AFFILIATE MARKETING

DATA EXPLORATION AND STRATEGY

BUSINESS RECOMMENDATION

BUSINESS OBJECTIVE

Optimization of member performance through segmentation & product recommendation

DATA SET EDA

Exploratory Data Analysis

DATA SOURCE

SALES DATA

- Sponsor(Upline), Member(ent, downline)
- Full year 2021(first half), 2022, 2023(half ?)
- Product master (own generate)

DATA PREPARATION

Data Sanity

Data Preparation for ML

- **Is data ready for use for implement business impact ?**
  1. Create Product master table (SKU, Product name, Price per unit)
  2. Join Transaction table with product master table and validate sales data each year.

DATA PREPARATION

Feature engineering

New features

- Sku_penetrate
- sku_last3m
- sku_last6m
- Sku_amount
- Total_amount
- Total_last_3m
- Total_last_6m
- Ticket_size_3m
- Ticket_size_6m
- Ticket_size

- Transaction_last3m
- Tansaction_last6m
- Total_transaction
- Total_last_3m_online
- Total_last_6m_online
- Total_last_3m_offline
- Total_last_6m_offline
- Total_online
- Total_offine
- total _network
- Mem_duration (months)

Data Sanity & Validation

Explore quality of data each year

1. Missing values ?
2. Duplicate value
3. Re-check sales performance after join product master table

```python
old = df2021['total_amount'].sum()
new = df2021_new['total_amount'].sum()

print('total_amount in 2021     : ' + str(old))
print('new total_amount in 2021 : ' + str(new))
print('diff = ' + str(((old-new)/old) * 100) + ' %')
```

```
total_amount in 2021     : 853354910410.0
new total_amount in 2021 : 852811245409.9998
diff = 0.06370913126157986 %
```

"sponsor" on Whole data COMPUTE - (76160 distinct)

CATEGORICAL

SUMMARY

Not empty 1,849,610 76.9 %
Empty 556,706 23.1 %

Validity cannot be computed on auto-detected meanings.

View variations over time...

Top 10 out of 76160 values	Count	%	Cum. %
No value	556706	23.1	23.1
TCC41ZZ41CB	2299	0.1	23.2
TCCERQCWRCB	2083	0.1	23.3
TCCEJW3EW3I	2010	0.1	23.4
TCCCCCCCCCG	1982	0.1	23.5
TCC4EC3QR40	1827	0.1	23.6
TCC41ZCRCJ0	1815	0.1	23.6
TCC41EQZRWU	1569	0.1	23.7
TWEE1J1K	1435	0.1	23.8
TCCEJZR1WJ2	1399	0.1	23.8

Transaction data

- Member = 77%
- Non-Member 23%

"sponsor" on Sample - (2487 distinct)

CATEGORICAL | VALUES CLUSTERING

SUMMARY

Valid	10,000	100.0 %
Hapax	1,525	15.3 %
Invalid	0	0.0 %
Empty	5,600	56.0 %

1525 HAPAXES 15.3 %

- T11W3J4F
- T13Q4CRB
- T13RWQZ2
- T1E1444F

0 INVALIDS 0.0 %

Top 50 out of 2487 values in sample	Count	%	Cum. %
No value	5600	56.0	56.0
TCC41ZZ41CB	14	0.1	56.1
TCC411R11EI	12	0.1	56.3
TCC4JW4Q1WK	12	0.1	56.4
TCC4QZEQCZU	12	0.1	56.5
TCC4R1JRJJI	12	0.1	56.6
TCC4QJWEZEI	11	0.1	56.7
TCC4J3WR1C7	10	0.1	56.8
TCC4Q4R1WCI	10	0.1	56.9
TCC4Q4RR432	10	0.1	57.0
TCC4QCR31W2	10	0.1	57.1

Free item as 0 Paid Amount

- By Member 44%
- By Non-Member 56%

```
df_trans_new['total_amount'].sum()-df_trans['total_amount'].sum()
✓ 0.0s

-5319555900.0
```

comparing

```
(df_trans_new['total_amount'].sum()-df_trans['total_amount'].sum())/df_trans['total_amount'].sum()
✓ 0.0s

-0.0026475969749030284
```

% dif

```
df_trans[df_trans.duplicated()]
```

✓ 5.7s

	payment_date	ent	center	product_json	total_amount	discount	paid_amount	trans_origin_type	payment_ym
239	2021-01-01	TCC4W4RE31I	T2CEQ1	[{"product":"BC4CC4","qty":1}]	175000.0	0.0	0	online	2021-01
501	2021-01-01	TCC434J33CF	TDCCJE	[{"product":"5C4C4Q","qty":1}]	189000.0	0.0	0	online	2021-01
503	2021-01-01	TCC434J33CF	TDCCJE	[{"product":"5C4C4Q","qty":1}]	189000.0	0.0	0	online	2021-01
671	2021-01-01	TZJRRJRP	TDCCJ4	[{"product":"6CECC4","qty":1}]	1170000.0	0.0	0	online	2021-01
787	2021-01-01	TCC4WRJ43EI	T7C141	[{"product":"5C4CC4","qty":1}]	341000.0	0.0	0	online	2021-01
...	...	...	...	...	...	...	...	...	...
362082	2023-07-06	TCC4ZJRWRE7	TUC1CJ	[{"product":"5C4CCE","qty":2}]	1170000.0	0.0	1170000	online	2023-07
362085	2023-07-06	TCC4ZJRWRE7	TUC1CJ	[{"product":"5C4CCE","qty":2}]	1170000.0	0.0	1170000	online	2023-07
362091	2023-07-06	TCCEC3R14ZU	TUC1CJ	[{"product":"5C4CCE","qty":2}]	1170000.0	0.0	1170000	online	2023-07
362094	2023-07-06	TCCEC3R14ZU	TUC1CJ	[{"product":"5C4CCE","qty":2}]	1170000.0	0.0	1170000	online	2023-07
362107	2023-07-06	TCCEQEZJQ4F	TKC1Z4	[{"product":"KCQCER","qty":1}]	2500.0	NaN	2500	offline	2023-07

14482 rows × 9 columns

Check dup.

```
df_member.isna().sum()
```

✓ 0.1s

```
ent                0
original_status    0
join_month         0
join_year          0
sponsor            0
join_ym            0
dtype: int64
```

```
df_trans.isna().sum()
```

✓ 0.7s

```
payment_date         0
ent                  0
center               0
product_json         9
total_amount         0
discount           140
paid_amount          0
trans_origin_type    0
payment_ym           0
dtype: int64
```

KEY DISCOVERIES FROM EDA

PAID AMOUNT = 0

THAT MUCH PRODUCTS GIVEN FOR FREE?

FREE PRODUCTS > SALES FOR ALL YEARS

All non members receive free items, where all free item to members and nonmembers are 44 / 56

NULL VALUE

There are transactions that once we join with 'members', the value NULL.

JSON FORMAT

FILE FORMATTING

KEY ASSUMPTIONS

PAID AMOUNT REMOVED

Paid amount not taken into account when performing ML

NON MEMBER TRANSACTION

There are instances where there are no members, so we assume these are ‘non-members’

Feature engineering

New features

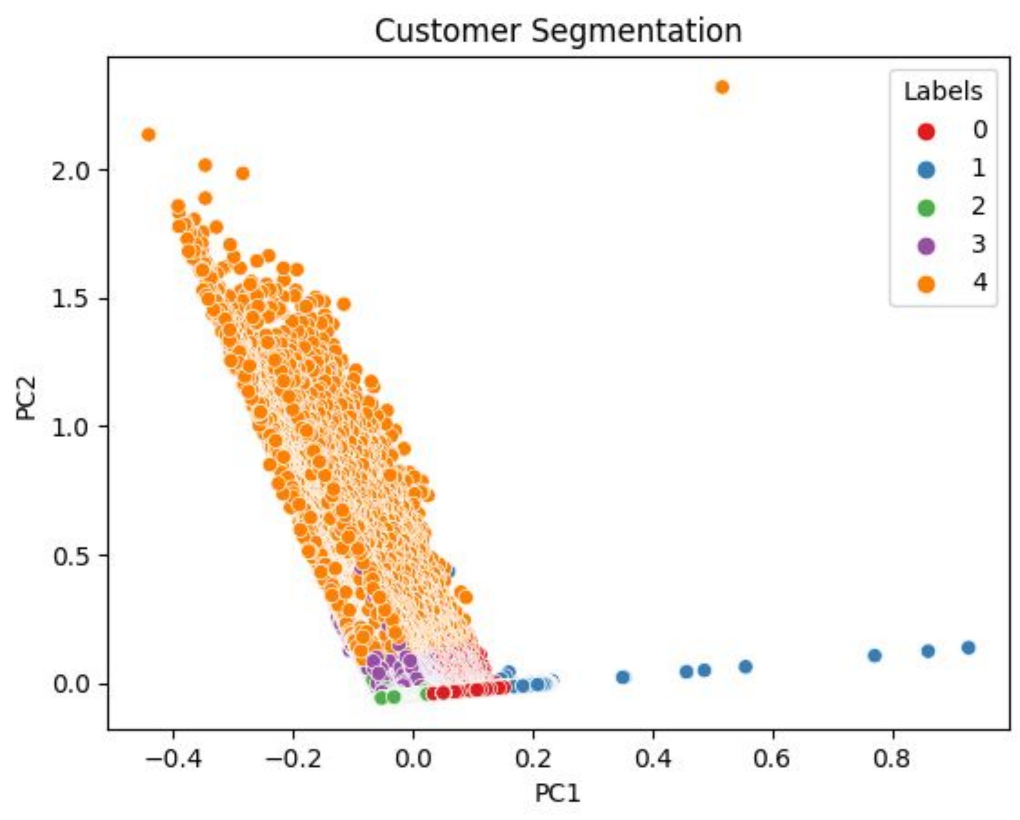
- Sku_penetrate
- sku_last3m
- sku_last6m
- Sku_amount
- Total_amount
- Total_last_3m
- Total_last_6m
- Ticket_size_3m
- Ticket_size_6m
- Ticket_size
- Transaction_last3m
- Tansaction_last6m
- Total_transaction
- Total_last_3m_online
- Total_last_6m_online
- Total_last_3m_offline
- Total_last_6m_offline
- Total_online
- Total_offine
- total _network
- Mem_duration (months)

Machine Learning

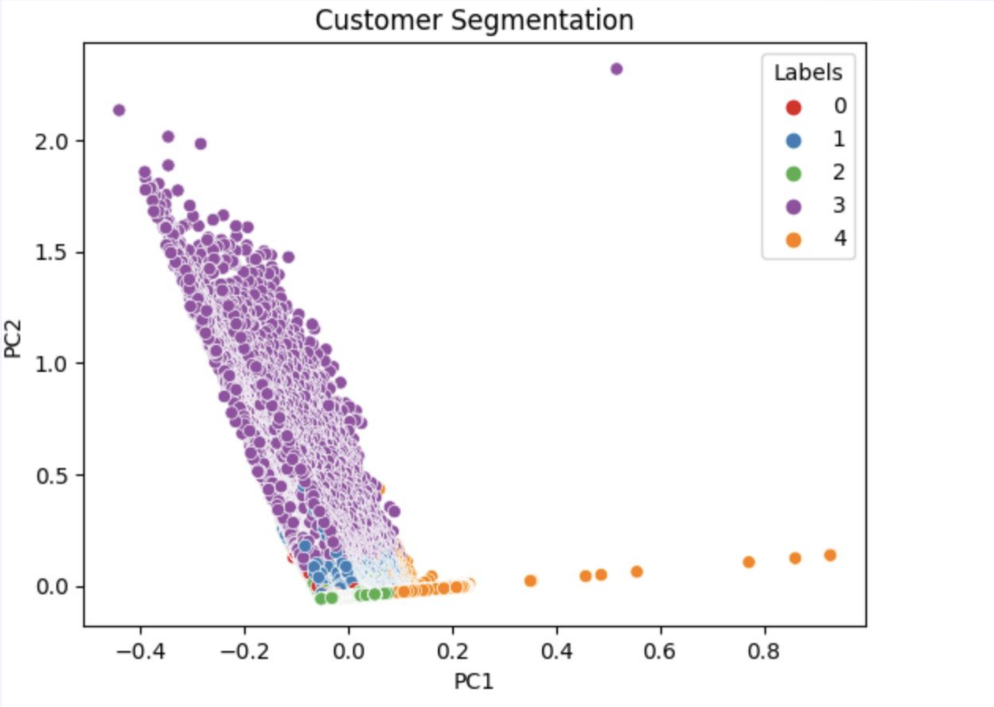
K-Median, keeping outlier to do clustering

1. Alternative way to clustering, extract outliers to do analysis separately or clapping outlier and do K-means

Customer segmentation



K Median


Customer Segmentation
Labels
0
1
2
3
4
2.0
1.5
1.0
0.5
0.0
PC2
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
PC1

Further steps

Dynamics rule based

With clustering can not find definition of each cluster cleary.

Separate sales channel analysis

Clustering in order to separate offline and online

Silhouette Analysis

Unable to run Silhouette Analysis. Therefore, we use Elbow to run K Median.

Clustering Characteristics

	CLUSTER 0	CLUSTER 1	CLUSTER 2	CLUSTER 3	CLUSTER 4
	Major Major	New Star	Survivor	Passive Incomer	Lovely Introductor
Membership Duration		newest		x	
Sum of Sales		x			
Product Variety	x				x
Number of Transaction	x	x			
Offline Sales					x
Online Sales	x	x			
Number of Downline					
Number of Member			x		

Business Strategies

	CLUSTER 0	CLUSTER 1	CLUSTER 2	CLUSTER 3	CLUSTER 4
	Major Major	New Star	Survivor	Passive Incomer	Lovely Introductor
TRAINING PROGRAM			Product Training		
SELL MORE PRODUCT CAT.			X		
GENERATE OWN DOWNLINE					X
EMPLOYEE RECOGNITION		X Achiever Trip		X Award Ceremony Event	
TEAM BUILDING PROCESSES	X Motivate DL				

Further steps

No. of Cluster join back to ‘sponsor’

Integrate clusters back to sponsors to segment sponsors and their downlines

Separate sales channel analysis

Clustering in order to separate offline and online

Silhouette Analysis

Unable to run Silhouette Analysis. Therefore, we use Elbow to run K Median.

Product Recommendation

- Data Discovery
- Product master table
- Basket Analytic
- Cross product
- Personalize promotions
- Further steps

Machine Learning

Each business strategy, which algorithm will we activate, why ?

1. Data for Product recommender (APRIORI)

   Ideally we want to do CosMF, but there isn't enough information [eg. ratings of each products] , we can do rating from sales with decile (1-10) by SKU but too much SKU to do this way.

DATA PREPARATION

For product recommendation

- Create Master Table by using JSON format in order clean the transaction value and further use data mart
- This allows us to acknowledge pricing of SKUs

Show screenshot APRIORI by cluster

```
df_trans_2 shape: (29901, 197)
Association rules for df_cluster_2:
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric

```
df_trans_3 shape: (1046244, 490)
Association rules for df_cluster_3:
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(5C4CCE)	(6CQC41)	0.147148	0.164278	0.030482	0.207154	1.260996	0.006309	1.054079	0.242687
2	(5C4CC4)	(6CQC41)	0.105202	0.164278	0.021256	0.202050	1.229924	0.003974	1.047336	0.208921
1	(6CQC41)	(5C4CCE)	0.164278	0.147148	0.030482	0.185553	1.260996	0.006309	1.047155	0.247662
3	(6CQC41)	(5C4CC4)	0.164278	0.105202	0.021256	0.129391	1.229924	0.003974	1.027783	0.223689

```
df_trans_4 shape: (766618, 544)
Association rules for df_cluster_4:
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
2	(8C4CCR)	(6CQC41)	0.069938	0.227909	0.024412	0.349056	1.531561	0.008473	1.186110	0.373170
6	(7C4CC4)	(6CQC41)	0.074283	0.227909	0.022304	0.300262	1.317464	0.005375	1.103400	0.260302
0	(5C4CCE)	(6CQC41)	0.127657	0.227909	0.037733	0.295584	1.296938	0.008639	1.096072	0.262458
4	(5C4CC4)	(6CQC41)	0.084823	0.227909	0.023464	0.276624	1.213747	0.004132	1.067344	0.192427
1	(6CQC41)	(5C4CCE)	0.227909	0.127657	0.037733	0.165563	1.296938	0.008639	1.045427	0.296537
3	(6CQC41)	(8C4CCR)	0.227909	0.069938	0.024412	0.107115	1.531561	0.008473	1.041636	0.449521
5	(6CQC41)	(5C4CC4)	0.227909	0.084823	0.023464	0.102954	1.213747	0.004132	1.020212	0.228088
7	(6CQC41)	(7C4CC4)	0.227909	0.074283	0.022304	0.097866	1.317464	0.005375	1.026141	0.312095

We can use this data for cluster to collaboratively crossell.

Further steps

Missing “RATING”

Unable to calculate CO-SINE Similarity

Recommend by Category

Too much effort to rate by SKUs for each member. **Better to have “Product Sub-Category**

Cal rating by sales, SKUs → Decile on Sales for Criteria → Set up Criteria for rating by SKUs → Rating by member by SKUs

THANK YOU

MEMBERS

TOTTHONG LERTVANARIN 6510424032
NICHA RONGRAM 6510424013
CHANAPAT CHAINGAM 6510424010
NUTCHAPONG LERTSITHIKARNKOSOL 6510424204
CHANAWUTH WUTHITHADA 6510424014
JIRAPAT ATIKOMTRIRAT 6510412009
PUNNATORN MINGKWAN 6510412003

GENERAL GOALS OF DATA ANALYTICS FOR AFFILIATE MARKETING

Possible Strategies

- **Increase profit**
  - Up sales, Cross sales
  - Optimization campaign or promotions
  - Some product related with some customer cluster or not?

- **Decrease cost**
  - Discount Rate?
  - Free item : 0 amount ?
  - Training cost from Turnover Rate employee?

13-7-66

Customer Segment

- ทำ Clustering ของ ent
- โดยใช้ R(3เดือน,6เดือน,12เดือน) F M
- ใช้ KMean (Mark that less effort, Less sensitive outlier)
- ดู Shap Value แต่ละ Cluster
-

Next Step#1

- ค้นหาความแตกต่างระหว่าง Online,Offline
- แยกวิเคราะห์ Online Offline
- เช่น AVG Price per Unit

Next Step#2

- เอา Cluster ที่ได้ไป join กับ Sponsor
- เพื่อจัดกลุ่ม Sponsor อีกที

13-7-66

Product Segment

- Basket Analysis
- Cross Product or Bundle Product
- Make it personalize promotions

CONTENTS

In this slide, we will explore various aspects of HDI Holding's data

DATA SOURCE	To view this template correctly in PowerPoint, download and install the fonts we used
DATA PREPARATION	An assortment of graphic resources that are suitable for use in this presentation
DATA ANALYTICS	You must keep it so that proper credits for our design are given
BUSINESS RECOMMENDATIONS	All the colors used in this presentation
MOVING FORWARD	These can be used in the template, and their size and color can be edited

LIVE
LEARN
LOVE

Data glossary

- Declare clear cut of business pain point
- Declare definition of each element on data (If any unclear definition, make assumption eg. original status then ML or Segment by performance each original status)

Insight Data

ALL NON MEMBER RECEIVED FREE ITEM

ALL FREE ITEM DELIVERED TO
MEMBER AND NONMEMBER 44/56

Results

CLUSTER 0

High no. of visits → goal is to increase spending per user

We can bundle products to increase spending per tnx

CLUSTER 1

Less than positive results → goal is to shut the store down

Or

Decrease stockage to prevent spoilage

CLUSTER 2

Customers are willing to pay higher prices → goal is to increase customer base

Possible campaigns:
First purchase promotion, loyalty program, targeted marketing

CLUSTER 4

Optimize supply chain and inventory management

Lower cost by implementing cost saving technologies that would reduce labour costs

Trial launch of products