EBA 5003 Practice Module in Customer Analytics

Title of Final Report

# Creating the next BIG HIT


Date of Report

23 Apr 2023

Team Name

Group 4

Team Members

LIN FANGZHOU(A0261850H)

YAN ZIHAN(A0261738X)

CHEN YUMENG (A0261899H)

SHI KECHEN (A0261672A)

SU CHEN (A0261760H)

# Table of Contents

# List of Figures

# List of Tables

# 1 Introduction

Tesco is a traditional supermarket located in UK which has suffered significant revenue loss as the offline retail industry has taken a downturn and customer churn. The data analytics team from Tesco decided to data mining existing customer data to enhance the market awareness such as customer segmentation and develop a new online campaign to maintain its customers to stabilize its market share and expand.

# 2 Industry Overview

The supermarket industry is a highly competitive and dynamic sector. It is characterized by the presence of several large, multinational corporations, as well as numerous regional and local players. The industry has been impacted by the rise of e-commerce and online grocery shopping, as well as changes in consumer behavior and preferences. The industry has responded to these challenges by investing in technology and supply chain efficiency, as well as expanding their product offerings and in-store experiences. As the economy recovered in 2023, physical stores receive an onslaught from competitors, there is an urgent need to broaden online sales, to survive the competition, data analytics teams need to dig deeper into customer data and come up with insightful solutions. With the outcome, senior executives could develop a new market strategy, define the target customer, and provide potential business insights to the management of the company as required.

# 3 Business Problem & Objectives

## 3.1. Business Problem

- Advances in digital technologies and improved logistics and distribution systems have transformed virtually every aspect of the retail industry - from consumer demand to shopping habits.
- At the same time, as rapidly increases of material standard of living, consumers are placing more and more emphasis on the omnichannel shopping experience and corporate brand identity, and companies can only win in the long run by being the best at all aspects of the consumer journey.
- As a result, it has become increasingly difficult for retailers to keep their business growing and retain loyal customers.

## 3.2. Business Value-Add

We are a global data scientist team specializing in customer analytics for large supermarket stores. The business analytics project aims to deliver data-driven insights into customers' behavior and transaction patterns, which will help supermarket owners with:

- A guideline to customer understanding, allowing store owners have a clear view for customers composition, expectation, and contribution to the total revenue for business.

- Building successful customer relationship from customer loyalty analysis, strengthen loyal customer group while reviving the hibernating group.
- Increasement in potential profits and growth opportunities by offering effective campaign design, help in building scientific measurement to monitor campaign's objectives.

## 3.3. Project Objectives

- Conduct customer segmentation by using clustering analysis on users' behavioural and demographic data. Enhance understanding for customers and help to develop future strategies for specialized groups.
- Maintain a successful customer relationship through customer loyalty analysis. Implement RFM model based on customer recency, frequency, and monetary data to help companies track customers and build a relationship that can increase sales and productivity.
- Deploy propensity model to identify target group for the next campaign based on past customer responses to previous initiatives. Analyse the result from lift chart to obtain a desired response rate.
- Design campaigns on next business cycle based on interpretation of models and customers' segmentations to better promote merchandising and maintain relationships with customers. Establish measurement to monitor campaign's future ROI for campaign tunning and optimization.

# 4 Project Design

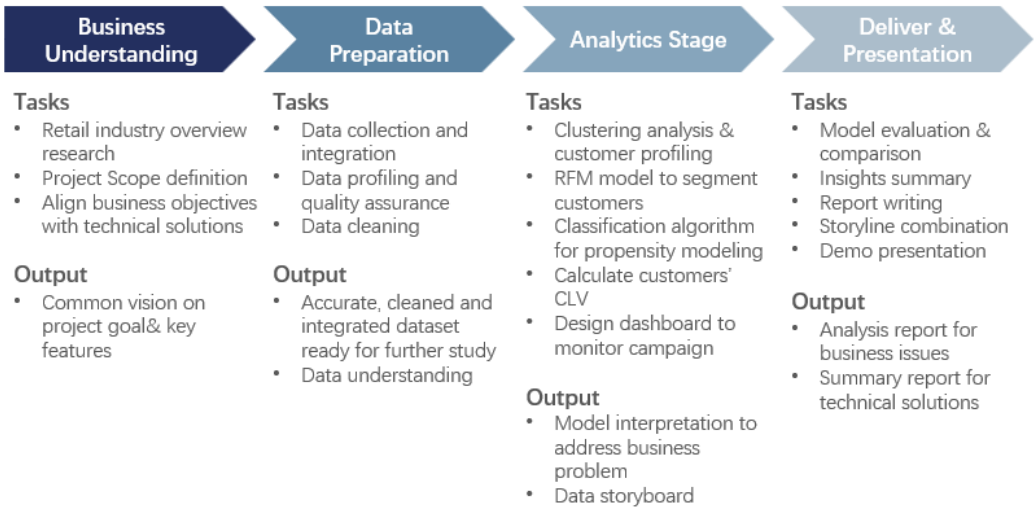


Figure 1

# 5 Scope of work

- Compare two scenarios based on consumers' demographic and behavioural variables. Then select the optimal solution based on the two performances.
- Derive principal components after dimension reduction and then clustering.
- Direct clustering based on existing variables.
- Build RFM model, using customer recency, frequency, monetary data to find loyal groups.
- Build various models such as logistic regression, neural network and SVM, compare the predictive performance of the models on the results of customer response, and select the most suitable model as the basis of Propensity score
- Based on consumer purchase records and loyalty information, deploy BTYD models to predict consumer lifecycle values.
- Conduct campaign design and visualize the process based on the models and segments.

# 6 Effort Estimates and Timeline

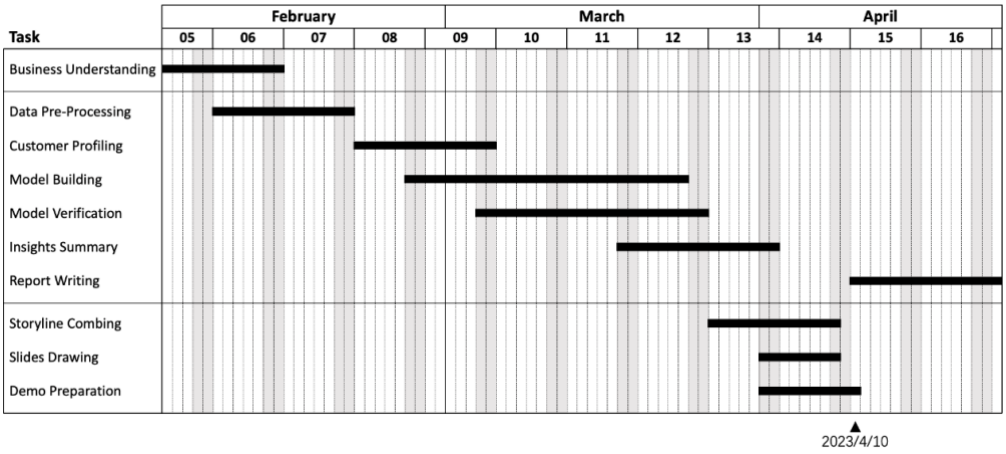


Figure 2

# 7 Dataset Used & Data Description

The dataset contains customers' behavioral and transaction data of Tesco from 2020 to 2022. It has the dimension of 2240*29, and the 29 attributes can be categorized into different groups: customers' demographic data, customers' purchase records, customers' response to previous campaigns, customers' loyalty, and customers' behavioral information. As each row represents an unique customer, the credential of the dataset is promising. In addition, the overall quality of the dataset is good, except for a few duplicated rows.

Some key features of the dataset are shown in ***Table 1*** below.

Table 1

| Column Group Name | Description | Columns Included | Type | Structured or Unstructured | Sample Value |
|---|---|---|---|---|---|
| Demo_info | Customers' ID and demographic information | 'ID','Year_Birth', 'Education', 'Marital_Status', 'Income', 'Kidhome', 'Teenhome' | Int & Char | Structured | 16, 1999-Aug-01, Primary School, Single, 58138.0, 1, 0 |
| Buying_info | Customers' buying amount for different categories of goods | 'MntWines','MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds' | Int | Structured | 635, 289, 88, 290, 283, 309 |
| Promo_info | Customers' response to previous campaigns | 'AcceptedCmp1', 'AcceptedCmp2', 'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5', 'Response' | Int | Structured | 0,0,0,0,1,1 |
| Loyal_info | Customers' joining date of Tesco, and number of days since the last purchase | 'Dt_Customer', 'Recency', 'Complain' | Int & Char | Unstructured | 2000-01-27, 80, 0 |
| Purchase_behave_info | Customers' behavioral data on different purchase channels | 'NumDealsPurchases', 'NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases', 'NumWebVisitsMonth' | Int | Structured | 10, 8, 9, 3, 8 |

# 8 Data Pre-processing

## 8.1. Data Cleaning

In the data cleaning phase, the "is.na" function was used to check for null values in the dataset. It was found that only the [income] column had 24 missing values, which were imputed with the mean value. Outliers were also checked, and any observations with more than three standard deviations were removed, leaving 29 columns and 2240 rows in the cleansed dataset.

## 8.2. Data Transformation

The following steps are used to transform data:

- Rename categories in the [Marital_Status] column: 'Married' and 'Together' to 'In relationship', 'Divorced', 'Widow', 'Absurd', 'Alone', and 'YOLO' to 'Single'
- Rename '2n Cycle' category in the [Education] column to 'Master'
- Create a new variable [Kids] by adding 'Kidhome' and 'Teenhome'
- Create a new variable [Expenses] by summing up [MntWines], [MntFruits], [MntMeatProducts], [MntFishProducts], [MntSweetProducts], and [MntGoldProds]
- Create a new variable [Time_Enrolled_Days] to represent the number of days between the customer's registration date and a fixed date.

## 8.3. Data Privacy and Security

Since the used dataset is an open-source dataset on Kaggle, without private data, there is no need to mask or decrypt any columns or rows of the exploratory dataset. In addition, there is no third-party access to our project, thus we do not need to set the confidential level for a different role of the project.

# 9 K-Means Clustering & Customer Profiling

## 9.1. K-means Clustering

For the K-means clustering, only Demo_info and Buying_info data were used, and categorical variables such as [Education] and [Marital Status] were excluded. By using the elbow method and the optimal CCC (Cubic Clustering Criterion), a cluster number of 4 was chosen, as shown in ***Figure 3*** and ***Table 2***.

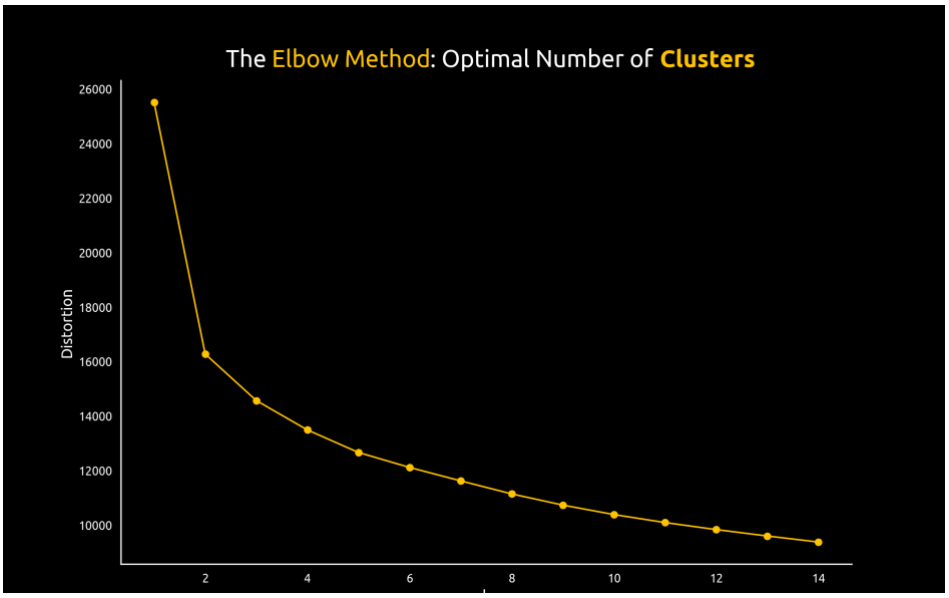


Figure 3

Table 2

| Clustering Comparison | | |
|---|---|---|
| N Cluster | CCC | Best |
| 3 | 1.83 | |
| 4 | 2.01 | Optimal CCC |
| 5 | -1.73 | |
| 6 | 2.75 | |

The final result of K-means are presented below(***Figure 4***):

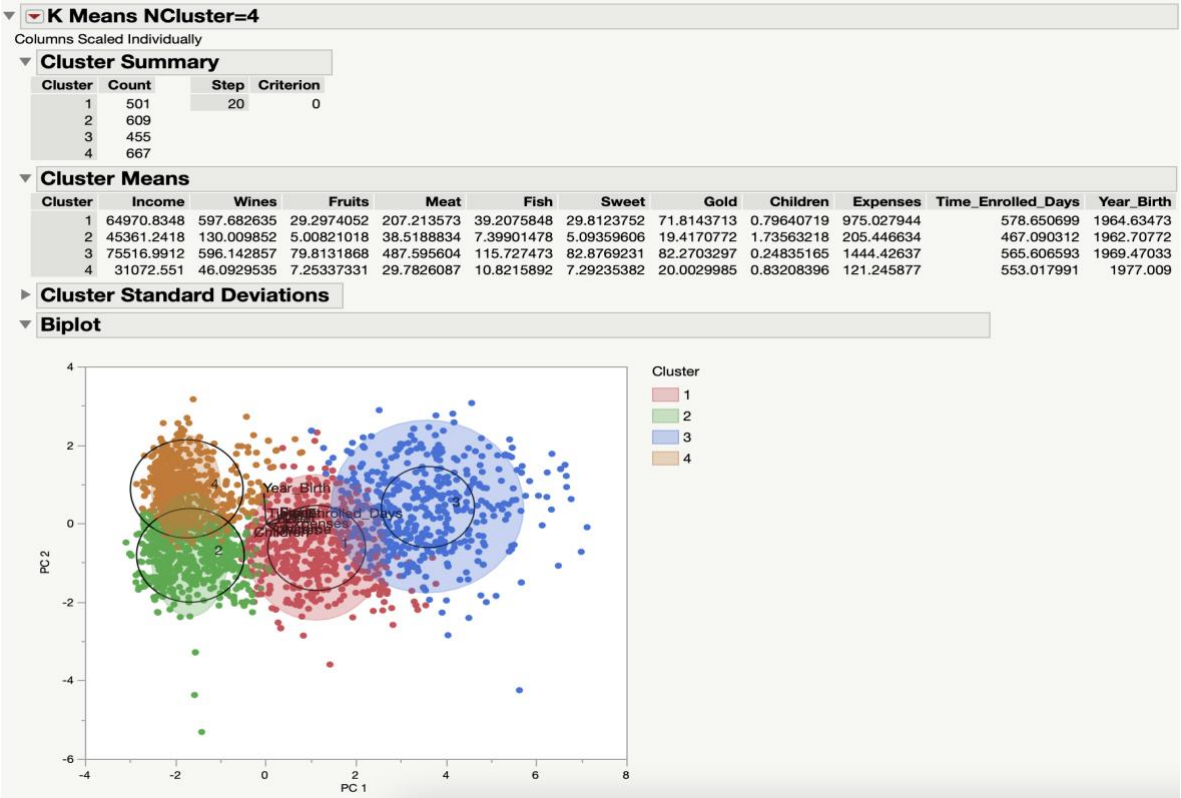


Figure 4

## 9.2. Customer Profiling

The 2-dimensional diagram below plotted customers by their expense and income. Based on these two dimensions, customers can be identified as economical customers group, cheap customers group, good customers group also elite group.

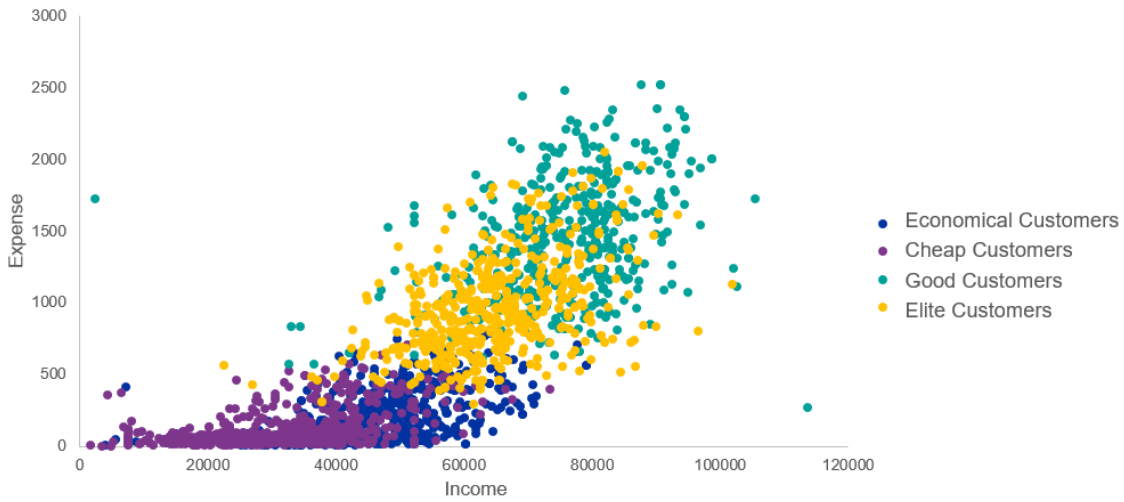


Figure 5

From the Demographic information, Cheap customers can be concluded as the youngest; Elite customers mainly include elites with high education and high income.

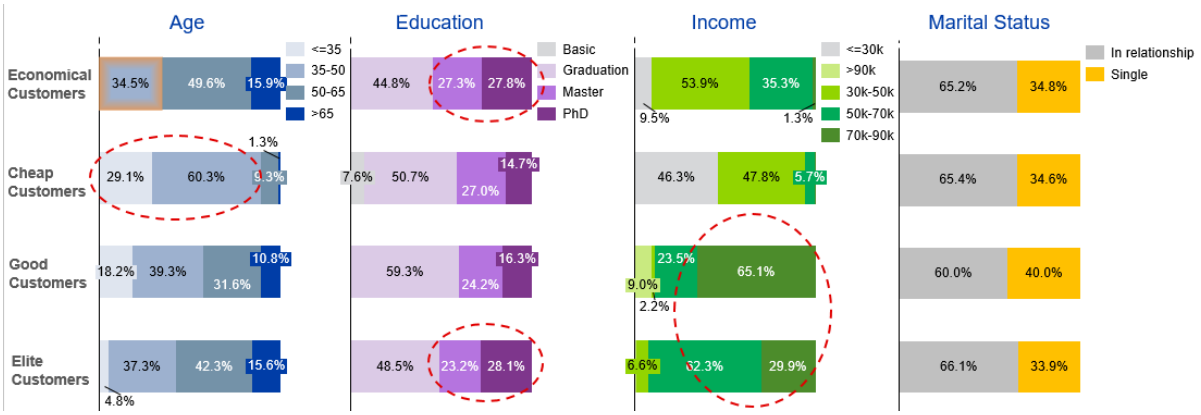


Figure 6

The below diagram shows the proportion of each category that is allocated to a particular cluster. From the Radar diagram, few points that can be concluded:

- Good Customers are the main consumers of daily necessities, such as meat, fruit, fish, etc.
- Elite Customers are the major buyers of gold and wine
- Generally, Economical and cheap customer groups show low spending habits

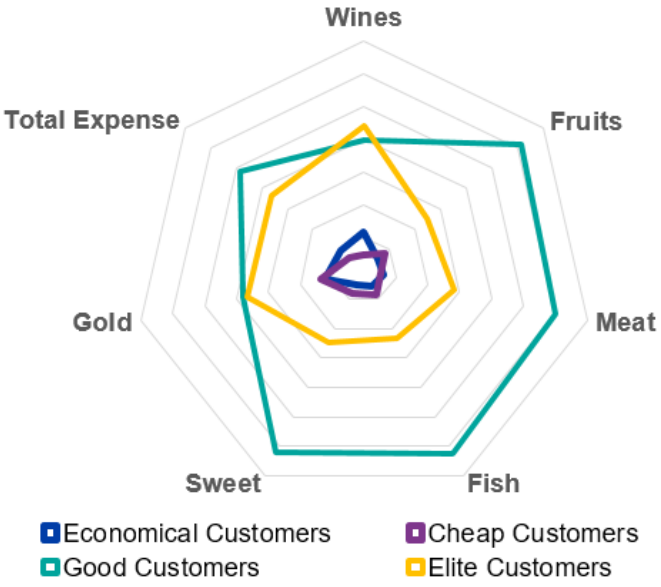


Figure 7

In general, there are four main types of customer group. The Good customers group consists of individuals who have high spending habits as well as higher education. They are the main consumers of daily necessities. The elite group consists of the highest incomes and highest spending habits group, also they are the major buyers of gold and wine.


Figure 8

# 10 RFM Model

## 10.1. Data preparation

RFM segmentation is a scoring technique used to better quantify customer behavior. During marketing campaigns, not all customers should be contacted with the same effort. Direct marketing segmentation enables to group of customers into different segments and analyze their profitability accordingly.

RFM metrics are closely related to the Customer Lifetime Value as frequency and monetary value affect directly CLV and recency affect retention. In short words, RFM metrics are described below:

- Recency : Time since last order
- Frequency : Total number of transactions
- Monetary : Total transactions value

These metrics are very important to understand customer behavior: the more recent the purchase, the more responsive the customer is to promotions; the more frequently customers buy, the more engaged they are.

In the Tesco dataset, recency, frequency, and monetary value can be extracted from loyal_info, purchase_behave_info, and buying_info. After setting the hurdle rule as more than one purchase and spending over ten, each of the values is divided into five equal parts, yielding a 5*5*5-cluster for all customers. ***Table 3*** shows the RFM score after data processing.

Table 3

| ID | Recency | Frequency | Monetary | recency_score | frequency_score | monetary_score | RFM_SCORE |
|---|---|---|---|---|---|---|---|
| **5524** | 58 | 25 | 1617 | 3 | 5 | 5 | **35** |
| **2174** | 38 | 6 | 27 | 4 | 1 | 1 | **41** |
| **4141** | 26 | 21 | 776 | 4 | 4 | 4 | **44** |
| **6182** | 26 | 8 | 53 | 4 | 2 | 1 | **42** |
| **5324** | 94 | 19 | 422 | 1 | 4 | 3 | **14** |

## 10.2. Customer Segmentation

In order to keep a manageable number of segments, the segments are created using only the recency and frequency scores. The monetary score is often viewed as an aggregation metric for summarizing transactions and will be used in subsequent analyses.

We will use the below ten segments to divide customers:

- Champions---------->Bought recently, buy often and spend the most
- Loyal Customers---------->Buy on a regular basis. Responsive to promotions.
- Potential Loyalist---------->Recent customers with average frequency.
- Recent Customers---------->Bought most recently, but not often.
- Promising---------->Recent shoppers, but haven't spent much.
- Customers Needing Attention------>Above average recency, frequency and monetary values. May not have bought very recently though.
- About To Sleep---------->Below average recency and frequency. Will lose them if not reactivated.
- At Risk---------->Purchased often but a long time ago. Need to bring them back!
- Can't Lose Them---------->Used to purchase frequently but haven't returned for a long time.
- Hibernating---------->Last purchase was long back and low number of orders. May be lost.

## 10. 3. Result and analysis

After segmentation, a tree map is constructed to understand how customers are distributed across the two axes. In ***Figure 9***, the horizontal axis represents the frequency value, whereas the vertical axis stands for the recency value. The ideal customer group, the champions, lie in the top-right corner of the map, next to the potential loyalist and new customers. As for hibernating customers, they stand next to someone who is at risk and about to sleep.

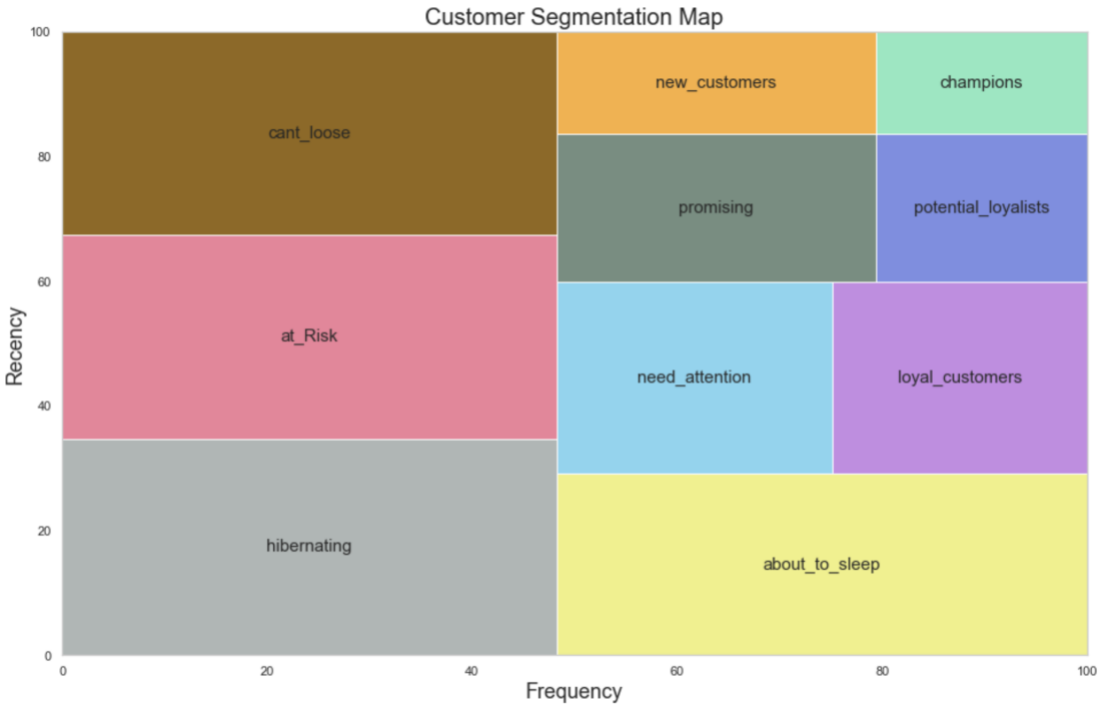


Figure 9

***Table*** 4 gives out the statistical summary for all the segments. Clearly distinctions can be seen between groups. Champions group, for example, shows an average day of 9 since their last purchase, in contrast, hibernating group's last purchase record can date back to three months ago. Nonetheless, champions group come to Tesco for shopping over 22 times and spend money above 1000.

Table 4

| | Recency | | Frequency | | Monetary | |
|---|---|---|---|---|---|---|
| | mean | count | mean | count | mean | count |
| **segment** | | | | | | |
| **about_to_sleep** | 48.912088 | 182 | 7.252747 | 182 | 80.510989 | 182 |
| **at_Risk** | 78.086455 | 347 | 17.976945 | 347 | 930.123919 | 347 |
| **cant_loose** | 79.611111 | 162 | 26.000000 | 162 | 1199.253086 | 162 |
| **champions** | 9.166667 | 168 | 22.845238 | 168 | 1067.440476 | 168 |
| **hibernating** | 79.883152 | 368 | 7.527174 | 368 | 102.524457 | 368 |
| **loyal_customers** | 39.478261 | 345 | 23.156522 | 345 | 1120.031884 | 345 |
| **need_attention** | 50.243243 | 74 | 15.364865 | 74 | 719.094595 | 74 |
| **new_customers** | 9.460177 | 113 | 5.460177 | 113 | 48.840708 | 113 |
| **potential_loyalists** | 18.882530 | 332 | 12.638554 | 332 | 442.027108 | 332 |
| **promising** | 29.747664 | 107 | 5.448598 | 107 | 39.102804 | 107 |

***Figure 10*** shows the proportions for ten groups. Good news is that champions contribute to the 7 percent of total customers, not a small portion. However, hibernating group plus risk group take up to above 30 percent of the total population, suggesting there will be huge potential lost for customers if Tesco do not take action.

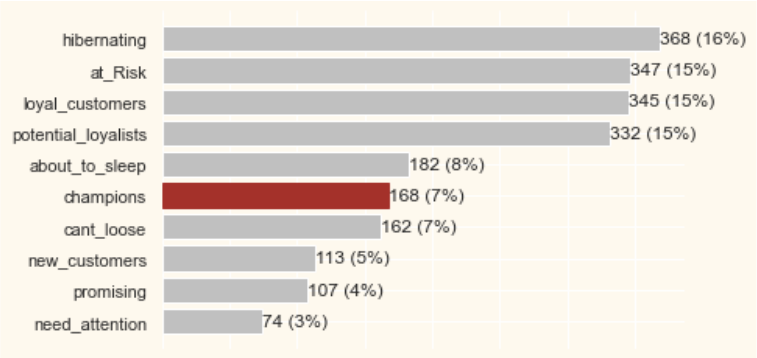


Figure 10

# 11 Propensity Model

## 11.1. Model Construction and Selection

To construct the propensity model, the first step is to generate a model that provides propensity scores. As this is a classification problem, three models were developed, namely, the Logistic Regression Model, Neural Network Model, and Support Vector Machine (SVM) Model. The primary comparison criteria for these models were the ROC Curve, AIC, and BIC values. The comparison results are presented below (***Figure 11*** & ***Table 5***):

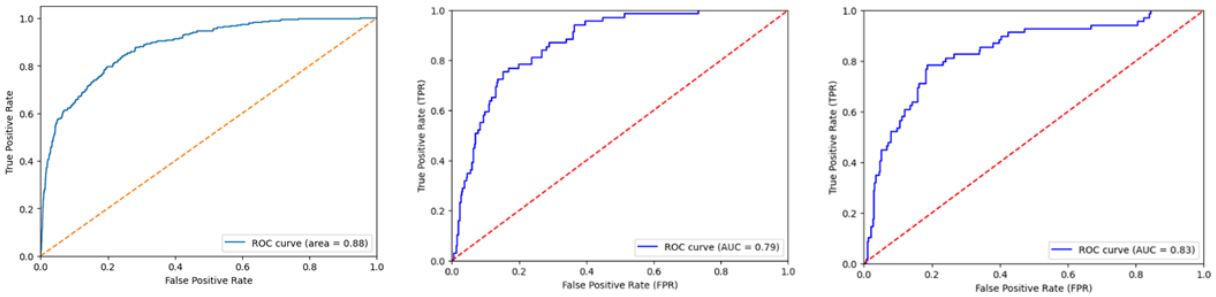


Figure 11
Comparison of ROC curves for 3 models
(Left: logistic regression, Middle: neural network, Right: SVM)

Table 5

| Model | Logistic Regression | Nueral Network | SVM |
|---|---|---|---|
| AIC | 849 | 8517 | 669 |
| BIC | 950 | 32721 | 965 |
| AUC | 0.88 | 0.79 | 0.83 |
| **ACCURACY** | **86.3%** | **81.1%** | **86.5%** |

As observed from the results, the Logistic Regression model yields a higher AUC value and relatively low AIC and BIC values. Given the commercial context, it is more appropriate to select a model that is both concise and effective. Therefore, the Logistic Regression model was ultimately chosen as the foundation for generating propensity scores.

The model was constructed using customer response to the forthcoming campaign as the target variable (Y). Given the relatively small dataset and the fact that 14% of the response values were 1, logistic regression was applied to both the original dataset and an over-sampled dataset. Several influential variables were selected for inclusion in the model, based on their statistical significance and their impact on the model's performance. A comprehensive overview of the model, including the chosen variables and their corresponding coefficients, is provided in ***Figure 12*** & ***Figure 13.***

**Parameter Estimates**

| Term | Estimate | Std Error | ChiSquare | Prob>ChiSq |
|---|---|---|---|---|
| Intercept | -0.0413795 | 0.5134614 | 0.01 | 0.9358 |
| Education[2n Cycle] | -0.2747986 | 0.2825717 | 0.95 | 0.3308 |
| Education[Basic] | -1.1708524 | 0.6060354 | 3.73 | 0.0534 |
| Education[Graduation] | 0.07266119 | 0.1902723 | 0.15 | 0.7026 |
| Education[Master] | 0.42884705 | 0.2202584 | 3.79 | 0.0515 |
| Marital_Status[Single] | 0.58793359 | 0.0765318 | 59.02 | <.0001* |
| Teenhome[0] | 0.53625054 | 0.2165427 | 6.13 | 0.0133* |
| Teenhome[1] | -0.3418415 | 0.2165777 | 2.49 | 0.1145 |
| Recency | -0.0284437 | 0.0028565 | 99.15 | <.0001* |
| MntMeatProducts | 0.00277437 | 0.000382 | 52.74 | <.0001* |
| MntGoldProds | 0.00427867 | 0.001506 | 8.07 | 0.0045* |
| NumWebPurchases | 0.11017303 | 0.0300024 | 13.48 | 0.0002* |
| NumStorePurchases | -0.1242412 | 0.0302392 | 16.88 | <.0001* |
| NumWebVisitsMonth | 0.24872635 | 0.0379483 | 42.96 | <.0001* |
| AcceptedCmp3[0] | -0.8858651 | 0.1079043 | 67.40 | <.0001* |
| AcceptedCmp4[0] | -0.5305717 | 0.133168 | 15.87 | <.0001* |
| AcceptedCmp5[0] | -0.755583 | 0.1298613 | 33.85 | <.0001* |
| AcceptedCmp1[0] | -0.6118087 | 0.1301254 | 22.11 | <.0001* |
| AcceptedCmp2[0] | -0.6889942 | 0.2713383 | 6.45 | 0.0111* |

For log odds of 1/0

Figure 12
Model Parameter Estimate (Original)

**Parameter Estimates**

| Term | Estimate | Std Error | ChiSquare | Prob>ChiSq |
|---|---|---|---|---|
| Intercept | 0.5572826 | 0.4425055 | 1.59 | 0.2079 |
| Education[2n Cycle] | -0.141867 | 0.1875413 | 0.57 | 0.4494 |
| Education[Basic] | -1.395562 | 0.3599011 | 15.04 | 0.0001* |
| Education[Graduation] | 0.18100882 | 0.1202759 | 2.26 | 0.1323 |
| Education[Master] | 0.3234279 | 0.1502701 | 4.63 | 0.0314* |
| Marital_Status[Single] | 0.4836863 | 0.0562109 | 74.04 | <.0001* |
| Teenhome[0] | 1.04507725 | 0.1733486 | 36.35 | <.0001* |
| Teenhome[1] | -0.4397673 | 0.1647264 | 7.13 | 0.0076* |
| Recency | -0.0322755 | 0.0020919 | 238.04 | <.0001* |
| MntMeatProducts | 0.00284086 | 0.0003662 | 60.18 | <.0001* |
| NumDealsPurchases | 0.22262068 | 0.0341416 | 42.52 | <.0001* |
| NumWebPurchases | 0.07899229 | 0.0265903 | 8.83 | 0.0030* |
| NumCatalogPurchases | 0.16914388 | 0.0303102 | 31.14 | <.0001* |
| NumStorePurchases | -0.1952055 | 0.0247572 | 62.17 | <.0001* |
| NumWebVisitsMonth | 0.35207496 | 0.0372215 | 89.47 | <.0001* |
| AcceptedCmp3[0] | -0.9229113 | 0.0938012 | 96.81 | <.0001* |
| AcceptedCmp4[0] | -0.3612053 | 0.1135766 | 10.11 | 0.0015* |
| AcceptedCmp5[0] | -0.985746 | 0.1177557 | 70.08 | <.0001* |
| AcceptedCmp1[0] | -0.5499677 | 0.1127894 | 23.78 | <.0001* |
| AcceptedCmp2[0] | -0.711573 | 0.2558965 | 7.73 | 0.0054* |
| MntSweetProducts | 0.00498579 | 0.0018 | 7.67 | 0.0056* |

Figure 13
Model Parameter Estimate (Over-Sampling)

Model performance was assessed through the comparison of ROC Curve (***Figure 14*** & ***Figure 15***), AIC, and BIC (***Table 6***) values.

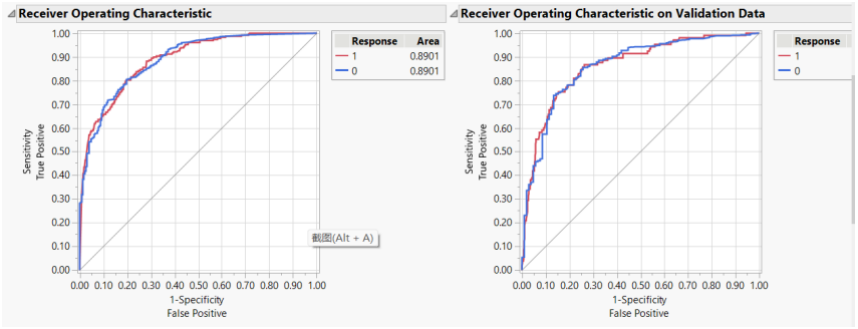


Figure 14
ROC Curve (Original)

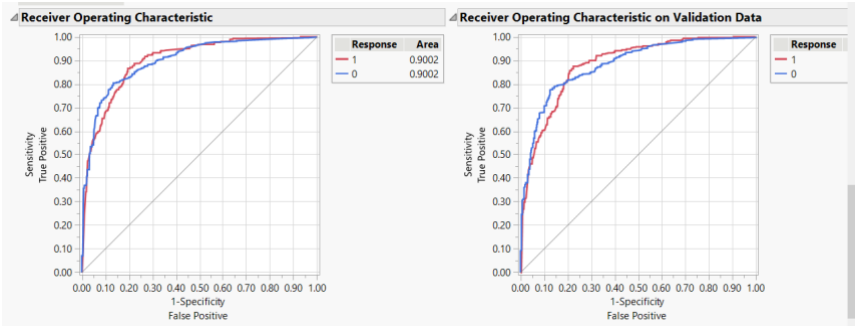


Figure 15
ROC Curve (Over-Sampling)

Table 6
AIC and BIC Values

| | Origin | Over-Sampling |
|---|---|---|
| AIC | 848.726 | 2194.49 |
| BIC | 949.816 | 2317.55 |
| **AUC** | **0.89** | **0.90** |

The classification performance of models trained on both sample types exhibited a high degree of similarity. However, upon considering the AIC and BIC values, the classification model derived from the original sample was ultimately selected in order to maintain a balance between the goodness of fit and model complexity. To further evaluate the model's performance, the confusion matrix, accuracy, recall, and precision values for the training set were calculated (***Table 7***).

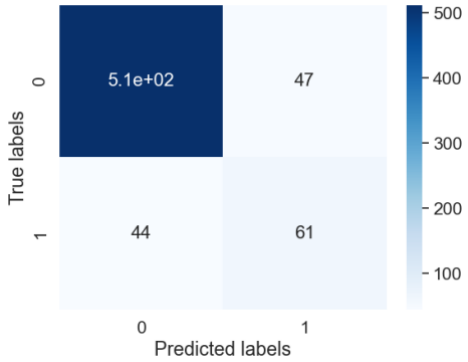


Figure 16

Table 7
Confusion Matrix and Performance Metrics

| | |
|---|---|
| Recall Rate | 58.1% |
| Precision | 56.5% |
| **Accuracy** | **86.3%** |

## 11.2. Customer Segmentation and Propensity Score Calculation

Based on the model's results, a propensity score reflecting the likelihood of each customer responding to the next campaign was obtained. This propensity score was used as the basis for the "likely response" column, representing each customer's likelihood of responding to the next campaign. A cutoff probability of 0.3 was chosen to determine whether a customer would respond to the campaign. To facilitate customized services, segmentation results, and CLV were also incorporated. The sample table is provided below (

***Table*** 8).

Table 8
Propensity Score Sample Table

| ID | Response | Likely Response | Probability | Segment | CLV |
|---|---|---|---|---|---|
| 5524 | 1 | 1 | 0.47411 | loyal_customers | 693.5018 |
| 2174 | 0 | 0 | 0.031584 | promising | 44.77978 |
| 4141 | 0 | 0 | 0.020632 | loyal_customers | 598.9745 |
| 6182 | 0 | 0 | 0.043139 | potential_loyalists | 73.49592 |
| 5324 | 0 | 0 | 0.039552 | at_Risk | 130.1774 |
| 7446 | 0 | 0 | 0.01743 | champions | 577.7868 |
| 965 | 0 | 0 | 0.068375 | loyal_customers | 280.5526 |
| 6177 | 0 | 0 | 0.233774 | potential_loyalists | 110.3127 |
| 4855 | 1 | 1 | 0.314662 | new_customers | 34.71968 |
| 5899 | 0 | 1 | 0.790744 | hibernating | 94.67202 |
| 387 | 0 | 0 | 0.013892 | hibernating | 32.41275 |
| 2125 | 0 | 0 | 0.034024 | at_Risk | 941.3801 |
| 8180 | 0 | 0 | 0.047543 | need_attention | 295.7758 |
| 2569 | 0 | 0 | 0.047652 | promising | 24.2971 |

## 11.3. Propensity Score Bands

Customers were sorted by their propensity scores from high to low and divided into 20 bands. The average customer value for each band was calculated by combining the results of CLV (***Table 9***). The first 18 bands covered 84.7% of customers with a high probability of responding, while the remaining two bands accounted for 65.5% of customers, with only about 15% of them likely to respond.

Table 9
Propensity Score Bands and Average Customer Value

| Score Band | Total no. of customers | % Total Customers | Cum % of Customers | Good Customer | Good Customer % | Cumulative Good Customer % | Lift | Good Customer Coverage % | Cumulative Good Customer Coverage | CLV |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 23 | 1.04% | 1.04% | 19 | 82.61% | 82.61% | 5.49 | 5.71% | 5.7% | 1380.178 |
| 2 | 18 | 0.81% | 1.85% | 15 | 83.33% | 82.93% | 5.54 | 4.50% | 10.2% | 1284.302 |
| 3 | 17 | 0.77% | 2.62% | 14 | 82.35% | 82.76% | 5.48 | 4.20% | 14.4% | 984.6935 |
| 4 | 20 | 0.90% | 3.52% | 18 | 90.00% | 84.62% | 5.99 | 5.41% | 19.8% | 1031.995 |
| 5 | 21 | 0.95% | 4.47% | 16 | 76.19% | 82.83% | 5.07 | 4.80% | 24.6% | 732.6851 |
| 6 | 16 | 0.72% | 5.19% | 10 | 62.50% | 80.00% | 4.16 | 3.00% | 27.6% | 920.3877 |
| 7 | 19 | 0.86% | 6.05% | 15 | 78.95% | 79.85% | 5.25 | 4.50% | 32.1% | 943.1882 |
| 8 | 34 | 1.53% | 7.58% | 23 | 67.65% | 77.38% | 4.50 | 6.91% | 39.0% | 749.3026 |
| 9 | 9 | 0.41% | 7.99% | 5 | 55.56% | 76.27% | 3.70 | 1.50% | 40.5% | 567.0026 |
| 10 | 29 | 1.31% | 9.30% | 15 | 51.72% | 72.82% | 3.44 | 4.50% | 45.0% | 811.8421 |
| 11 | 28 | 1.26% | 10.56% | 14 | 50.00% | 70.09% | 3.33 | 4.20% | 49.2% | 635.2837 |
| 12 | 38 | 1.72% | 12.28% | 21 | 55.26% | 68.01% | 3.68 | 6.31% | 55.6% | 455.1572 |
| 13 | 39 | 1.76% | 14.04% | 8 | 20.51% | 62.06% | 1.36 | 2.40% | 58.0% | 624.5998 |
| 14 | 41 | 1.85% | 15.89% | 11 | 26.83% | 57.95% | 1.78 | 3.30% | 61.3% | 429.1416 |
| 15 | 55 | 2.48% | 18.37% | 11 | 20.00% | 52.83% | 1.33 | 3.30% | 64.6% | 559.2074 |
| 16 | 81 | 3.66% | 22.03% | 18 | 22.22% | 47.75% | 1.48 | 5.41% | 70.0% | 505.9042 |
| 17 | 108 | 4.88% | 26.91% | 20 | 18.52% | 42.45% | 1.23 | 6.01% | 76.0% | 428.8541 |
| 18 | 167 | 7.54% | 34.45% | 29 | 17.37% | 36.96% | 1.16 | 8.71% | 84.7% | 452.3316 |
| 19 | 384 | 17.34% | 51.78% | 29 | 7.55% | 27.11% | 0.50 | 8.71% | 93.4% | 328.9143 |
| 20 | 1068 | 48.22% | 100.00% | 22 | 2.06% | 15.03% | 0.14 | 6.61% | 100.0% | 307.2795 |
| Total | 2215 | 100% | | 333.00 | 15.03% | | 1.00 | 100.00% | | 22 |

The ROC curve for the entire sample had an AUC of approximately 0.88, suggesting good classification performance (***Figure 17***). The lift chart for the 20 bands displayed a clear trend, with the lift for the first 18 bands being greater than 1 (***Figure 18***). This indicates the model's effectiveness in identifying customers with a high propensity to respond.

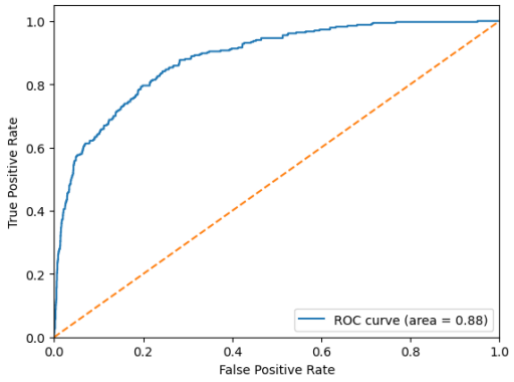


Figure 17
ROC Curve for the Entire Sample

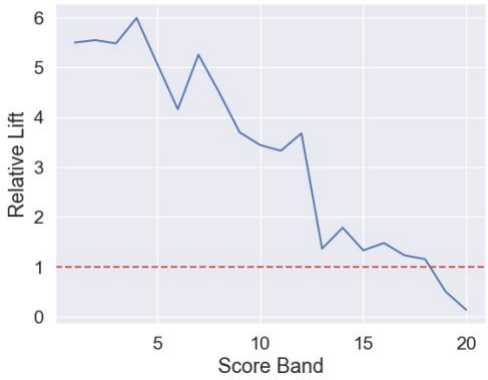


Figure 18
Lift Chart for the 20 Bands

## 11.4. Customer Quadrants

Customers were segmented into four quadrants by integrating both Response and CLV. The X-axis represented the threshold probability of Response at 0.3, and the Y-axis corresponded to the average CLV of all customers. A scatter plot of Response probability and CLV for all customers was generated.

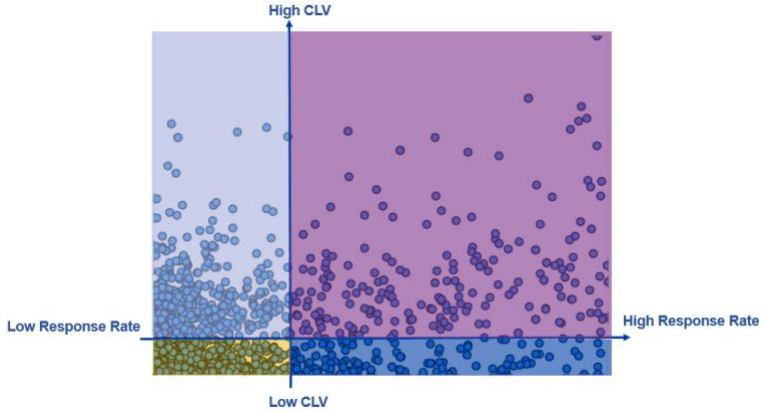


Figure 19
Scatter Plot of Response Probability and CLV

For this scatter plot, the Y-axis intercept (Y=0) can be interpreted as the average CLV of all customers, while the X-axis intercept (X=0) represents the response probability threshold of 0.3. By understanding these benchmarks, marketing teams can better tailor their strategies to effectively target different customer segments and maximize overall return on investment.

Upon examining the scatter plot, some observations can be made. For instance, Quadrant I (high Response and high CLV) exhibits a relatively sparse distribution of customers. Meanwhile, Quadrant III (low Response and low CLV) exhibits the highest concentration. The distinct sales strategies will be devised based on the specific quadrant in which a customer is situated.

Detailed strategies for each quadrant, including Quadrants I and III, will be further discussed and developed in Part 13.

# 12 CLV Model

## 12.1. Model Definition

Customer Lifetime Value can be viewed as the economic value derived from the firm's relationship with its customers. CLV is defined as a measure of the present value of future cash flows attributed to the customer relationship. In other words, CLV measure the net profit a customer will bring to the firm over the future periods. Hence past customer transactions may be used as a predictive driver of the economic value of a firm's customer relationship.

The CLV formula can be written as:

$$CLV = \sum_{n=1}^{N} \frac{Value_n * Retention^n}{(1 + Discount\,Rate)^n}$$

The Buy Till You Die(BTYD) model is built on 4 metrics that are closely related to the ones used for RFM segmentation:

- Frequency: The number of repeated purchases the customer made after his first date of first purchase
- Age (Time): The period the customer has been enrolled in the company, expressed in days, weeks, or even months. Age = Last date in dataset - first customer purchase date
- Recency: The age of the customer when he made his last purchase. Recency = Last customer purchase date - first customer purchase date
- Monetary value: The average amount spent by a customer

While there are several versions of BTYD models, the BG/NBD model is used in the Tesco dataset.
BG/NBD was introduced in 2004 by Peter Fader and stands for Beta Geometric/Negative Binomial Distribution.
The model distinguishes customer behavior in two parts:

- The buying process which models the probability a customer makes a purchase
- The dying process (or dropout) which models the probability a customer quit and never purchase again

## 12.2. CLTV Calculation

To calculate CLV, "DT_Customer" is created which will help to calculate the Recency* and Age variables in the BTYD model.

**Note that Recency used in the BTYD model is different that the one used in RFM segmentation.*
Next step, assuming a monthly discount rate of 1%, the Long Term Value for each customer over the next 12 months are calculated. ***Table 10*** plots the top 10 customers based on LTV.

Table 10

| ID | Frequency | Recency | Age | Monetary_value | segment | LTV |
|---|---|---|---|---|---|---|
| 1826 | 14 | 110 | 110 | 79.33 | potential_loyalists | 2806.89 |
| 8029 | 14 | 131 | 158 | 108.07 | potential_loyalists | 2832.24 |
| 9264 | 21 | 159 | 160 | 74.45 | champions | 2853.27 |
| 7959 | 10 | 109 | 128 | 124.55 | potential_loyalists | 2888.92 |
| 3005 | 22 | 149 | 156 | 71.74 | champions | 2925.99 |
| 10133 | 24 | 218 | 234 | 93.96 | champions | 2956.30 |
| 2186 | 21 | 178 | 208 | 102.59 | loyal_customers | 3089.57 |
| 5350 | 17 | 204 | 233 | 140.28 | loyal_customers | 3190.55 |
| 5735 | 17 | 204 | 233 | 140.28 | loyal_customers | 3190.55 |
| 477 | 21 | 78 | 109 | 98.05 | loyal_customers | 3904.41 |

From the table, the top 10 customers all belong to the Champions/Loyal customers/Potential loyalist segments.

Finally, the average Long Term Value of each RFM segment defined earlier is calculated for the next 12 months.

Table 11

| Segment | LTV |
|---|---|
| Champions | 801 |
| Loyal customers | 658 |
| Cant loose them | 403 |
| At risk | 355 |
| Need attention | 200 |
| Potential loyalist | 197 |
| Lost | 65 |
| New customers | 62 |
| Promising | 61 |
| About to sleep | 54 |

***Table 11*** suggests that the segments with the highest LTV values are the Champions, followed just after by the Loyal customers. The RFM segmentation along with the LTV model can be used to develop a classification model and determine which customers are most likely to be receptive to our next promotional marketing campaign.

# 13 Campaign Design

## 13.1. Scenario and objectives

After Covid 19, stores face risks of which competitors cannibalizing their market share. To tackle this issue, the commercial team wants to announce a new campaign aiming to promote customer relationships, enhance customer loyalty and exploit high net worth value customers under the $10,000 budget limitation.

## 13.2. Campaign strategy

For the most profitable tier 1 customers, the marketing team will hand over the promo code via both email and SMS. This promo code will help analysts to understand which tier they belong to, and which campaign is in effect. In this new campaign scoring in April, teams will send $100 voucher when spent more than $200, with no time limit. The voucher expires in September, so the performance period ends in October as one month lagging.

To-be-maintained customers will have same format but only received Email reminders with the voucher $30 off when spent more than $100. Same expiration date as tier 1 customers.

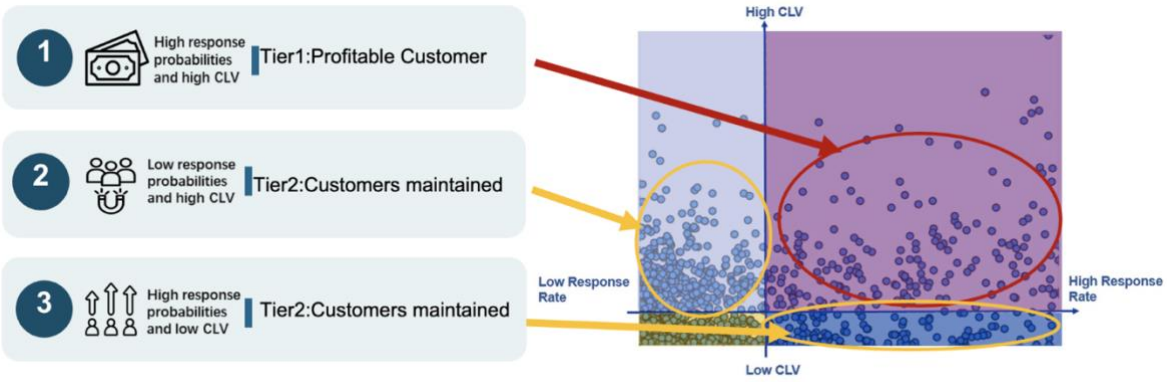


Figure 20

## 13.3. Classification and targeting

### 13.3.1. Objective 1: Exploit high net worth value customers

To achieve this objective, CLTV will help us have a deep understanding of the value of the customers with consideration of the discount rate of time. And RFM analysis will give us the quality of the orders from each customer, team will utilize both models to determine whether it is a profitable customer.

### 13.3.2. Objective 2: Increase sales and increase customer loyalty

With the help of the RFM model and propensity model which has the response rate as the dependent variable, it helps to reach the customers with high response rates and worth to spend on, customers fall into this classification will be our to-be-mentioned tier 2 customers.

### 13.3.3. Objective 3: Optimize the budget usage (Assume $10k)

To maximize the utility of the budget, all the classification models will help to reduce the unnecessary cost spend and enhance the return of investment.

### 13.3.4. Measurement Criteria:

1. Return of Investment > $1.3 per dollar

2. Cost Per Conversion < $30

3. Total budget of this campaign will be $10,000

## 13.4. Time planning for campaign

- Observation Period: Oct 22 – Mar 23
- Scoring month: Apr 23 (Now)
- Performance: May 23 - Oct 23

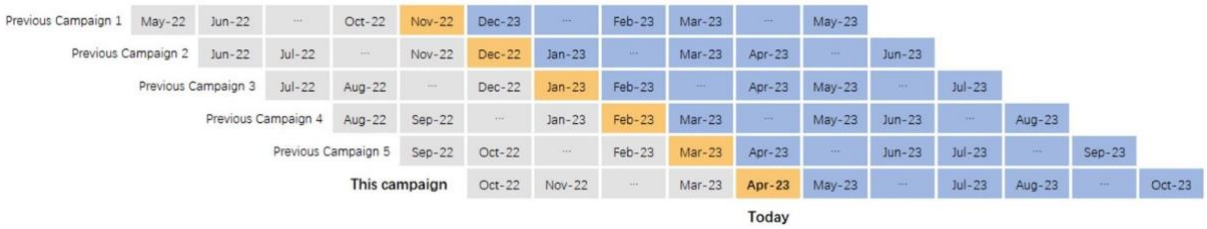


Figure 21

Tagging during the performance:

- Good (Tag 1): New purchases and entered the promo code we sent (Accept new campaign)
- Bad (Tag 0): New purchases without entering the promo code we sent (Not brought by the new campaign)

## 13.5. Test and Experimentation

Before launching the campaign, the commercial team will simulate the campaign by splitting the data into control and test group as shown below. This will help the leader to get a better understanding of the effectiveness of this campaign and the models we used.

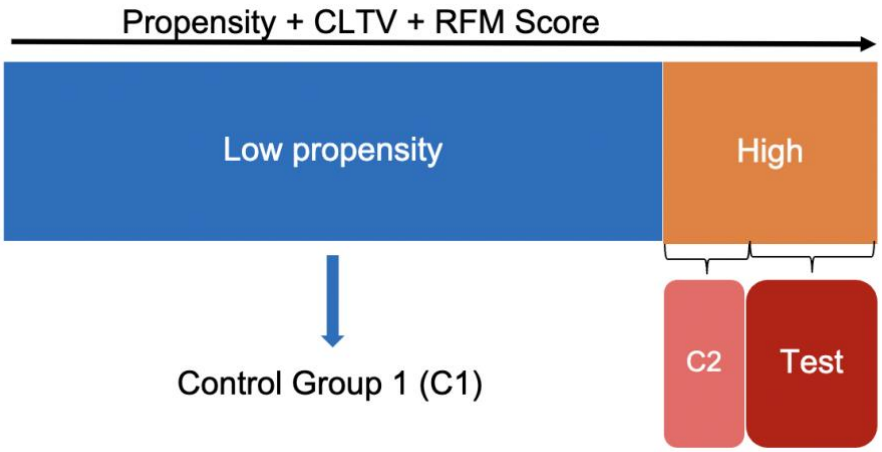


Figure 22

This experimentation will take all these models into consideration, and to examine the Model Effectiveness, the team will treat the control group 1 and 2 business as usual (BAU), by comparing these two results we will have a basic understanding of the performance of this classification.

To examine the campaign's Effectiveness, take the contrast of the test group and the second control group (C2), the lift of the campaign will be recorded.

## 13.6. Dashboarding and Result monitoring

Final deliverable of the campaign is the dashboards which monitoring all the crucial performance index of the campaign, including:

AB testing for the sales vs the conversion rate as the figure shows below.

Key index monitoring:

- ROI=Campaign Rev-Campaign Cost
- CPC= Cost/#conversion


Figure 23

Basic Metrics (In time):

Volume, Reach, Response Profitability, Viewability, Change%

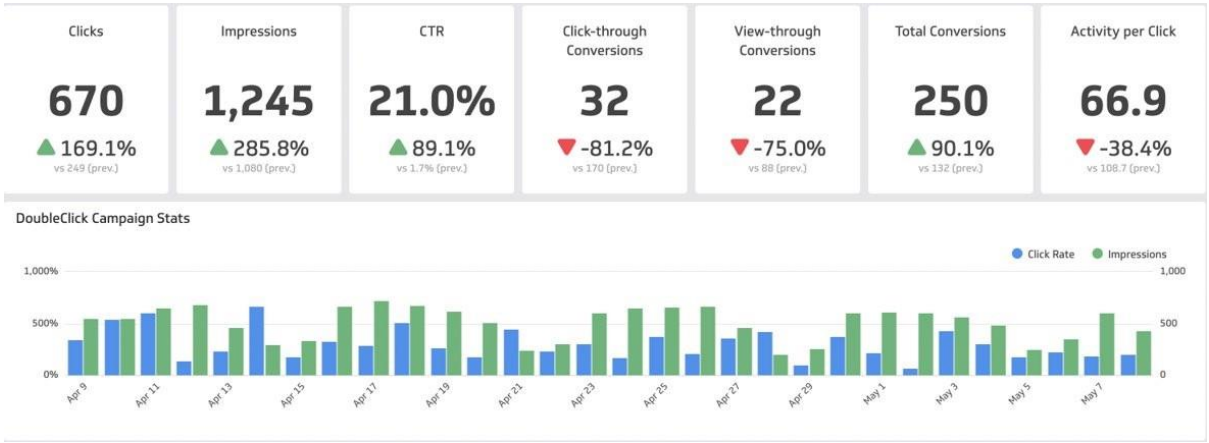


Figure 24

# 14 Conclusions

- There are four main types of customer group. Tesco should focus on the good customer group and the elite group. The previous group consists of individuals who have high spending habits as well as higher education. They are the main consumers of daily necessities. The elite group consists of the highest incomes and highest spending habits group, also they are the major buyers of gold and wine.

- From the RFM model and CLTV calculation, champions and loyal customers are crucial to Tesco's total revenue as they contribute more than half of the money. On the one hand, Tesco should maintain a good relationship with loyal customers. However, for sake of long-term planning, customers that fall into the about-to-sleep and the hibernating group have to be reactivated through a subsequent campaign or other marketing activities.
- In the quadrant chart consisting of Response and CLV, the majority of customers possess low campaign response rates and CLV, falling within the fourth quadrant. Conversely, customers with high response rates and CLV in the first quadrant are more dispersed, yet they contribute most profits. Therefore, the company should implement customized strategies targeting this segment of customers.
- The commercialised product of this analysis is the campaign plan, which utilize the result from all the models of this project despite the customer profiling which is required by management team. Campaign aggregates all the analysis to achieve the objectives set in the beginning and proposed the feasible measurement to monitor the performance of both models and campaigns.
- Risk and mitigation
  a) Model only use the historical data which may contain bias and ad hoc spikes, to mitigate this bias, team may need continually updates the models and campaigns.
  b) Due to the limited size of the dataset and slight class imbalance in sample distribution, the recall rate of the model's classification results is not particularly high.
  c) The response rate in reality could be much lower than the experimentation due to cannibalization of competitors, once happened may require much strong campaign may cost more budget.

# 15 References

Daniel McCarthy, Edward Wadsworth, November, 2014 ,"Buy 'Til You Die - A Walkthrough"
https://www.kaggle.com/datasets/ahsan81/superstore-marketing-campaign-dataset