The algorithm trains several decision trees on different subsamples of data and uses the average to improve model prediction accuracy



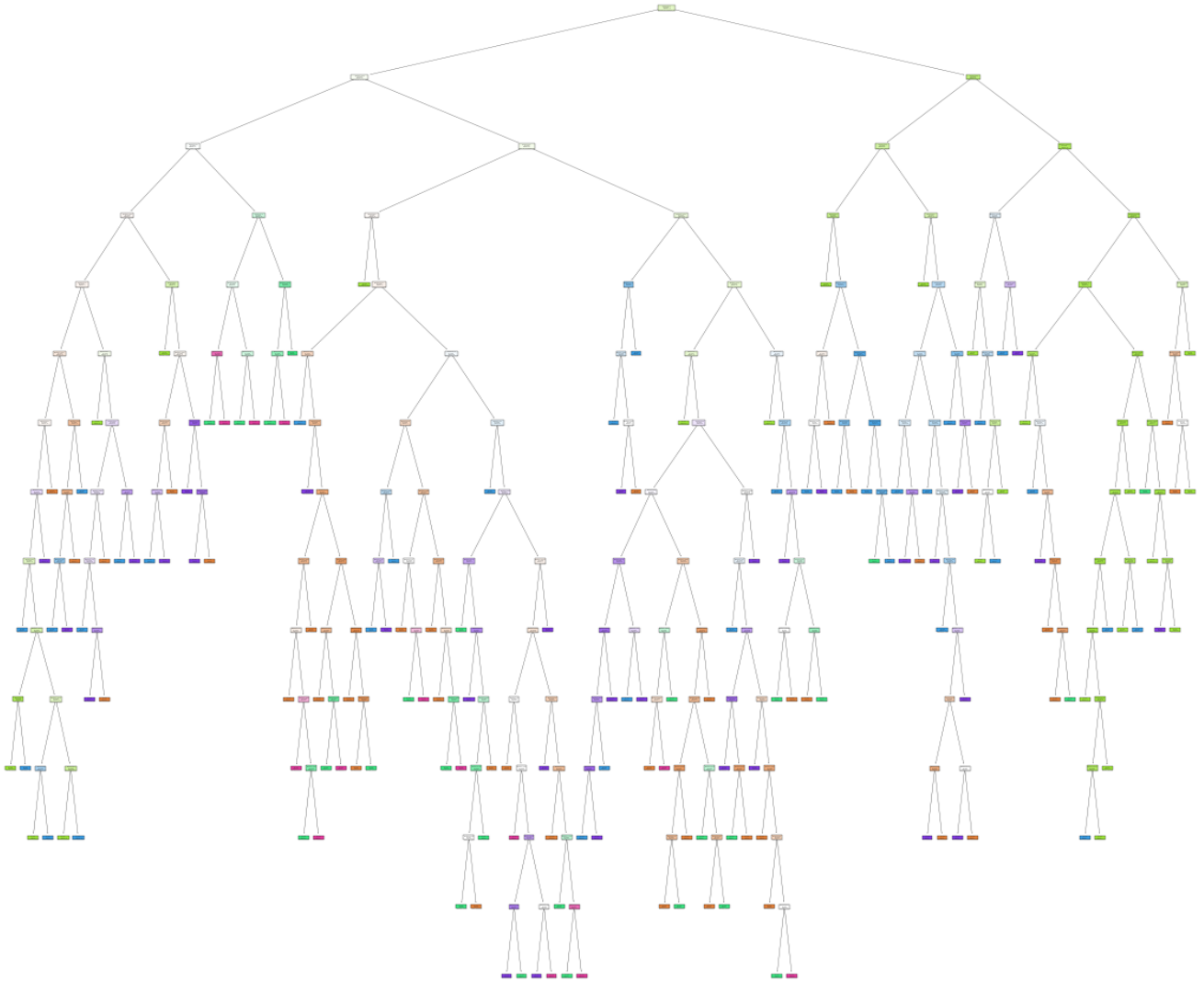
Random Forest Classification Steps



Random Forest Results : First Tree (Single Decision Tree)



Random Forest Results : Single decision tree



Model Evaluation

- To evaluate model built using KNN , CART and Random forest algorithms, we will use Classification reports and 10-k fold Cross validation.
- The Classification report gives :
    1. Precision — The ability of a classifier not to label an instance positive that is actually negative.
    2. Recall — The ability of a classifier to find all positive instances.
    3. F1-score — A weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0.
    4. Support — The number of actual occurrences of the class in the specified dataset.
- The confusion matrix gives a clear overview of the actual labels and the prediction of the model. It represents the accuracy visualization of the predicted model.
- Cross Validation is a very useful tool of a data scientist for assessing the effectiveness of the model, especially for tackling overfitting and underfitting. In addition, it is useful to determine the hyper parameters of the model, in the sense that which parameters will result in lowest test error

KNN Model Evaluation

```
Test Result:
================================================
Accuracy Score: 98.39%

_______________________________________________
CLASSIFICATION REPORT:
                    1            2           3           4           5  \
precision    0.952087     1.000000    0.959732    0.984241    0.982422
recall       0.980892     0.994777    0.907937    1.000000    0.996040
f1-score     0.966275     0.997381    0.933116    0.992058    0.989184
support    628.000000  1723.000000  315.000000  687.000000  505.000000

                    6  accuracy    macro avg  weighted avg
precision    0.987654  0.983949     0.977689      0.984043
recall       0.944882  0.983949     0.970754      0.983949
f1-score     0.965795  0.983949     0.973968      0.983860
support    254.000000  0.983949  4112.000000   4112.000000
_______________________________________________
Confusion Matrix:
 [[ 616    0    4    0    8    0]
 [   0 1714    0    9    0    0]
 [  25    0  286    0    1    3]
 [   0    0    0  687    0    0]
 [   0    0    0    2  503    0]
 [   6    0    8    0    0  240]]
```

- It can be observed that class 2 has a highest and class 1 has the lowest precision.
- That means KNN model best fitted for 2nd cluster i.e Active loyal buyers compared to the 1st cluster i.e. Retail buyers.
- Overall Test accuracy of this model is 98.39%

CART Testing Set Results:

```
|--- feature_0 <= 0.08
|   |--- class: 2
|--- feature_0 >  0.08
|   |--- feature_0 <= 0.24
|   |   |--- feature_0 <= 0.09
|   |   |   |--- feature_2 <= 0.01
|   |   |   |   |--- feature_1 <= 0.00
|   |   |   |   |   |--- class: 4
|   |   |   |   |--- feature_1 >  0.00
|   |   |   |   |   |--- class: 4
|   |   |   |--- feature_2 >  0.01
|   |   |   |   |--- class: 2
|   |   |--- feature_0 >  0.09
|   |   |   |--- class: 4
|   |--- feature_0 >  0.24
|   |   |--- feature_0 <= 0.44
|   |   |   |--- class: 5
|   |   |--- feature_0 >  0.44
|   |   |   |--- feature_0 <= 0.63
|   |   |   |   |--- class: 1
|   |   |   |--- feature_0 >  0.63
|   |   |   |   |--- feature_0 <= 0.82
|   |   |   |   |   |--- class: 3
|   |   |   |   |--- feature_0 >  0.82
|   |   |   |   |   |--- class: 6
```

Feature 0- Recency
Feature 1- Frequency
Feature 2- Monetary

CART Model Evaluation

```
Test Result:
================================================
Accuracy Score: 99.32%

_______________________________________________
CLASSIFICATION REPORT:
                    1            2           3           4            5  \
precision    1.000000     1.000000    0.960366    0.978632     1.000000
recall       0.992038     0.991875    1.000000    1.000000     0.998020
f1-score     0.996003     0.995921    0.979782    0.989201     0.999009
support    628.000000  1723.000000  315.000000  687.000000  505.000000

                    6  accuracy    macro avg  weighted avg
precision    1.000000  0.993191     0.989833      0.993394
recall       0.968504  0.993191     0.991739      0.993191
f1-score     0.984000  0.993191     0.990653      0.993217
support    254.000000  0.993191  4112.000000   4112.000000
_______________________________________________
Confusion Matrix:
 [[ 623    0    5    0    0    0]
 [   0 1709    0   14    0    0]
 [   0    0  315    0    0    0]
 [   0    0    0  687    0    0]
 [   0    0    0    1  504    0]
 [   0    0    8    0    0  246]]
```

- In case of CART model, It can be observed that class 1,2 ,5 and 6 has a highest and class 3 has the lowest precision.
- That means CART model gives best prediction results for cluster 1, 2 5 and 6 type of clients.
- Overall Test accuracy of this model is 99.32%

Random Forest Model Evaluation

```
Test Result:
================================================
Accuracy Score: 98.64%
________________________________________________
CLASSIFICATION REPORT:
                    1       2           3           4           5           6  \
precision    0.950376     1.0    0.948630    0.997093    0.990099    0.995885
recall       0.978328     1.0    0.905229    1.000000    0.994036    0.960317
f1-score     0.964150     1.0    0.926421    0.998544    0.992063    0.977778
support    646.000000  1719.0  306.000000  686.000000  503.000000  252.000000

              accuracy    macro avg  weighted avg
precision     0.986381     0.980347      0.986433
recall        0.986381     0.972985      0.986381
f1-score      0.986381     0.976493      0.986317
support       0.986381  4112.000000   4112.000000
________________________________________________
Confusion Matrix:
 [[ 632    0   10    0    4    0]
 [   0 1719    0    0    0    0]
 [  27    0  277    0    1    1]
 [   0    0    0  686    0    0]
 [   1    0    0    2  500    0]
 [   5    0    5    0    0  242]]
```

- It can be observed that class 2 has a highest precision.
- That means KNN model best fitted for 2nd cluster i.e Active loyal buyers compared to the other types of clients.
- Overall Test accuracy of this model is 98.64%

Model Evaluation : K Fold Cross Validation

- In the Cross-validation the dataset is randomly split up into ‘k’ groups. One of the groups is used as the test set and the rest are used as the training set.
- The model is trained on the training set and scored on the test set. Then the process is repeated until each unique group as been used as the test set.
- Cross-validation gives the model an opportunity to test on multiple splits so we can get a better idea on how the model will perform on unseen data.
- In order to train and test our model using cross-validation, we will use the ‘cross_val_score’ function with a cross-validation value of 10.
- ‘cross_val_score’ takes in our model and our data as parameters. Then it splits our data into 10 groups and fits and scores our data 10 seperate times, recording the accuracy score in an array each time. We will save the accuracy scores in the ‘cv_scores’ variable.
- Using cross-validation, for our KNN model ,mean score is about 77.41%.
- Using cross-validation, for our CART model ,mean score is about 99.65%.
- Using cross-validation, for our Random Forest model ,mean score is about 98.35%.

Model Selection : Accuracy

	Test Accuracy Score	Root mean Square error	Hyperparameter Tunning Optimization	K- fold Cross Validation
KNN	98.39%	0.2009	0.87	77.41%
CART	99.32%	0.1420	99.4%	99.65%
Random Forest	98.64%	0.3308	97.5%	98.35 %

- The CART model showed the smallest error and highest accuracy in the training sample.
- The results of the customers' distribution by classes are presented in next slide.
- On the test sample, this algorithm showed an accuracy of 99%, so CART was chosen to implement the classification process for the entire data sample .

Each cluster characterizes a specific group of customers that are similar in purchasing activity. At the same time, clients have a significant difference between clusters.

Cluster No	Cluster Name	Number Of Customers
Cluster 1	Retail Buyer	908
Cluster 2	Active Loyal Customers	2452
Cluster 3	Threshold Buyer	444
Cluster 4	New Buyer	987
Cluster 5	Occasional Buyer	725
Cluster 6	Most cheap buyer	358

Model Deployment

This is the final phase of the CRISP-DM process.

Since the model achieved high predictive performances ,it can be used to enhance campaigns

Here we will use the Model created and perform deployment.

After evaluating the results obtained from the Model, in the previous phase, deployment would be executed on a small set of data to ensure success of Model.

Simple report of results will be generated and the results will be checked to verify success.

The deliverables will be compared against the expected goal as discussed during the Business Understanding phase.

Finally the Model execution instructions will be provided to the company, who can now utilize this Model to categories clients, and build the effective marketing strategies.

This will in turn increase the chances of increasing the business and retaining client.

Conclusion

The online retailers wants to build their data-driven strategies investing heavily in developing their own intelligent decision-making systems for classification online store clients by their purchasing activity.

The analysis of different approaches allowed us to propose the solution of this problem in two stages.

At first, we segmented our e-commerce customers by RFM metrics using the k-means method. The algorithms automated selection of the number of clusters and there were 6 groups of clients highlighted.

Conclusion

In the second stage, with help of machine learning algorithms the customers' classification system was built. The presence of the second stage is conditioned by the need to take into account the constant updating of the client base and accumulation of new information.

Tenfold cross validation was performed to avoid retraining models. The analysis of calculations by 3 classification methods allowed us to give the advantage of the "CART" method.

Customer segmentation in 6 groups can lead to building better customer centric strategies.

References

Data Set Source : https://archive.ics.uci.edu/ml/datasets/Online+Retail+II#

Papers:

1. https://link.springer.com/article/10.1057/dbm.2012.17
2.https://searchdatamanagement.techtarget.com/definition/RFM-analysis
3. http://ceur-ws.org/Vol-2649/paper2.pdf

Other References:

1. Class Notes
2. https://scikit-learn.org/
3. https://www.analyticsvidhya.com/

Thank You!