Creating the next BIG HIT.

Team: Group 4

LIN FANGZHOU(A0261850H)
YAN ZIHAN(A0261738X)
CHEN YUMENG (A0261899H)
SHI KECHEN (A0261672A)
SU CHEN (A0261760H)

TESCO

Introduction

1. Industry Overview
2. Business Problems

1.1 Industry Overview

The supermarket industry is a highly competitive and dynamic sector

- Impacted by the rise of e-commerce, online shopping, as well as changes in consumer behavior
- As the economy recovered, physical stores receive an onslaught from competitors

Food retailer in Europe, with its headquarters in the **UK**

Voted Britain's Favorite Supermarket

7 years running

By customers in the Grocer Gold Awards

1.2 Business Problems

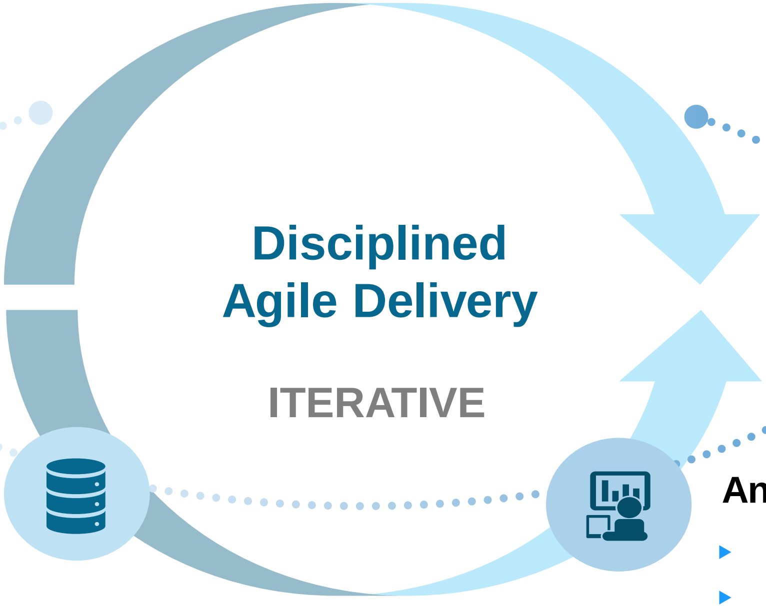
Keep business growing & Retain loyal customers

1. Deepen customer understanding
2. Discover potential profits and growth opportunities
3. Campaign design

Project Design

1. Project Management Plan
2. Data Preparation

2.1 Project Management Plan



Business Understanding

- Project scope definition
- Resourcing and scheduling
- Industry overview

Data Preparation

- Data exploration
- Data cleaning
- Data transformation

Analytics Stage

- Clustering analysis & customer profiling
- RFM model to segment customers
- propensity modeling
- Calculate customers' CLV
- Design dashboard to monitor campaign

Delivery & Presentation

- Insight summary
- Demo presentation
- Report writing

2.2 Data Preparation

1. Data Exploration

Determine which data to use

2. Data Cleaning

Correct, impute or remove erroneous value

3. Data Transformation

Derive new attributes for subsequent analysis

2.2 Data Exploration

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

Data Analysis

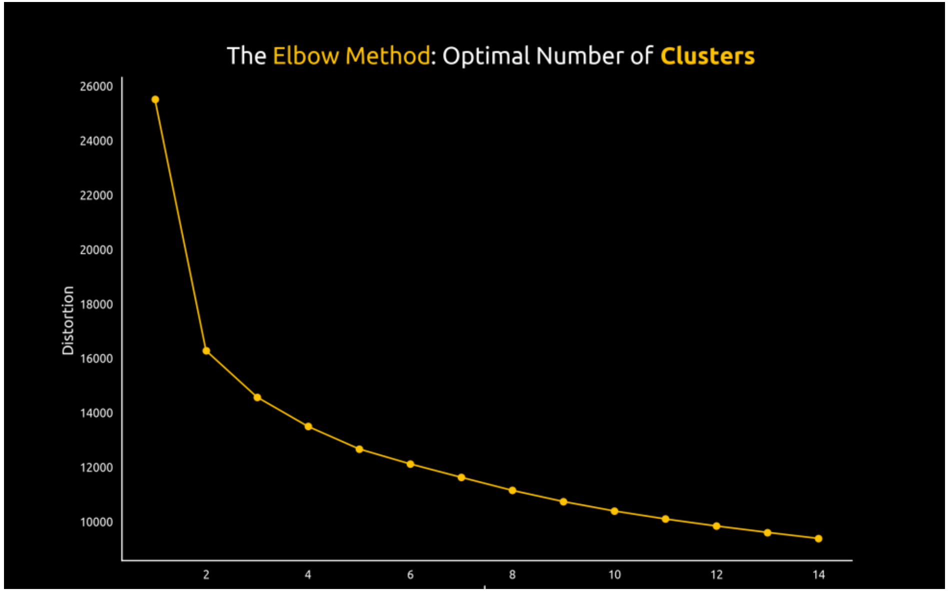
1. K-Means
2. Customer Profiling
3. RFM Model
4. Propensity Model
5. CLTV Model

3.1 K-Means

- Just use Demo_info and Buying_info in clustering
- Exclude categorical variables when clustering: [Education] , [Marital Status]

	Year_Birth	Income	Wines	Fruits	Meat	Fish	Sweet	Gold	Children	Expenses	Time_Enrolled_Days
count	2240	2240	2240	2240	2240	2240	2240	2240	2240	2240	2240
mean	1969	52247	304	26	167	38	27	44	0.95	606	538
std	12	25038	337	40	226	55	41	52	0.75	602	232
min	1893	1730	0	0	0	0	0	0	0	5	26
25%	1959	35539	24	1	16	3	1	9	0	69	367
50%	1970	51742	174	8	67	12	8	24	1	396	539
75%	1977	68290	504	33	232	50	33	56	1	1046	711
max	1996	666666	1493	199	1725	259	263	362	3	2525	1089

3.1 K-Means



Cluster Comparison		
N cluster	CCC	Best
3	1.83052	
4	2.01093	Optimal CCC
5	-1.7331	
6	2.74849	

Choose cluster=4

3.1 K-Means

K Means NCluster=4

Columns Scaled Individually

Cluster Summary

Cluster	Count
1	501
2	609
3	455
4	667

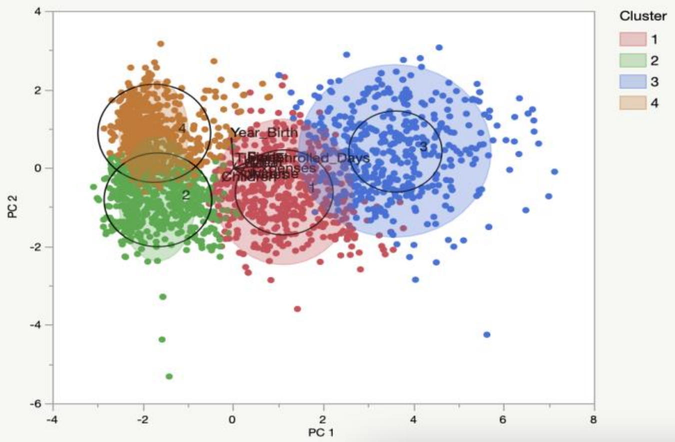
Step	Criterion
20	0

Cluster Means

Cluster	Income	Wines	Fruits	Meat	Fish	Sweet	Gold	Children	Expenses	Time_Enrolled_Days	Year_Birth
1	64970.8348	597.682635	29.2974052	207.213573	39.2075848	29.8123752	71.8143713	0.79640719	975.027944	578.650699	1964.63473
2	45361.2418	130.009852	5.00821018	38.5188834	7.39901478	5.09359606	19.4170772	1.73563218	205.446634	467.090312	1962.70772
3	75516.9912	596.142857	79.8131868	487.595604	115.727473	82.8769231	82.2703297	0.24835165	1444.42637	565.606593	1969.47033
4	31072.551	46.0929535	7.25337331	29.7826087	10.8215892	7.29235382	20.0029985	0.83208396	121.245877	553.017991	1977.009

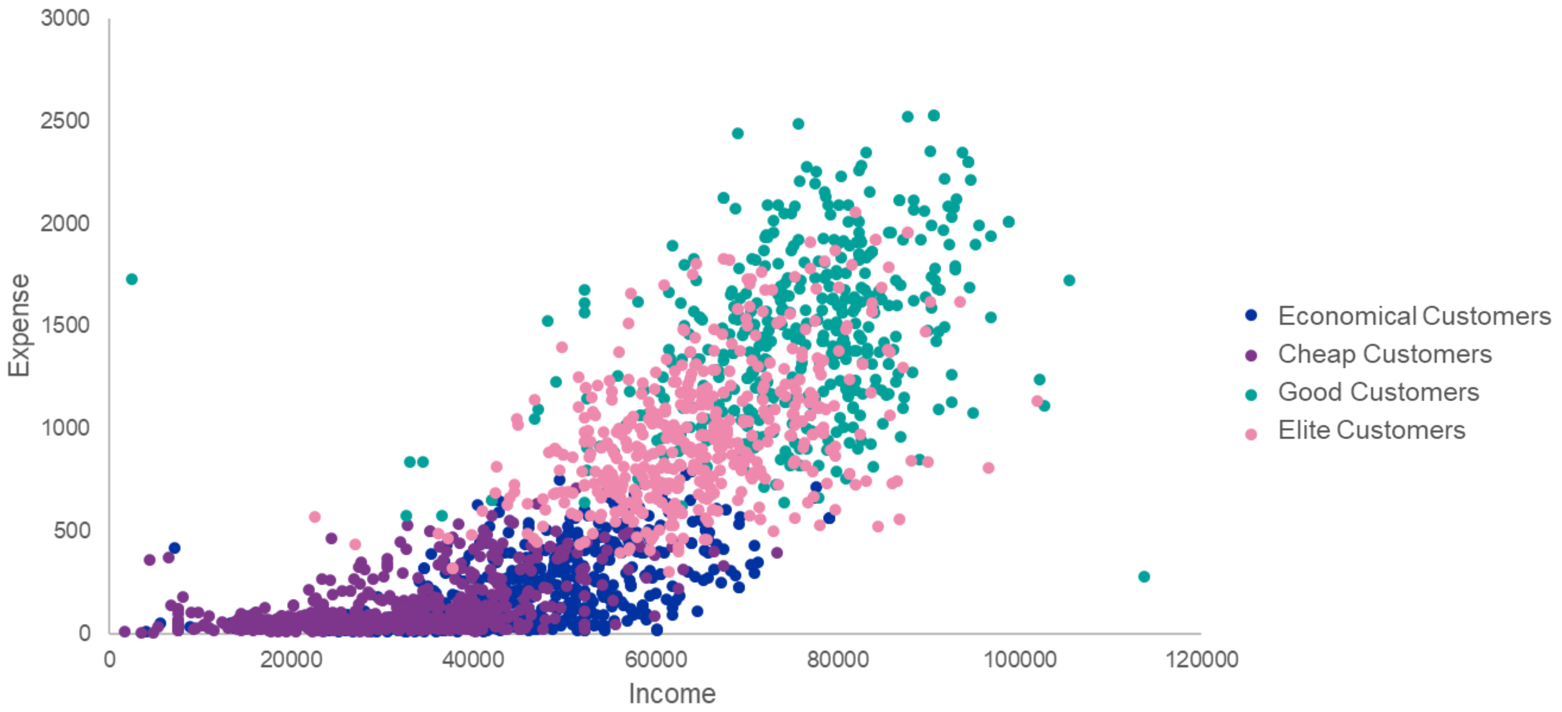
Cluster Standard Deviations

Biplot



3.2 Customer Profiling

Clusters by income & expense

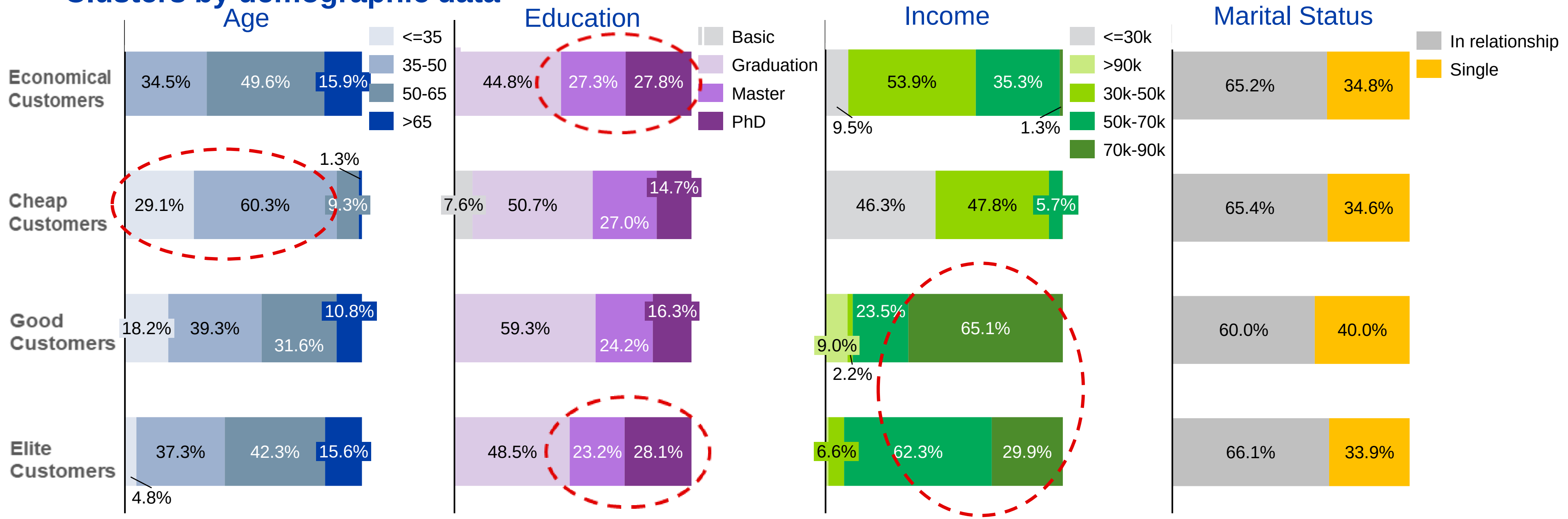


TESCO

3.2 Customer Profiling

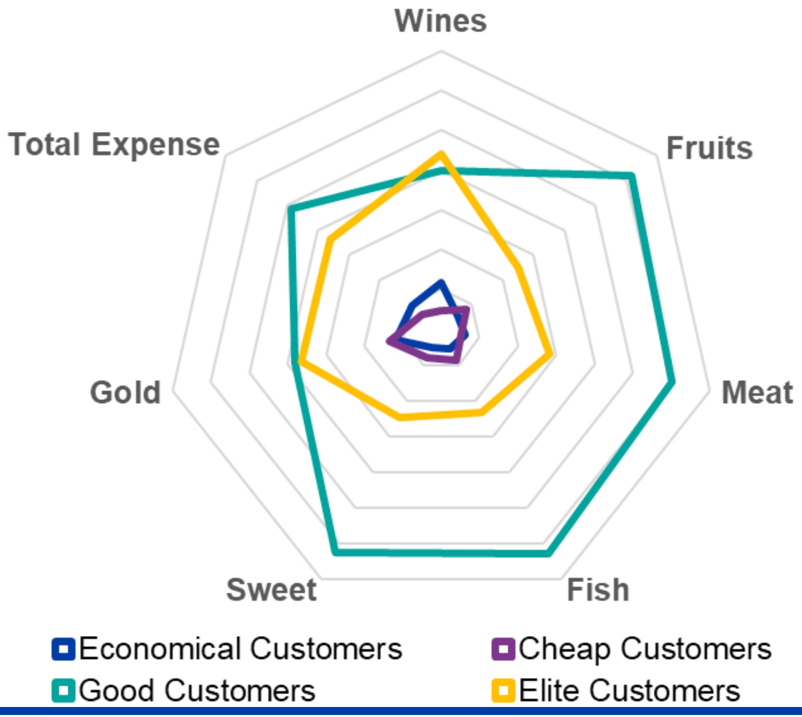
Cheap customers are the youngest; Elite customers mainly include elites with high education and high income

Clusters by demographic data



3.2 Customer Profiling

Value Spend Percentage (%) of Clusters (The proportion of each category that is allocated to a particular cluster)



- Good Customers are 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 group show low spending habits

3.2 Customer Profiling

Good Customers

- High incomes and high spending habits
- high education
- main consumers of daily necessities

1

Elite Customers

- Highest incomes and highest spending habits
- major buyers of gold and wine

2

Economical Customers

- Lower incomes and low spending habits

3

Cheap Customers

- Lowest spending habits
- youngest group

4

3.3 RFM Model

Data Preparation

Data used: loyal_info, purchase_behave_info, buying_info

- Recency - Number of days since customer's last purchase
- Frequency - Total number of transactions
- Monetary - Total transactions value

```
def get_rfm_scores(dataframe):

    dataframe["recency_score"] = pd.qcut(dataframe["Recency"], 5, labels=[5, 4, 3, 2, 1])
    dataframe["frequency_score"] = pd.qcut(dataframe["Frequency"], 5, labels=[1, 2, 3, 4, 5])
    dataframe["monetary_score"] = pd.qcut(dataframe["Monetary"], 5, labels=[1, 2, 3, 4, 5])
    dataframe["RFM_SCORE"] = dataframe["recency_score"].astype(str) + dataframe["frequency_score"].astype(str)

    return dataframe

get_rfm_scores(rfm_data)
rfm_data.head()
```

	ID	Recency	Frequency	Monetary	recency_score	frequency_score	monetary_score	RFM_SCORE
0	5524	58	25	1617	3	5	5	35
1	2174	38	6	27	4	1	1	41
2	4141	26	21	776	4	4	4	44
3	6182	26	8	53	4	2	1	42
4	5324	94	19	422	1	4	3	14

RFM Scoring

5 clusters for each RFM metrics leading to 5x5x5 clusters

Hurdle Rules

More than 1 purchase, spend>10

3.3 RFM Model

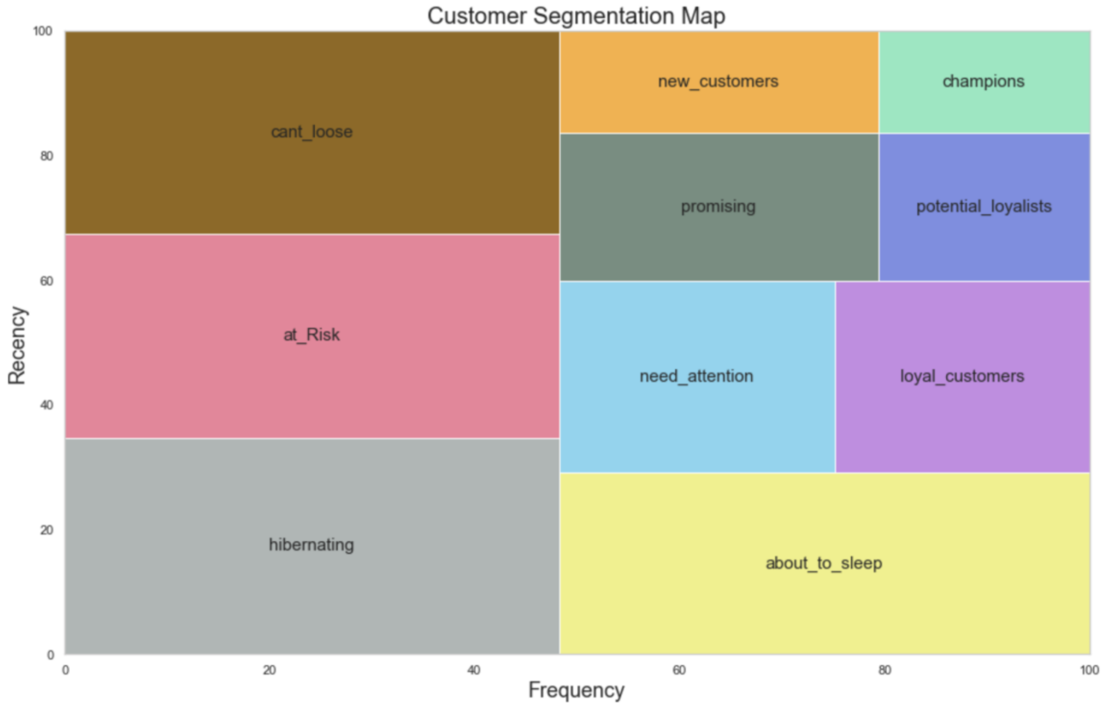
Customer Segmentation

Customer belongs to one of the ten segments

- *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.*

3.3 RFM Model

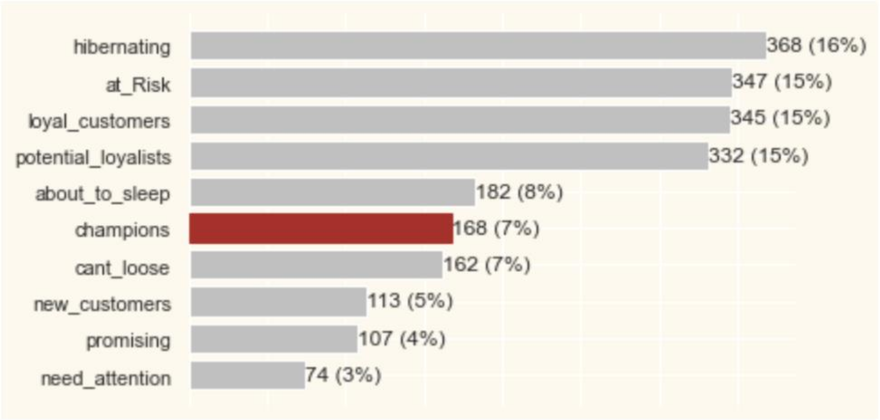
Segment Visualization



3.3 RFM Model

Segment Visualization

	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



3.4 CLTV Model

Model Introduction*

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

BG/NBD Model

- 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*

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

**more assumptions and details for the model can be found in the appendix*

3.4 CLTV Model

CLTV Estimation

Top 10 CLTV customers
(LTV for next 12 months, assume a monthly discount rate of 1%)

	ID	Frequency	Recency	Age	Monetary_value	segment	LTV
333	1826	14	110	110	79.33	potential_loyalists	2806.89
686	8029	14	131	158	108.07	potential_loyalists	2832.24
1486	9264	21	159	160	74.45	champions	2853.27
1138	7959	10	109	128	124.55	potential_loyalists	2888.92
1046	3005	22	149	156	71.74	champions	2925.99
1431	10133	24	218	234	93.96	champions	2956.30
1838	2186	21	178	208	102.59	loyal_customers	3089.57
1544	5350	17	204	233	140.28	loyal_customers	3190.55
1159	5735	17	204	233	140.28	loyal_customers	3190.55
1681	477	21	78	109	98.05	loyal_customers	3904.41

3.4 Propensity Model

We use **Logistic Regression** to generate the **Propensity Score**

Model1: Origin Dataset

Source	LogWorth	PValue
Recency	16.084	0.00000
Marital_Status	15.479	0.00000
AcceptedCmp3	9.942	0.00000
Teenhome	8.282	0.00000
MntMeatProducts	7.775	0.00000
Education	7.568	0.00000
AcceptedCmp5	7.560	0.00000
NumWebVisitsMonth	7.114	0.00000
AcceptedCmp4	3.962	0.00011
AcceptedCmp1	3.527	0.00030
NumDealsPurchases	3.313	0.00049
MntGoldProds	2.966	0.00108
NumStorePurchases	2.256	0.00554
AcceptedCmp2	1.469	0.03397

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

Model2: Over Sampled - SMOTE

Source	LogWorth	PValue
Recency	63.051	0.00000
AcceptedCmp3	26.026	0.00000
Teenhome	25.826	0.00000
NumWebVisitsMonth	22.920	0.00000
AcceptedCmp5	18.123	0.00000
Marital_Status	17.717	0.00000
NumStorePurchases	15.441	0.00000
MntMeatProducts	14.881	0.00000
Education	12.330	0.00000
NumDealsPurchases	10.684	0.00000
NumCatalogPurchases	7.709	0.00000
AcceptedCmp1	6.372	0.00000
AcceptedCmp4	2.843	0.00144
AcceptedCmp2	2.713	0.00194
NumWebPurchases	2.571	0.00269
MntSweetProducts	2.273	0.00533

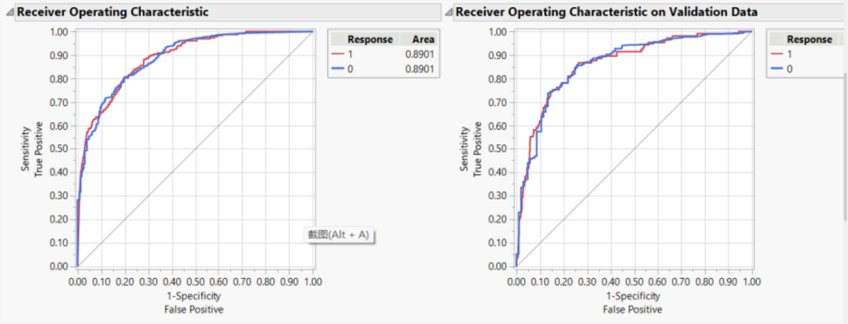
Parameter Estimates

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*

The dataset is relatively small and imbalance (Response=1 accounts for 14%)

3.4 Propensity Model

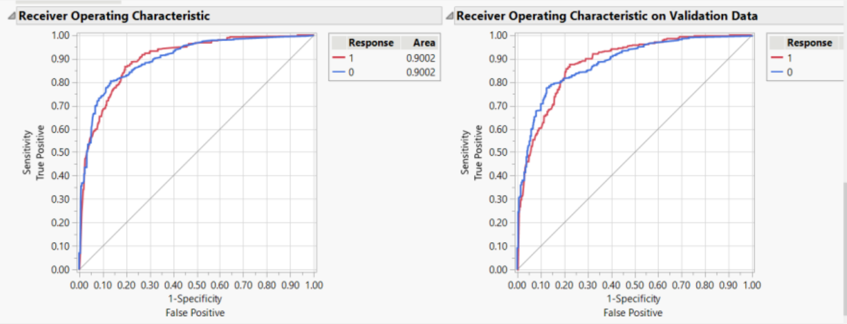
Model1: Origin Dataset



AICc	848.726
BIC	949.816

A balance between better goodness of fit and model complexity.

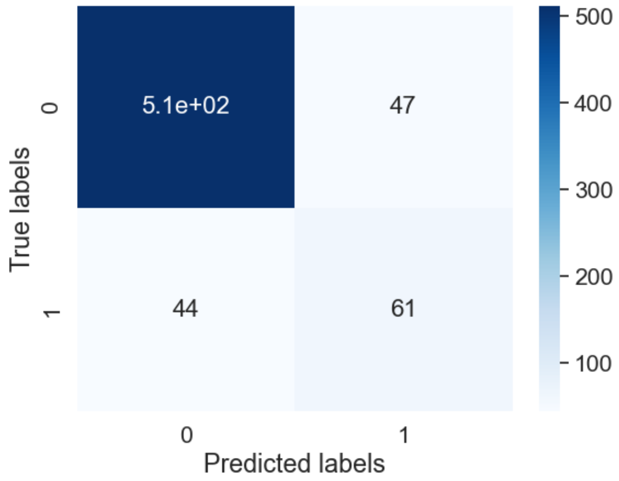
Model2: Over Sampled - SMOTE



AICc	2194.49
BIC	2317.55

3.4 Propensity Model

Confusion Matrix of Test Data (Cutoff: 0.3)



Accuracy	86.3%
Recall Rate	58.1%
Precision	56.5%

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	
8180	0	0	0.047543	need_at	
2569	0	0	0.047652	promisin	

According to the calculated CLV and the Segment

TESCO

3.4 Propensity Model

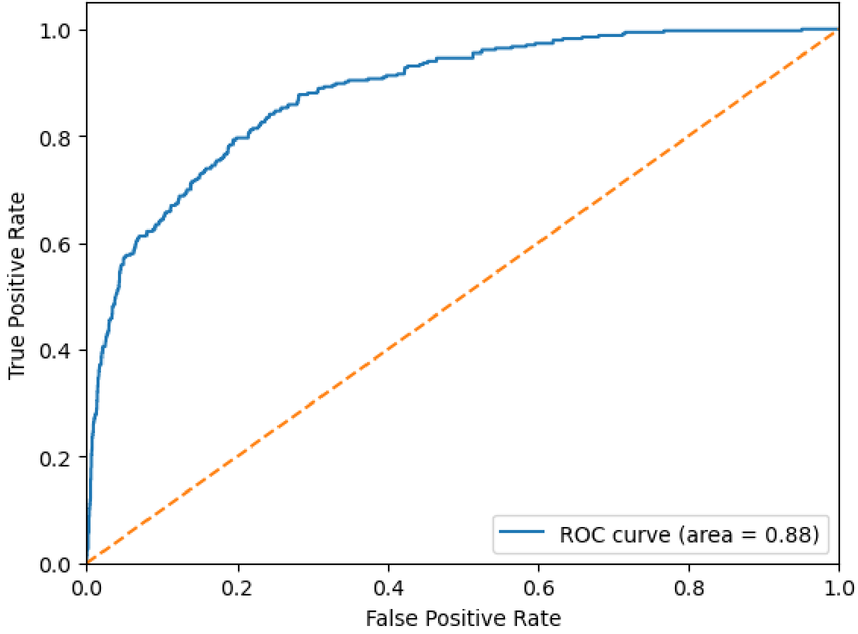
Score Band	Total no. of customers	% Total Customers	Cum % of Customers	Good Customer	Good Customer %	Cumulative Good Customer %	Lift	Good Customer Coverage %	Cumulative Good Customer Coverage	CLV
1	23	1.04%	1.04%	19	82.61%	82.61%	5.49	5.71%	5.7%	1380.178
2	18	0.81%	1.85%	15	83.33%	82.93%	5.54	4.50%	10.2%	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%		
8	34	1.53%	7.58%	23	67.65%	77.38%	4.50	6.91%		
9	9	0.41%	7.99%	5	55.56%	76.27%	3.70	1.50%		
10	29	1.31%	9.30%	15	51.72%	72.82%	3.44	4.50%		
11	28	1.26%	10.56%	14	50.00%	70.09%	3.33	4.20%		
12	38	1.72%	12.28%	21	55.26%	68.01%	3.68	6.31%		
13	39	1.76%	14.04%	8	20.51%	62.06%	1.36	2.40%		
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%		
16	81	3.66%	22.03%	18	22.22%	47.75%	1.48	5.41%		
17	108	4.88%	26.91%	20	18.52%	42.45%	1.23	6.01%		
18	167	7.54%	34.45%	29	17.37%	36.96%	1.16	8.71%		
19	384	17.34%	51.78%	29	7.55%	27.11%	0.50	8.71%		
20	1068	48.22%	100.00%	22	2.06%	15.03%	0.14	6.61%		
Total	2215	100%		333.00	15.03%		1.00	100.00%		

The first 18 bands cover 84.7% of customers with a high probability of responding to the campaign

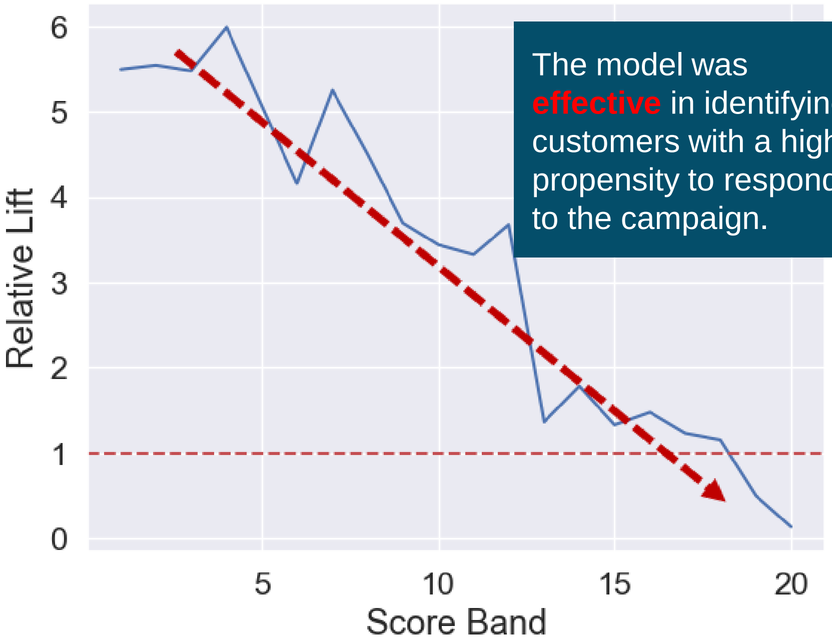
The remained 65.5% of customers only have about 15% of customers with a high probability of responding to the campaign

3.4 Propensity Model

ROC



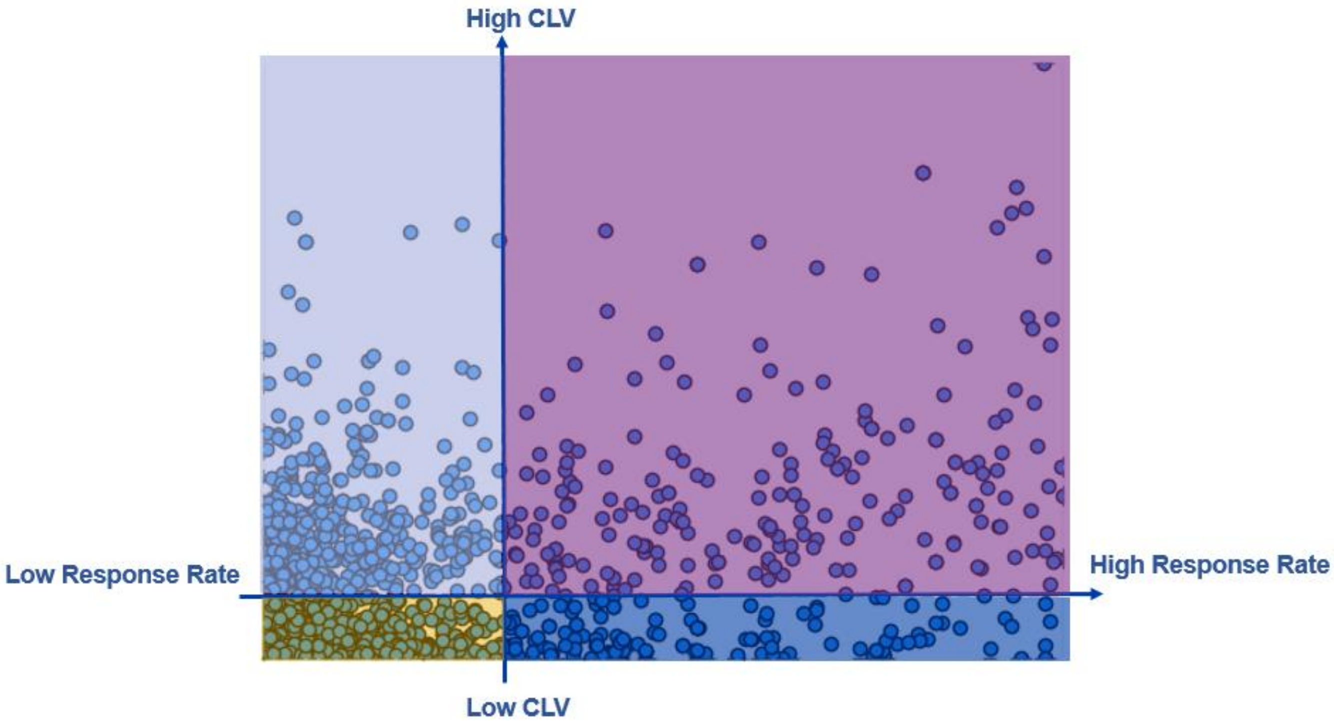
Lift Chart



The model was effective in identifying customers with a high propensity to respond to the campaign.

3.4 Propensity Model

We combined Response and CLV, divided all customers into four quadrants.



Deliverables

1. Campaign Design
2. Dashboard

4 Campaign Plan

1. Campaign Objectives
2. Customer Targeting
3. Campaign Test Experimentation
4. Performance Measurement

4.1 Campaign Objectives

1

Exploit high net worth value customers

- CLTV and RFM analysis

2

Increase sales and increase customer loyalty

- RFM analysis and propensity model (Y is response)

3

Optimize the budget usage (Assume $10k)

- Require customization for limited customers with high ROI(>$1.3) low CPC(<$30)
- Channel selection based on the CLTV of customers
  - Message + Email for Tier 1 customers
  - Email for the rest of target customers

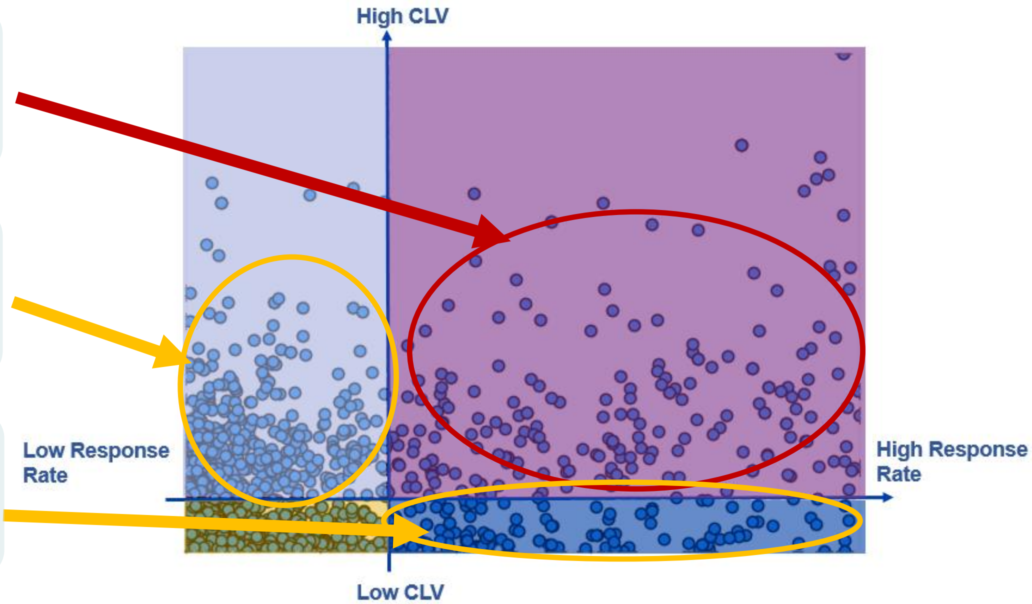
4.2 Customer Targeting

CLTV+Propensity Model

1. High response probabilities and high CLV | Tier1:Profitable Customer

2. Low response probabilities and high CLV | Tier2:Customers maintained

3. High response probabilities and low CLV | Tier2:Customers maintained



- Reduce the budget and enhance the ROI (Objective 3)

4.3 Campaign Design

Ideas: Store utilizes frequency of customer purchases and the average spend per order to extrapolate customer value to differentiate the customer bands. Given the response propensity and combined with RFM score, to customize our content (mainly Promo code) to target to-be-maintained customers to achieve the objectives 1,2: Promote customer relationship and increase sales from profitable customers.

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

Previous Campaign 1	May-22	Jun-22	...	Oct-22	Nov-22	Dec-23	...	Feb-23	Mar-23	...	May-23					
Previous Campaign 2		Jun-22	Jul-22	...	Nov-22	Dec-22	Jan-23	...	Mar-23	Apr-23	...	Jun-23				
Previous Campaign 3			Jul-22	Aug-22	...	Dec-22	Jan-23	Feb-23	...	Apr-23	May-23	...	Jul-23			
Previous Campaign 4				Aug-22	Sep-22	...	Jan-23	Feb-23	Mar-23	...	May-23	Jun-23	...	Aug-23		
Previous Campaign 5					Sep-22	Oct-22	...	Feb-23	Mar-23	Apr-23	...	Jun-23	Jul-23	...	Sep-23	
This campaign						Oct-22	Nov-22	...	Mar-23	**Apr-23**	May-23	...	Jul-23	Aug-23	...	Oct-23
										Today						

Dependent variable:
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)

4.4 Campaign Test Experimentation

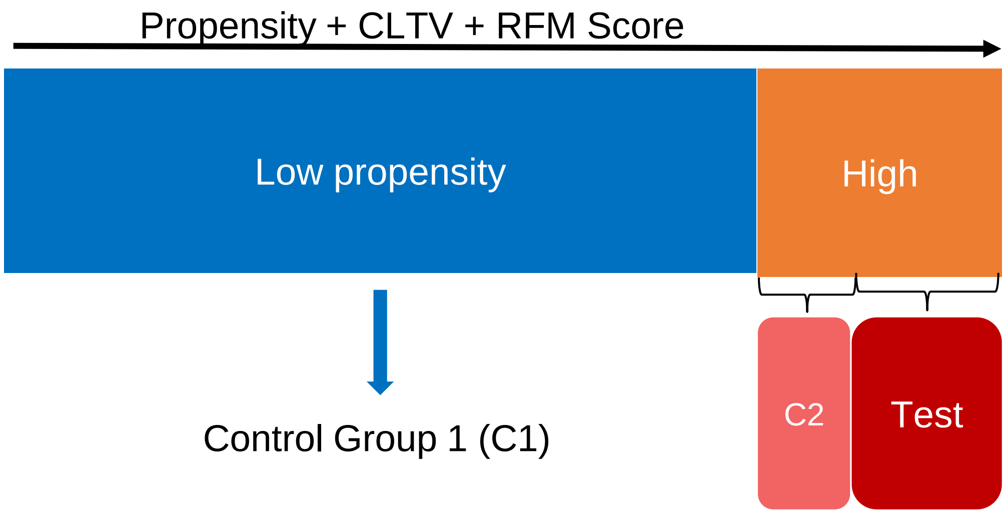
Test, Control Group

Model Effectiveness(BAU)

- Comparsion C1~C2

Campaign Effectiveness

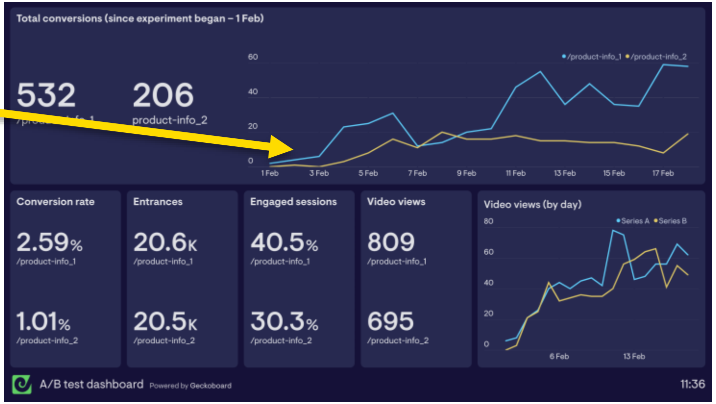
- Comparsion Test~C2



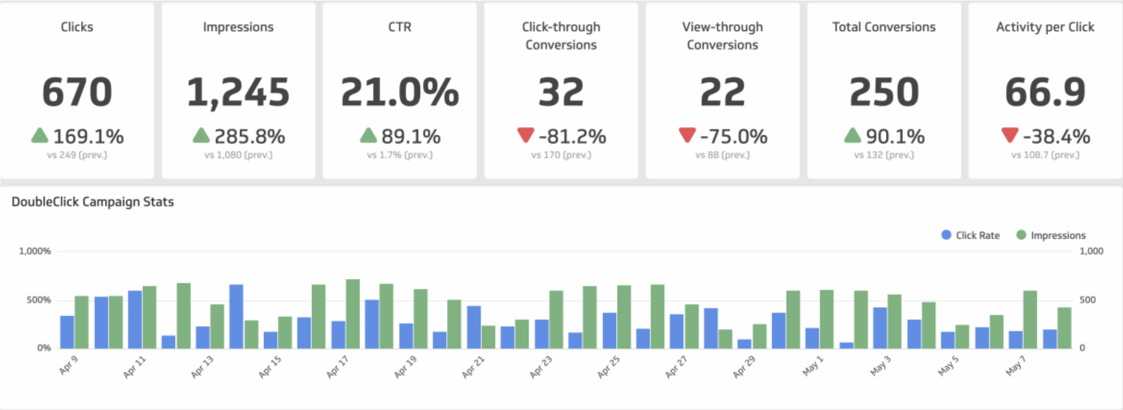
TESCO

4.4 Campaign Measurement

- Conversion Rate ~ Sales(AB testing)
- ROI=Campaign Rev-Campaign Cost
- CPC= Cost/#conversion



Ref: Geckoboard



- Basic Metrics(In time):

Volume, Reach, Response

Profirability, Viewability, Change%

Ref: Klipfolio

TESCO

Appendix

3.4 CLTV Model

Model Introduction

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

Buy Till You Die(BTYD) model is built on 4 metrics which 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 its last purchase. *Recency = Last customer purchase date - first customer purchase date*
- Monetary value : The *average* amount spent by a customer

3.4 CLTV Model

Model Introduction

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

Assumptions for BG/NBD model:

- While active, the number of transactions made by a customer follows a **Poisson distribution** with transaction rate λ
- Heterogeneity in transaction rate λ follows a **Gamma distribution** (each customer has its own probability of buying)
- After any transaction, a customer becomes inactive with probability p. The point at which a customer "drops out" (or "die") is distributed across the transactions according to a **Geometric distribution**
- Heterogeneity in p (dropout probability) follows a **Beta distribution**
- The transaction rate λ and the dropout probability p vary independently across customers

TESCO