Starbucks Customer Monthly Spending Analysis & Promotion Recommendation Model

Team Members:

Ang Li

Bo Yang

Hesheng Yang

Li Zhuang

Background Introduction

Chosen Topic

- Customer Behavior and Sales Analysis

Customer Behavior is a Broad Concept

- Purchase Decision Process
- Consumption Motivation
- Consumption patterns
- User Experience
- Brand Preferences etc.

Purchase Frequency
Interest
Preference

Goal

- Uncover underlying insights by data visualization
- Build predictive models based on user behavior

Project Overview

Dataset Instruction

Profile: 14,000+ Starbucks customers

Portfolio: 10 promotional schemes

Transcript: 300,000+ user purchase history, promotional participation

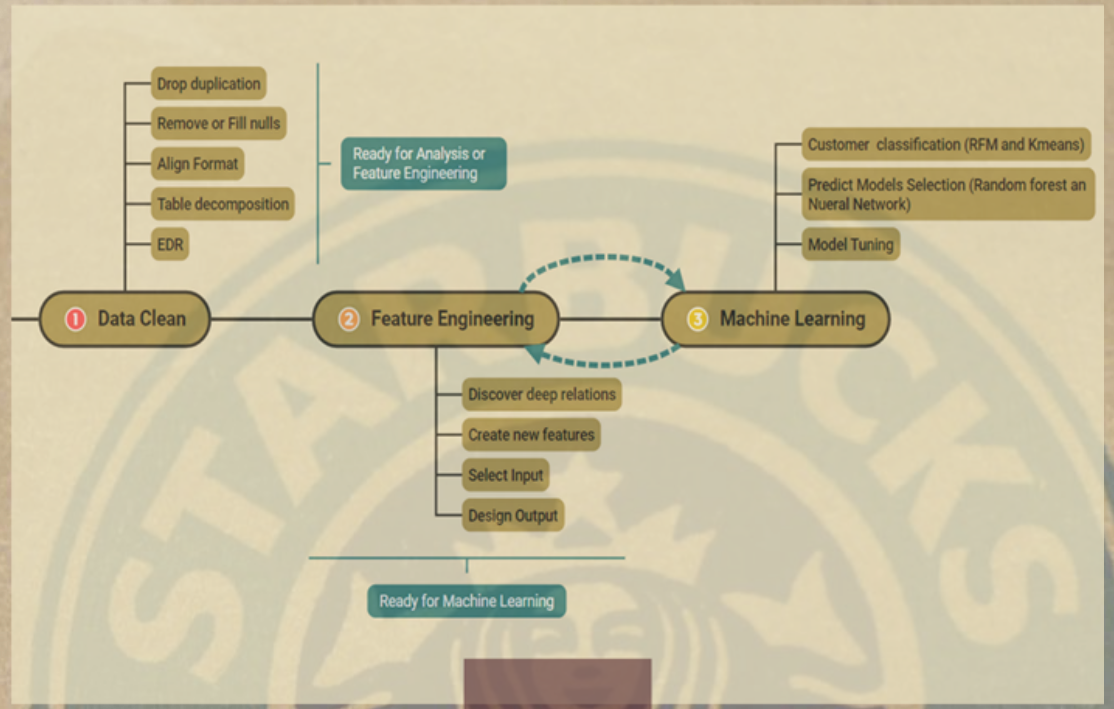
Data

- Age
- Gender
- Income
- Transaction Records
- Promotion Type
- Promotion Duration
- Min Spending

Data Analysis

Insight

- Customer Value
- Frequency
- Preference
- Loyalty
- Sensitivity
- Interesting Level
- Conversion Rate



Promotion Prediction Model

Tools and Techniques Used

	Power BI	Python
	• *Data clean* • *Feature Engineering* • *Visualization*	• *Customer Classification* • *Machine Learning* • *Visualization*
	• *DAX*	• *Pandas* • *Seaborn* • *Scikit-learn*

About Data Set

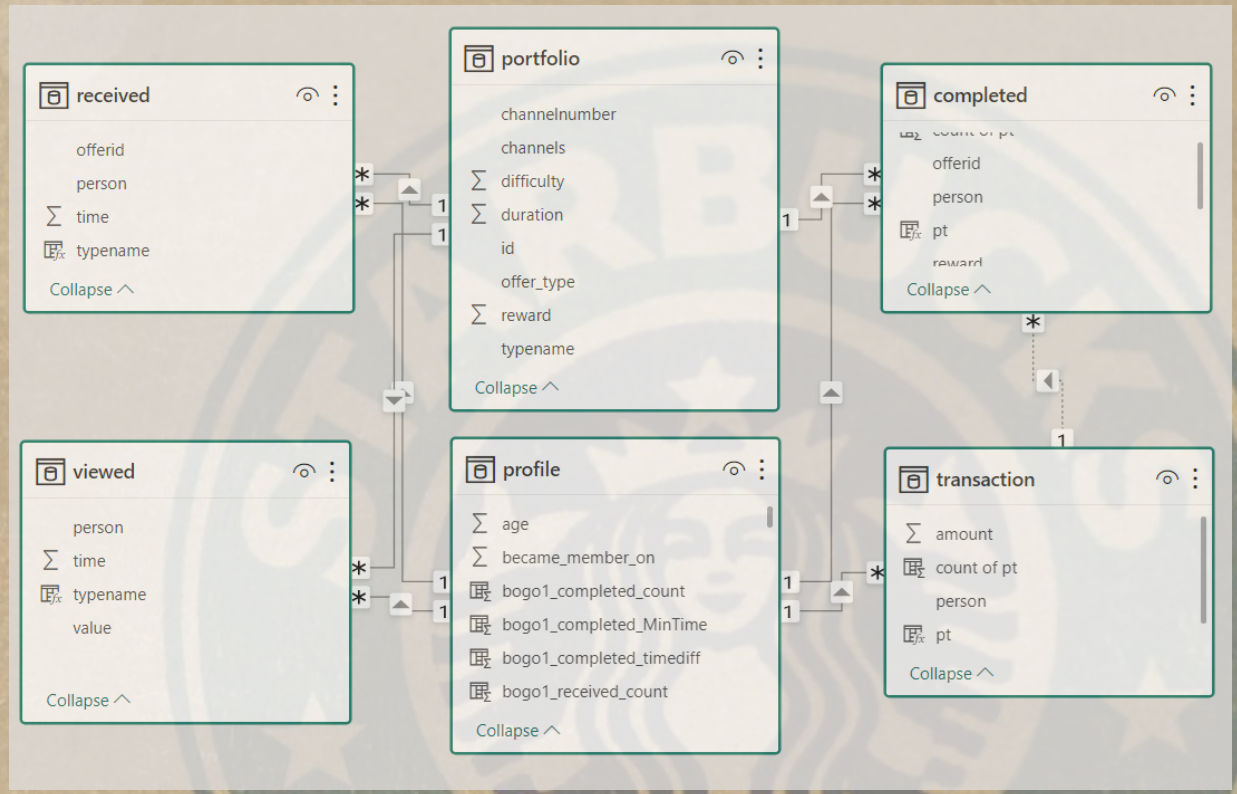
PROFILE	VALUES
Member ID	17,000+
Age	18 - 101
Gender	M/F/O
Income	30,000-120,000
Member Start Date	2013-2018

PORTFOLIO	DISTINCT VALUES
Promotion Type	3
Promotion ID	4
Promotion Duration	5
Minimum spending requirement	5
Promotion Delivery Channel	4
Promotion Reward	5

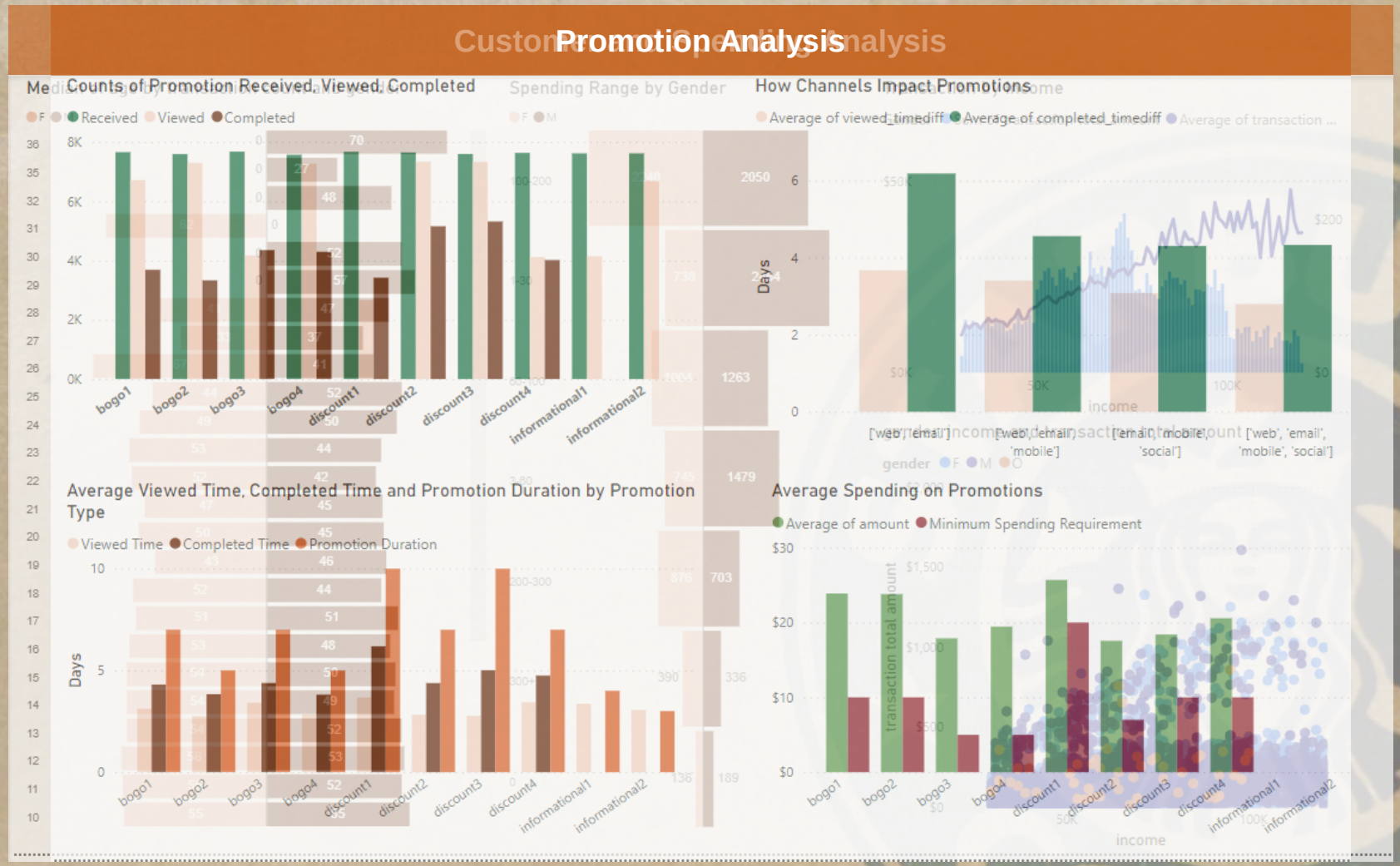
TRASCRIPT
Member ID
Event Type
Time
Promotion ID
Amount

Promotion Received
Promotion Viewed
Promotion Completed
Transaction

Data Modelling



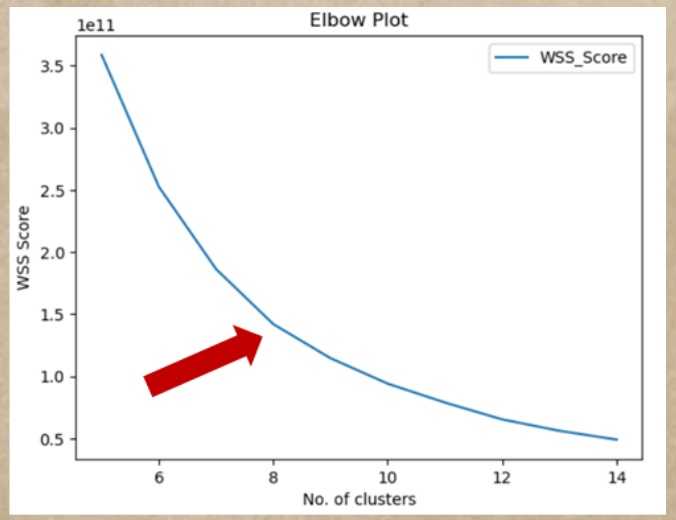
Data Exploration & Feature Engineering

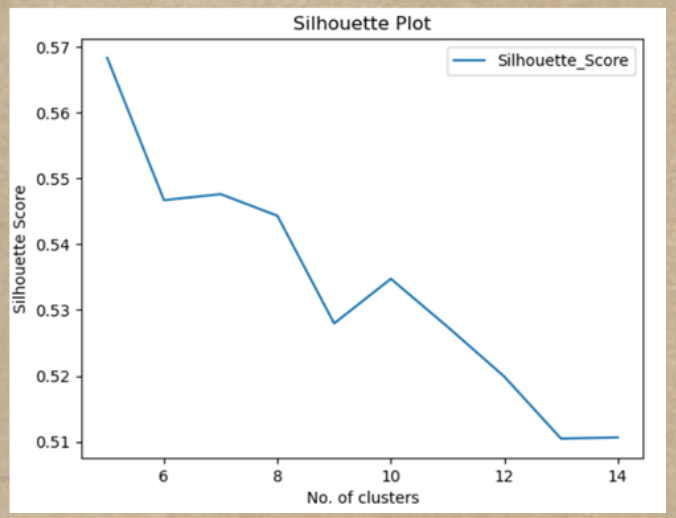


Customer Cluster - Kmean

Input

Age
Gender
Income
Enrollment Date
Transaction Counts
Total transaction Amount





K = 8

Customer Cluster - RFM Model

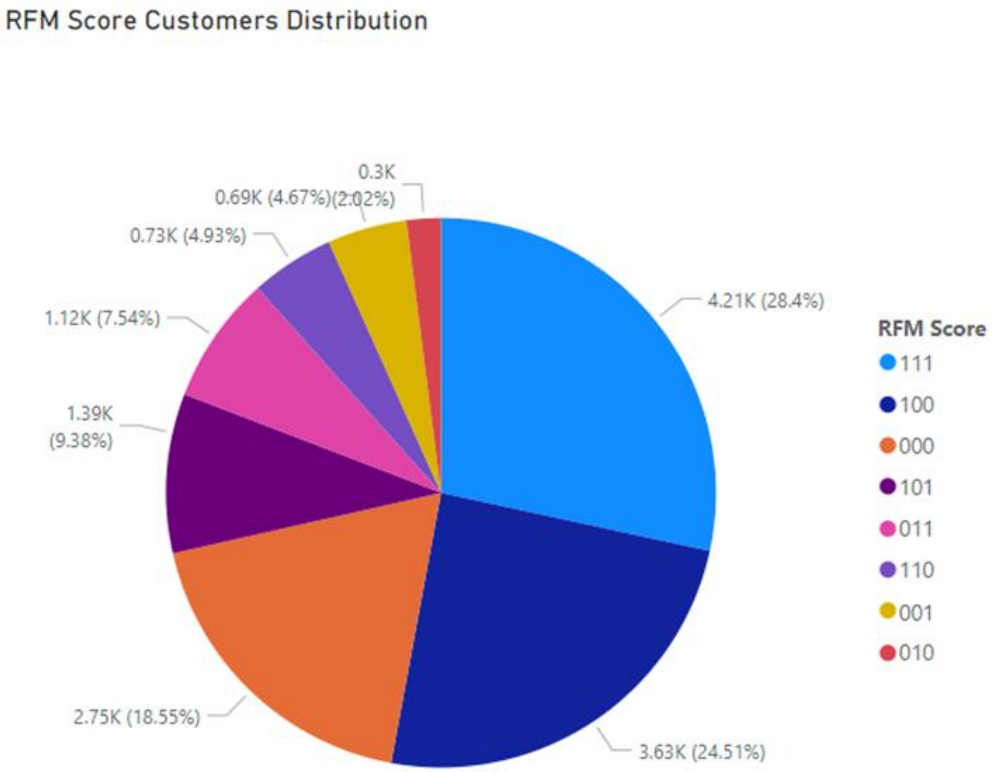
R=Recency
recently transaction time

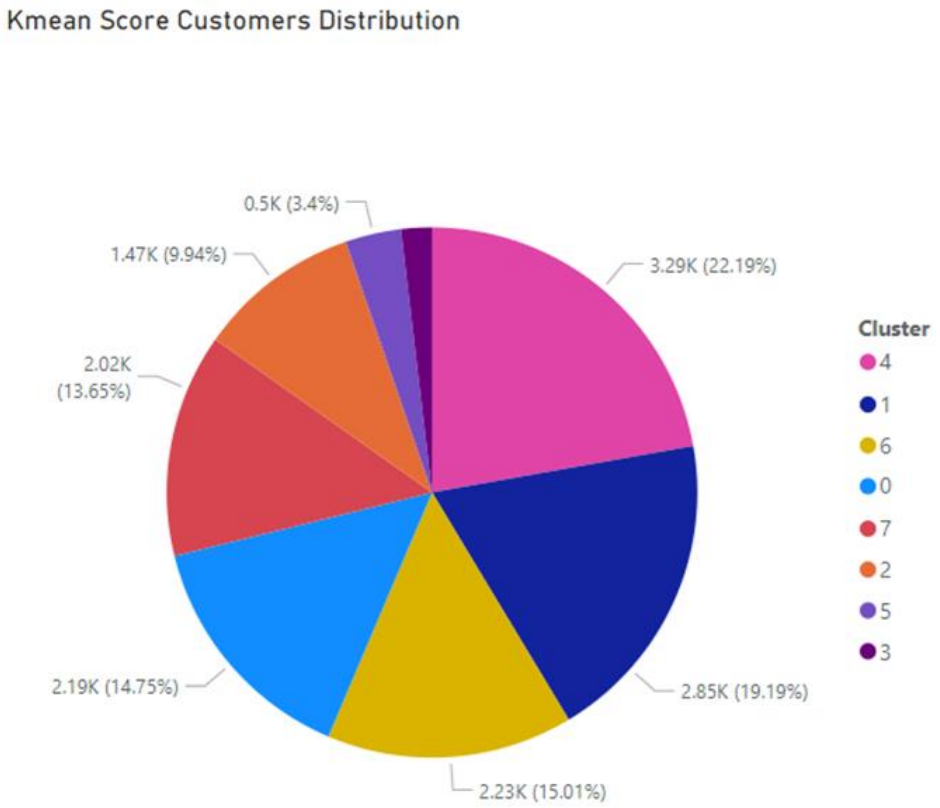
F=Frequency
coupon use frequency

M=Monetary
Sum Amount per person

R	F	M	Customer Type
0	0	0	Lost
1	0	0	New
0	1	0	Regular
0	0	1	Whales
1	1	0	Faithful
1	0	1	Rookie
0	1	1	Can’t Lose
1	1	1	VVVIP

RFM vs Kmean on Customer Cluster





Interest Score

Interest Score

Interest Score=0.7*Success rate+0.3*Response rate

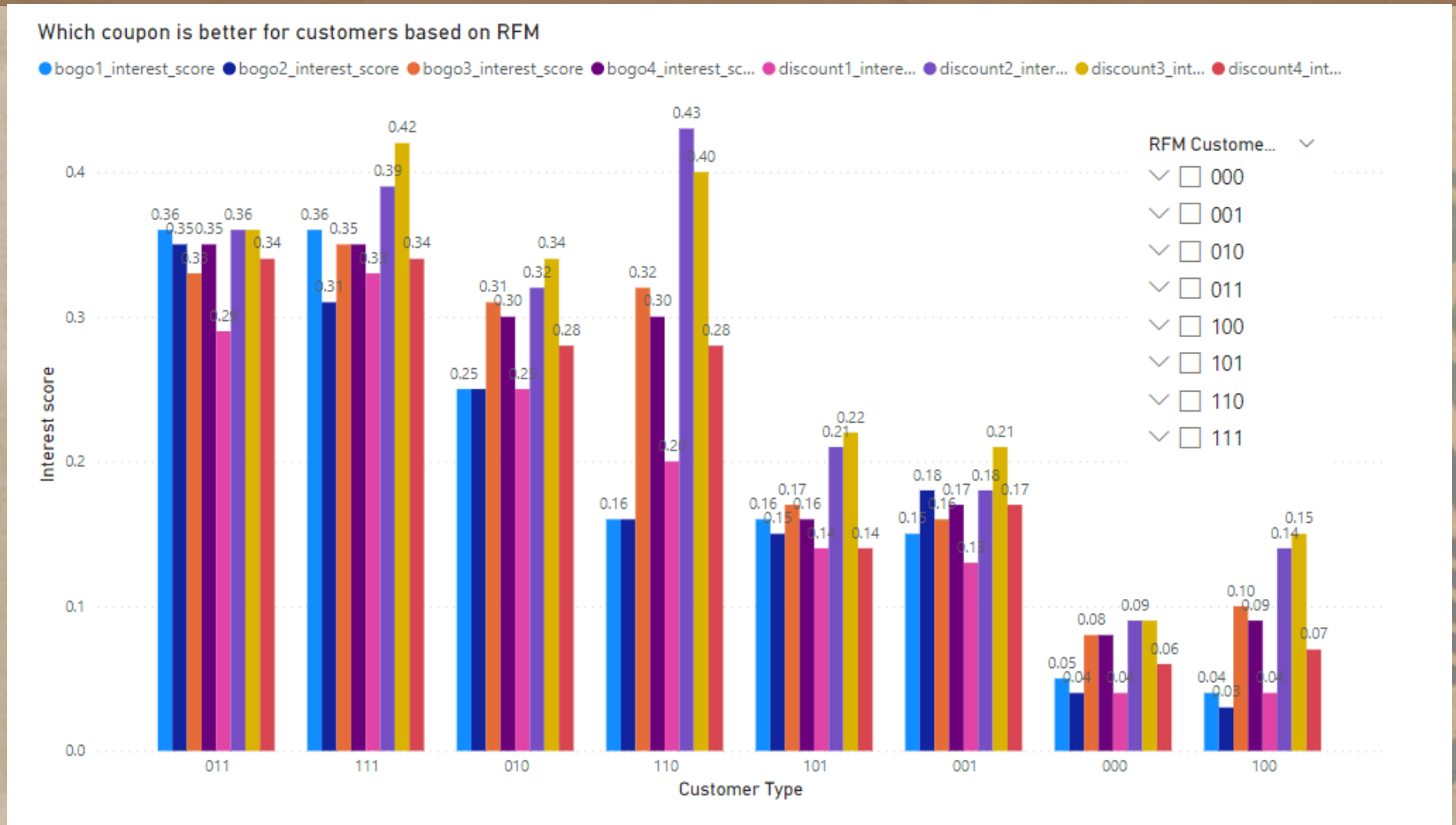
Success rate=

completed coupon / received coupon

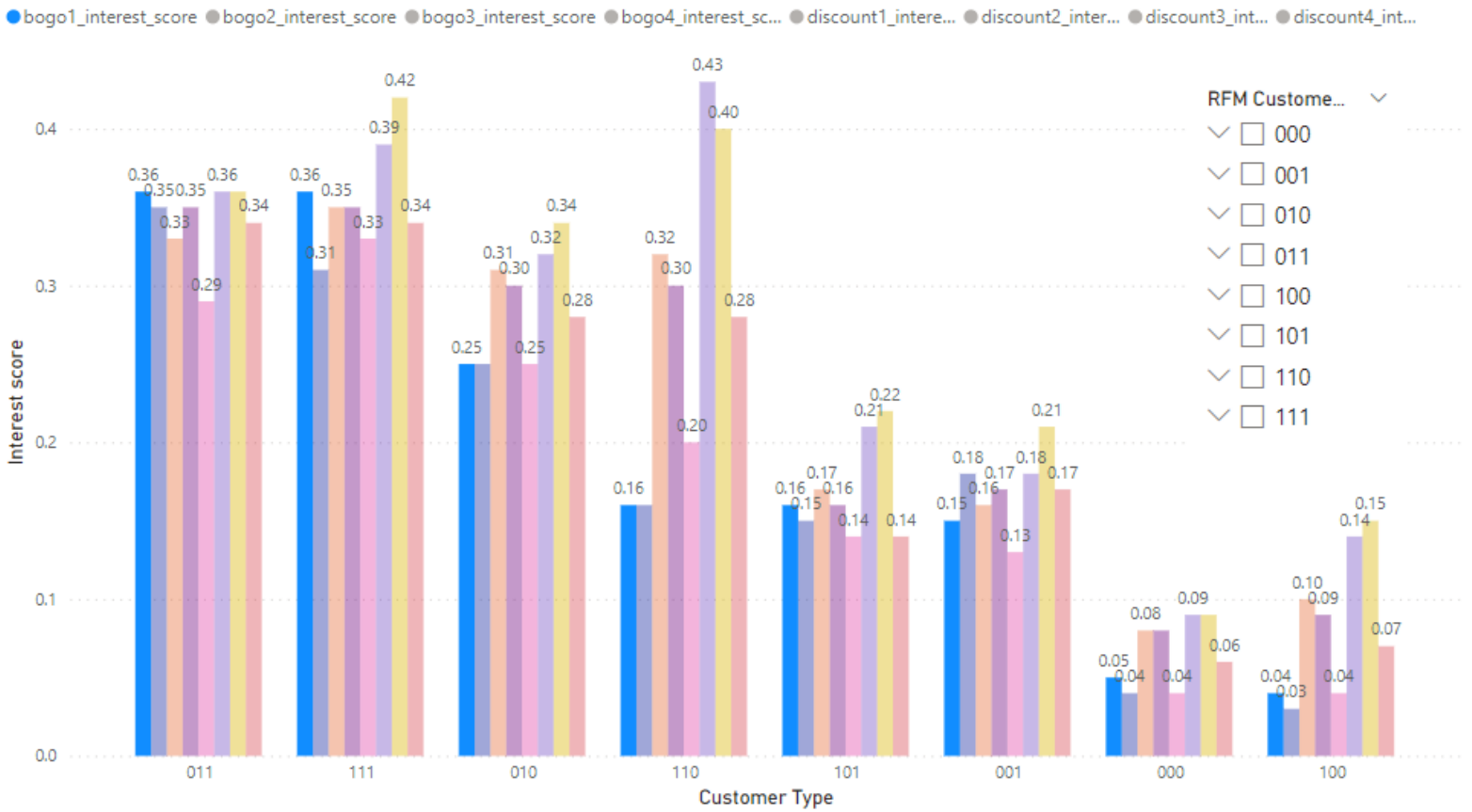
Response rate=

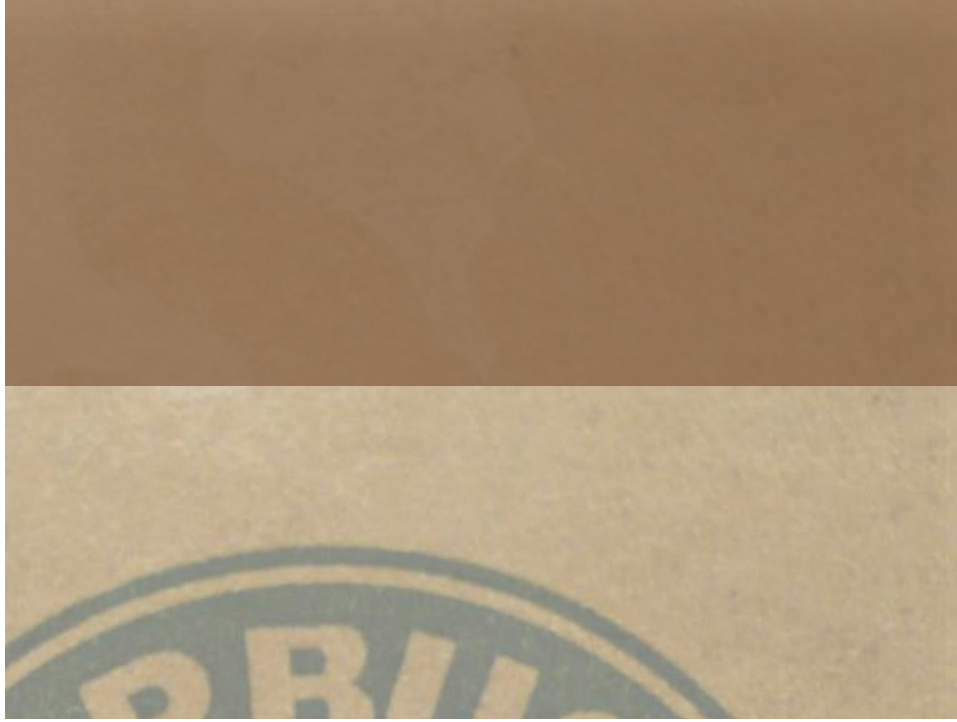
1 - customer response time/coupon duration

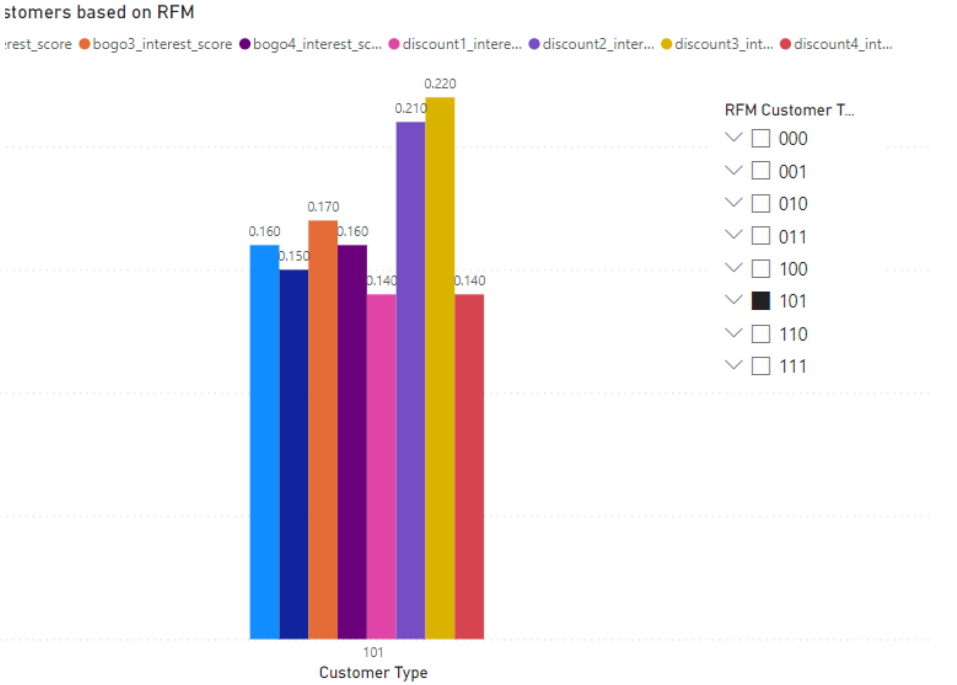
Which coupon is better for you? Based on RFM type



Which coupon is better for customers based on RFM





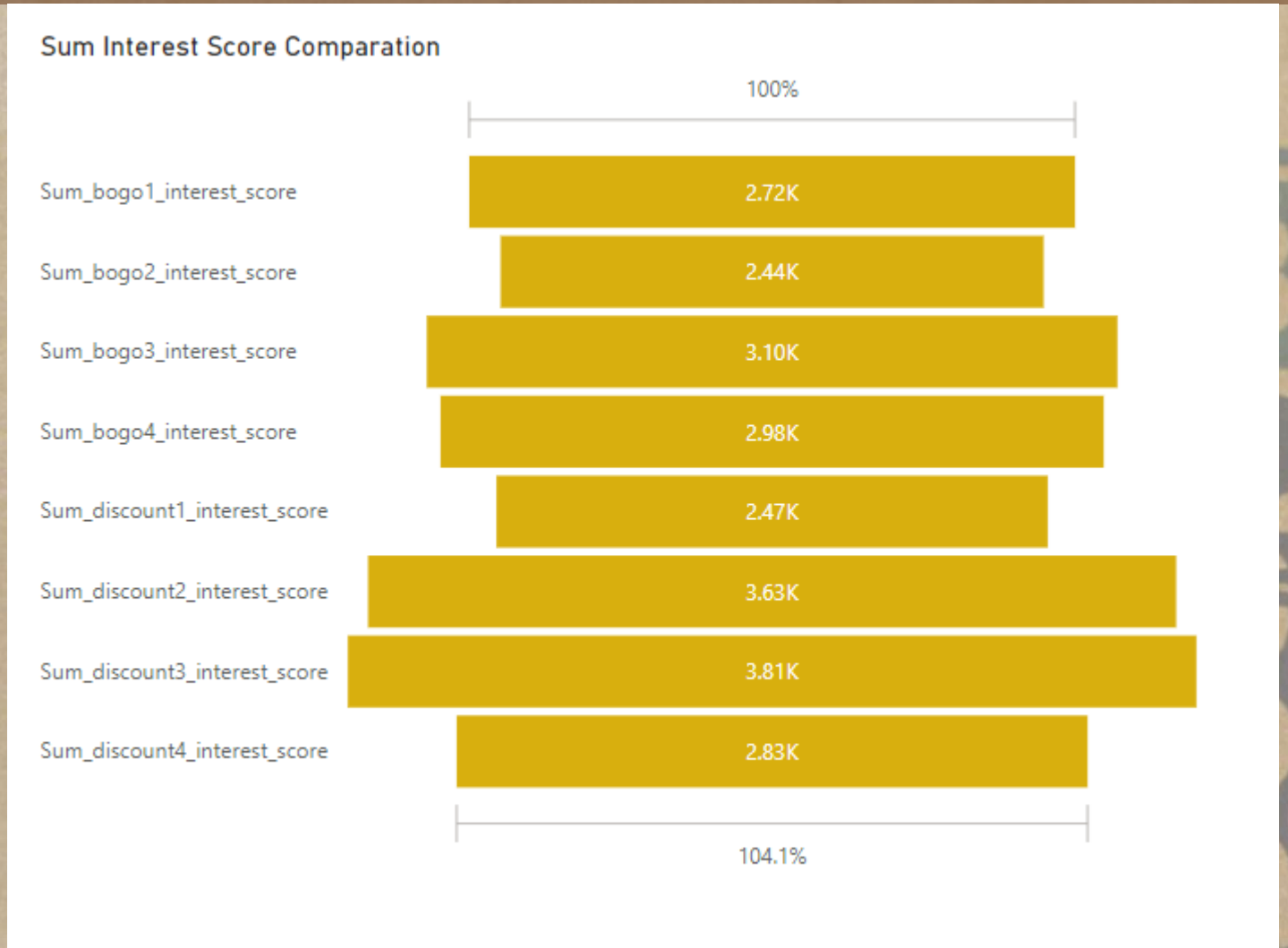


Strategy 1 - Send exact coupon to exact Customer Type

000	Discount 2 Discount 3 Bogo3	100	Discount3 Discount2 Bogo3
001	Discount3 Discount2 bogo2	101	Discount3 Discount2 bogo3
101	Discount3 Discount2 bogo3	110	Discount2 Discount3 Bogo3
011	Bogo1 Bogo4 Discount2	111	Discount3 Discount2 Bogo1

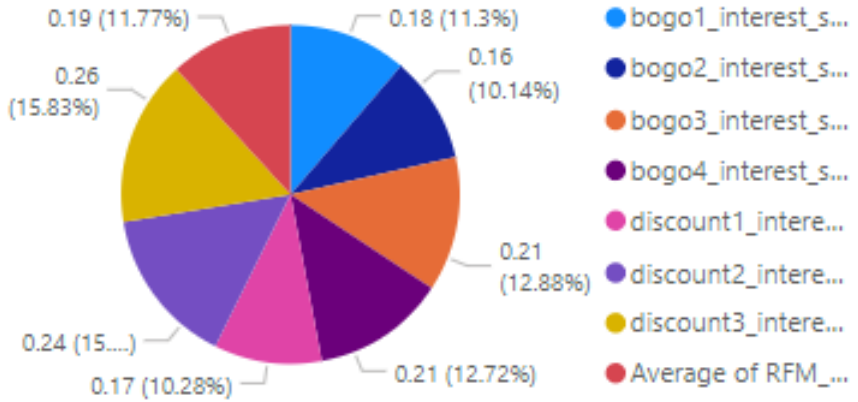
BOGO1	011 111 010	Discount1	111 011 010
BOGO2	011 111 010	Discount2	110 111 011
BOGO3	111 011 110	Discount3	111 110 011
BOGO4	011 111 010(110)	Discount4	011 111 010(110)

Strategy 2 - Send the most popular coupon on general public



Customer & Coupon Relation

Coupon ratio

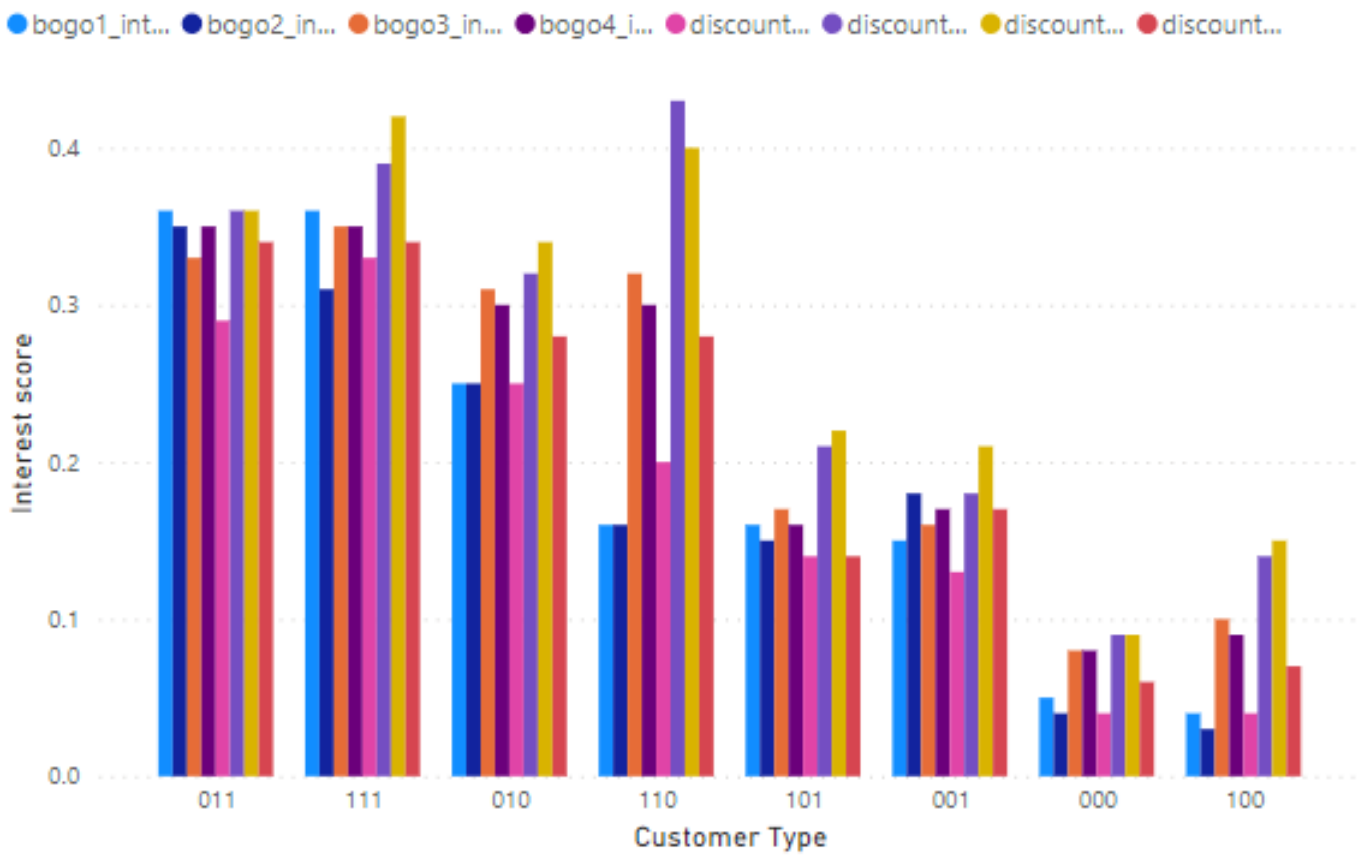


14.83K
Count of person

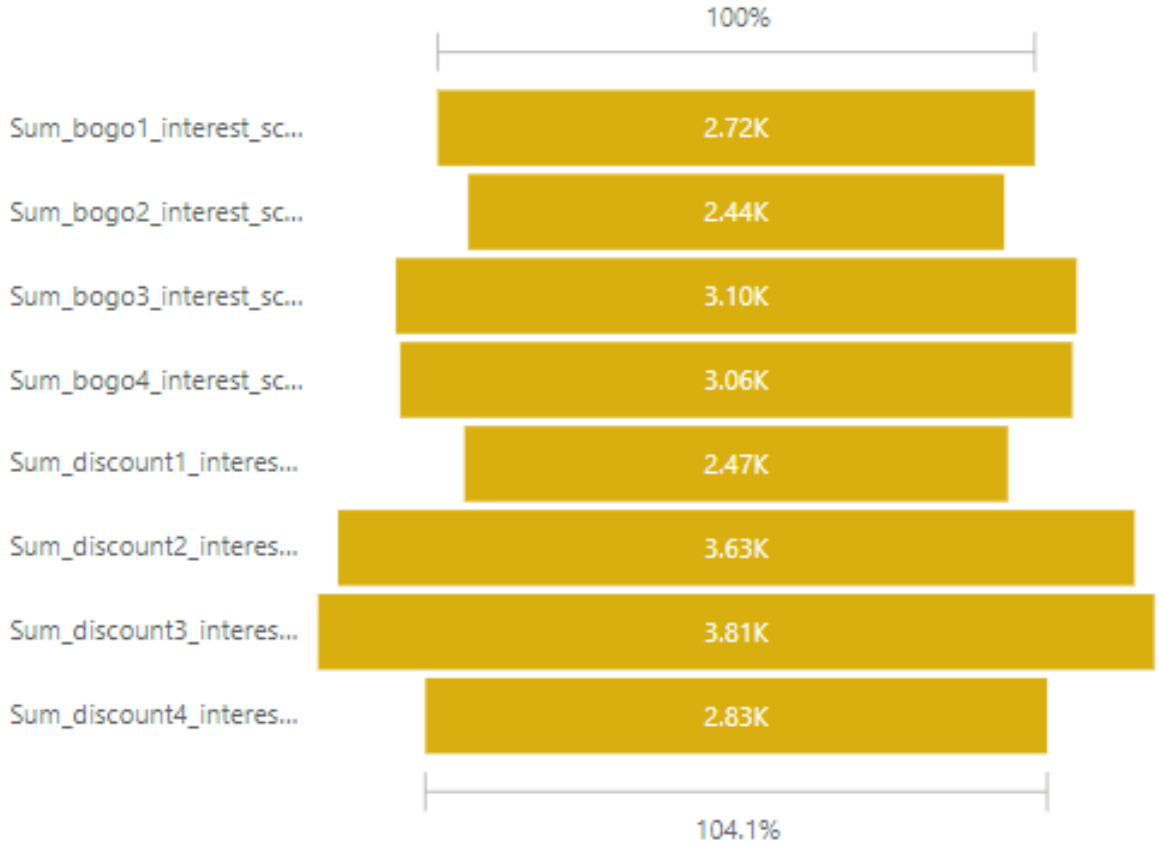
31
Count of r_day

RFM Custome...
- 000
- 001
- 010
- 011
- 100
- 101
- 110
- 111

Which coupon is better for customers based on RFM

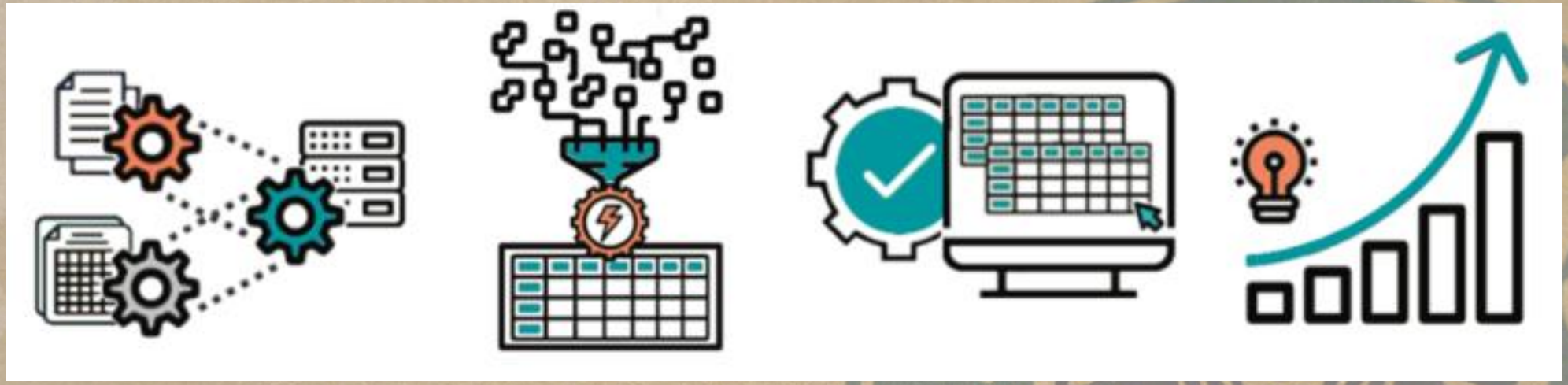


Sum Interest Score Comparation



Machine Learning

- Design - Purpose of the ML model



Prepare the data (Coupon Usages)

Create the model (Train & Test)

Prediction (With different plan)

Better Strategy

Machine Learning

INPUT

User Features	User Behavior	Coupon Features	Marketing Features
age gender income	Recency Frequency Monetary bogo1_completed_timediff bogo2_completed_timediff bogo3_completed_timediff bogo4_completed_timediff discount1_completed_timediff discount2_completed_timediff discount3_completed_timediff discount4_completed_timediff	**bogo1_difficulty** **bogo2_difficulty** **bogo3_difficulty** **bogo4_difficulty** **discount1_difficulty** **discount2_difficulty** **discount3_difficulty** **discount4_difficulty** **bogo1_duration** **bogo2_duration** **bogo3_duration** **bogo4_duration** **discount1_duration** **discount2_duration** **discount3_duration** **discount4_duration**	**bogo1_received** **bogo2_received** **bogo3_received** **bogo4_received** **discount1_received** **discount2_received** **discount3_received** **discount4_received**

OUTPUT

Result
bogo1_interest_score bogo2_interest_score bogo3_interest_score bogo4_interest_score discount1_interest_score discount2_interest_score discount3_interest_score discount4_interest_score

Machine Learning

Random Forest

A

Regression

INTEREST SCORE

B

Classification

Most interested coupon
Second interested coupon

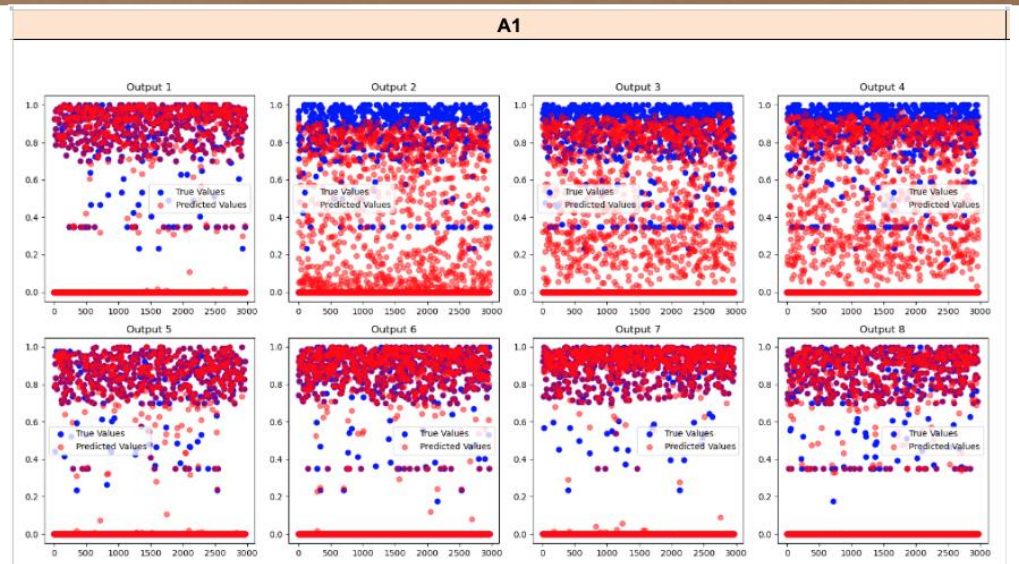
Performance Evaluation

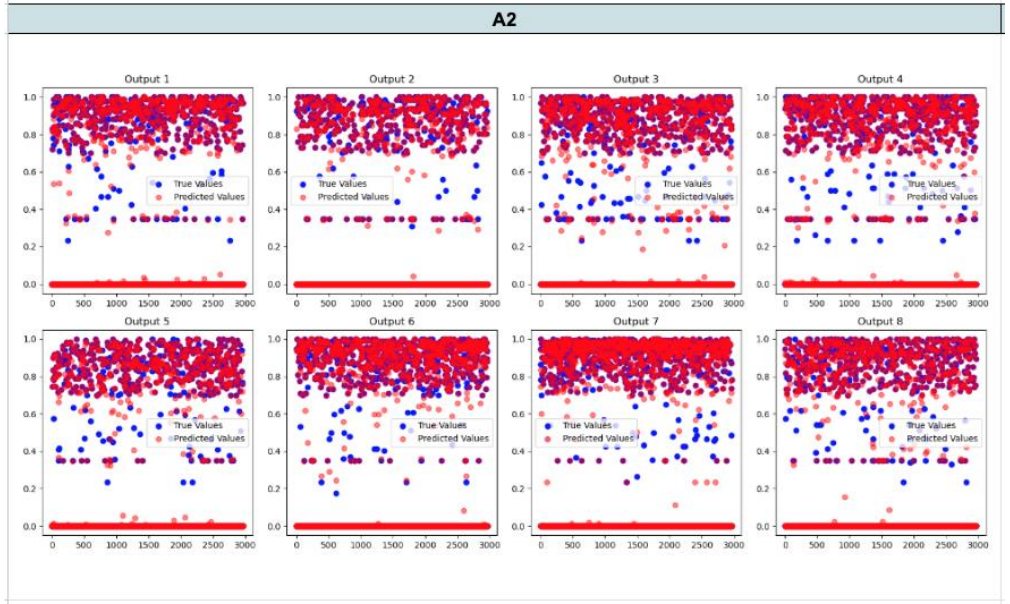
Model A - Regression

MSE / RMSE / MAE / R^2

	A1	A2
Input (different)	RFM	Kmean
Output	bogo1_interest_score' ,'bogo2_interest_score' ,'bogo3_interest_score' ,'bogo4_interest_score' ,'discount1_interest_score' ,'discount2_interest_score' ,'discount3_interest_score' ,'discount4_interest_score'	bogo1_interest_score' ,'bogo2_interest_score' ,'bogo3_interest_score' ,'bogo4_interest_score' ,'discount1_interest_score' ,'discount2_interest_score' ,'discount3_interest_score' ,'discount4_interest_score'
MSE	0.015	0.001
RMSE	0.090	0.030
MAE	0.036	0.005
Training R^2	0.984	0.999
Testing R^2	0.884	0.993
Average OOB Score	0.884	0.994

	Green means: LOWER value = BETTER performance
	Red means: HIGHER value = BETTER performance





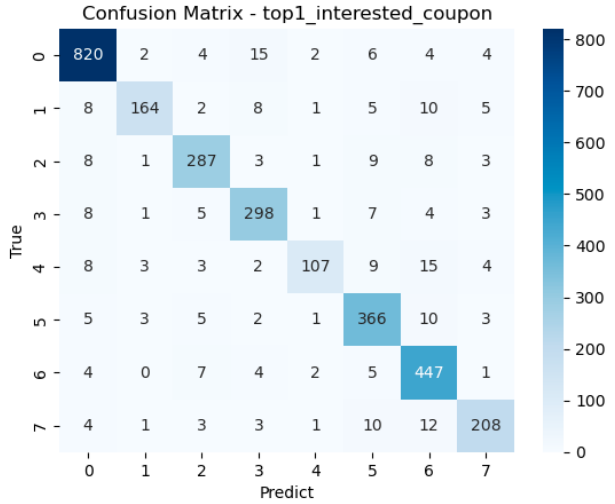
Performance Evaluation

Model B - Classification

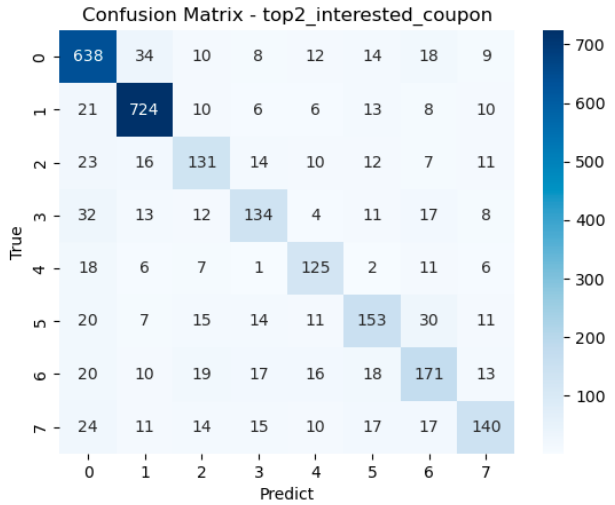
Hamming Loss / Jaccard Score / Accuracy
Confusion Matrix (Heat map) / Classification Report

	AA1	AA2
Input (different)	RFM	Kmean
Output	top1_interested_coupon' ,'top2_interested_coupon'	top1_interested_coupon' ,'top2_interested_coupon'
1st Hamming Loss	0.090	0.091
2nd Hamming Loss	0.253	0.253
avg Hamming Loss	***0.172***	***0.172***
1st Jaccard Score	0.834	0.834
2nd Jaccard Score	0.597	0.596
avg Jaccard Score	***0.715***	***0.715***
1st Accuracy	90.96%	90.93%
2nd Accuracy	74.74%	74.70%
avg Accuracy	***82.85%***	***82.82%***

	Green means: LOWER value = BETTER performance
	Red means: HIGHER value = BETTER performance



TP: 164, TN: 820, FP: 2, FN: 8



TP: 724, TN: 638, FP: 34, FN: 21

Classification Report - top1:

	precision	recall	f1-score	support
bogo1	0.95	0.96	0.95	857
bogo2	0.94	0.81	0.87	203
bogo3	0.91	0.90	0.90	320
bogo4	0.89	0.91	0.90	327
discount1	0.92	0.71	0.80	151
discount2	0.88	0.93	0.90	395
discount3	0.88	0.95	0.91	470
discount4	0.90	0.86	0.88	242
accuracy			0.91	2965
macro avg	0.91	0.88	0.89	2965
weighted avg	0.91	0.91	0.91	2965

Classification Report - top2:

	precision	recall	f1-score	support
bogo1	0.80	0.86	0.83	743
bogo2	0.88	0.91	0.89	798
bogo3	0.60	0.58	0.59	224
bogo4	0.64	0.58	0.61	231
discount1	0.64	0.71	0.68	176
discount2	0.64	0.59	0.61	261
discount3	0.61	0.60	0.61	284
discount4	0.67	0.56	0.61	248
accuracy			0.75	2965
macro avg	0.69	0.67	0.68	2965
weighted avg	0.74	0.75	0.74	2965

Explore the Neural Network

Python Sklearn.neural_network module

Two library:

1. MLPClassifier:

   Multi-layer Perception Classifier

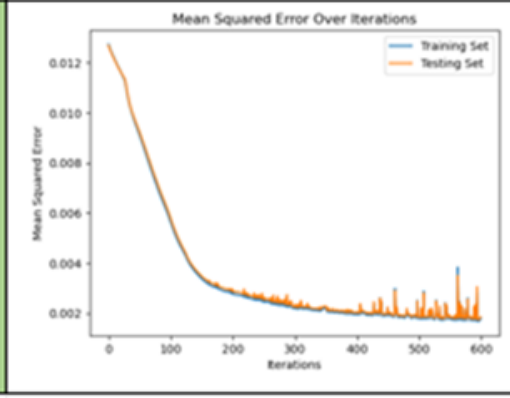
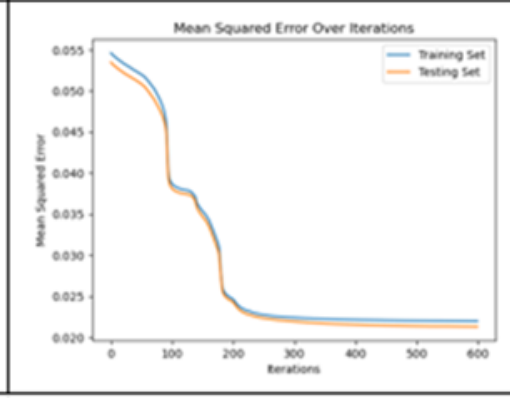
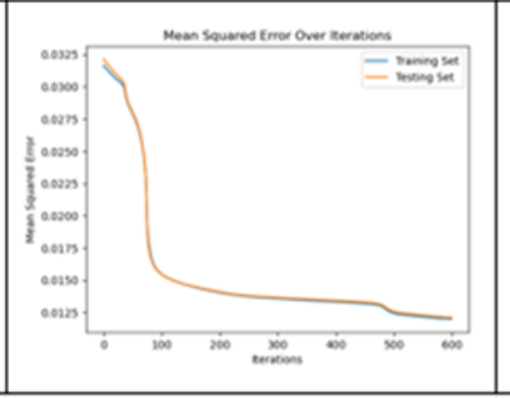
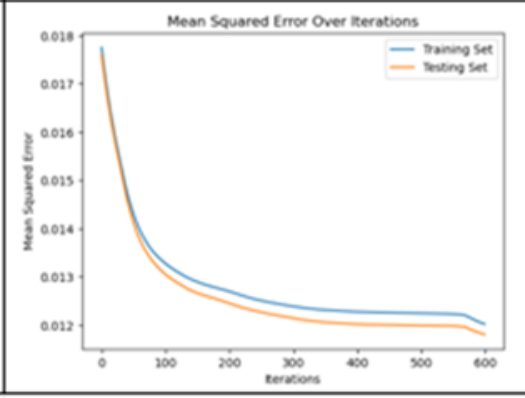
1. MLPRegressor:

   Multi-layer Perceptiopn Regressor

Results and Findings

1.Model performs well in both train data and test data, it means the model generalize effectively to unfamiliar data

2.Time consuming

hidden_layer_sizes	(30,30,30)	(10,10,10)	(10,10)	(10)
mse_train_pred	0.0018	0.0220	0.0120	0.0120
mse_test_pred	0.0018	0.0220	0.0121	0.0118
mae_train_pred	0.0169	0.0768	0.0558	0.0558
mae_test_pred	0.0174	0.0756	0.0558	0.0552
r2_train_pred	0.9869	0.8228	0.9130	0.9114
r2_test_pred	0.9862	0.8225	0.9104	0.9114
Mean Squared Error Over Iteration				

Conclusion

- After evaluation, all 3 models have shown outstanding performance
- meeting the expected testing results for the project
- able to help the marketing department to improve promotion strategy.

Findings：

- The interest score of each coupon is a bit more related to **Kmean**.
- For the coupon portfolio, the interest score is more related to **Channels** rather than difficulty and duration.
- With the increase of ranks, the accuracy went down.

Future Improvements and Outlook

Q1: User who never receive one type of coupon. (Manually give 0)

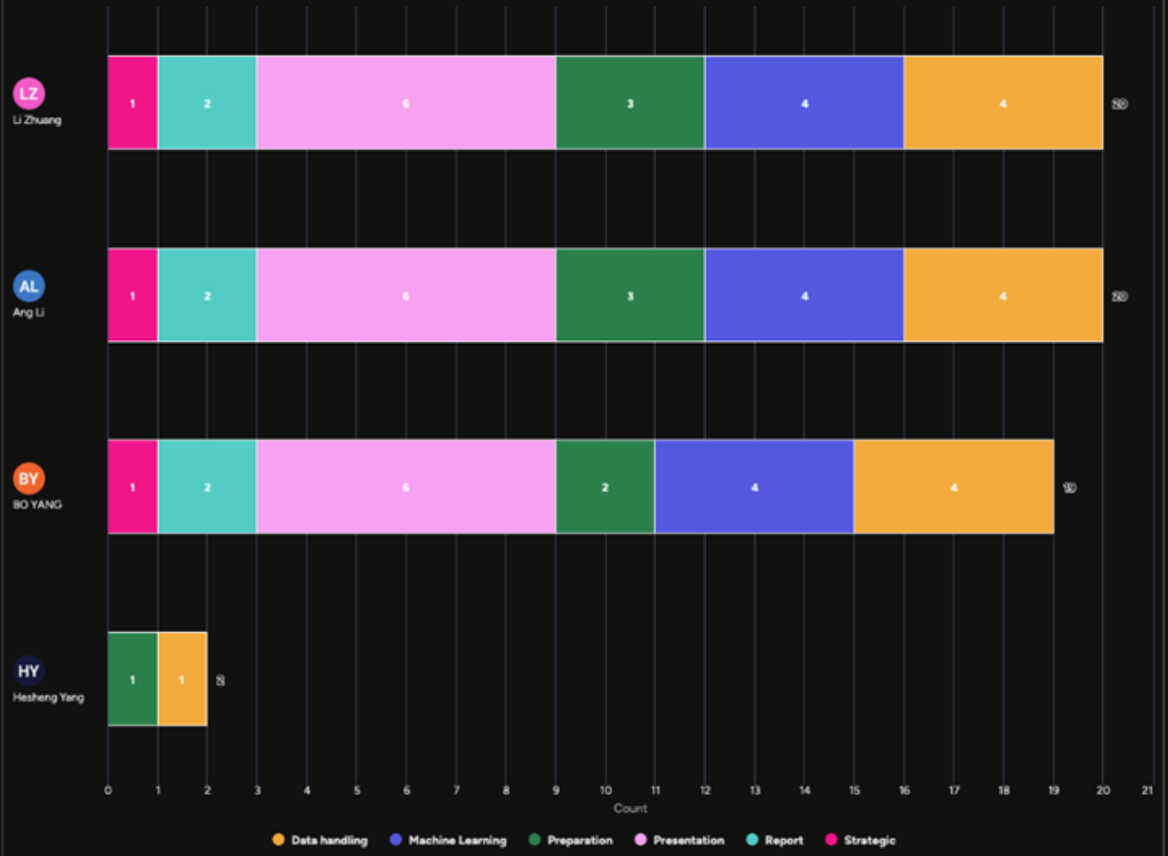
Solution: It should be adjusted to a more reasonable value through additional calculations.

Q2: Currently, for the promotion response_time, we only calculate the first response time after a user receives the coupon.

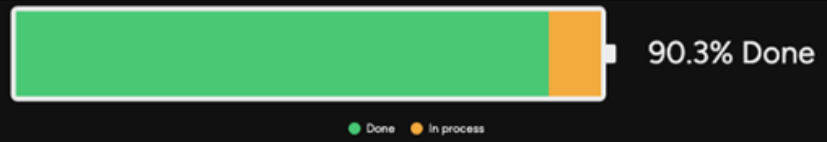
Solution: We can enhance the calculation by including every instance of the user's reaction time, computing the average and give to the response_time.

Task Rate

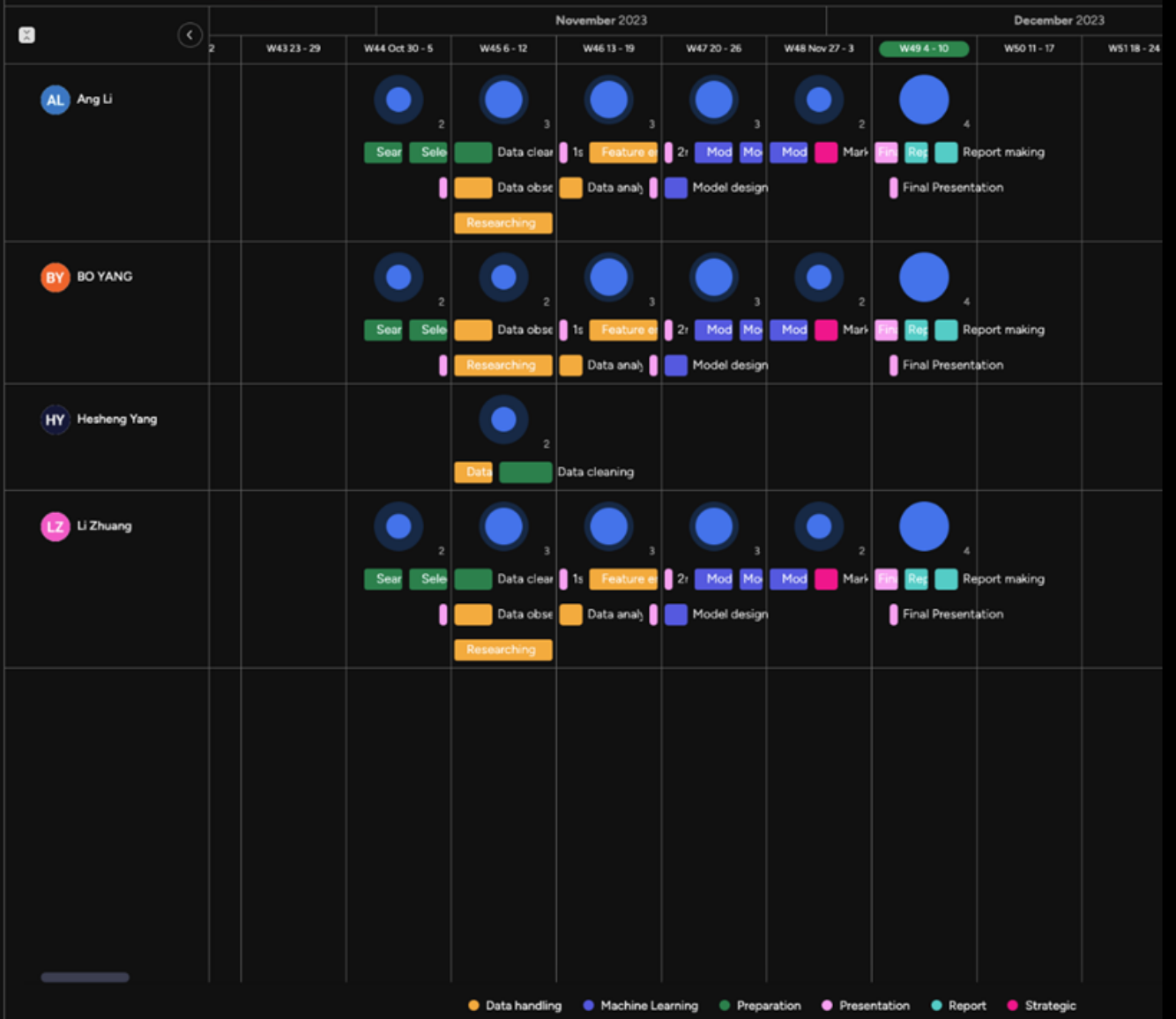
Task Contribution (Task participated amount)



Overall project completion



Weekly Overload detector



Week 1 (Oct 30 - Nov 5)

Epic	Category	Schedule	Status	Owner	Participants
Plan Meeting	Meeting	Nov 5	Done	BY	BY LZ AL HY
Searching for Dataset	Preparation	Oct 31 - Nov 2	Done	LZ	BY LZ AL
Selecting Dataset	Preparation	Nov 3 - 5	Done	LZ	BY LZ AL
Review Meeting	Meeting	Nov 5	Done	BY	BY LZ AL HY
1st Powerpoint	Presentation	Nov 5	Done	BY	BY LZ AL
+ Add Epic					
		Oct 31 - Nov 5		BY LZ	

Week 2 (Nov 6 - Nov 12)

Epic	Category	Schedule	Status	Owner	Participants
Plan Meeting	Meeting	Nov 6	Done	BY	BY LZ AL HY
Data cleaning	Preparation	Nov 9 - 12	Done	HY	HY
Data cleaning (re-assig...	Preparation	Nov 6 - 8	Done	AL	LZ AL
Data observation	Data handling	Nov 6 - 8	Done	AL	BY LZ AL HY
Researching	Data handling	Nov 6 - 12	Done	LZ	BY LZ AL
Review Meeting	Meeting	Nov 12	Done	BY	BY LZ AL HY
+ Add Epic					
		Nov 6 - 12		BY +3	

week 3 (Nov 13 - Nov 19)

Epic	Category	Schedule	Status	Owner	Participants
Plan Meeting	Meeting	Nov 13	Done	BY	BY LZ AL HY
1st Presentation	Presentation	Nov 13	Done	BY	BY LZ AL
Data analysis	Data handling	Nov 13 - 14	Done	LZ	BY LZ AL
Feature engineering 2	Data handling	Nov 15 - 19	Done	AL LZ	BY LZ AL

Subitem	Owner	Status
RFM	AL	Done
Kmean	LZ	Done
+ Add subitem		

Epic	Category	Schedule	Status	Owner	Participants
Review Meeting	Meeting	Nov 19	Done	BY	BY LZ AL HY
2nd Powerpoint	Presentation	Nov 19	Done	BY	BY LZ AL
+ Add Epic					
		Nov 13 - 19		BY +2	

week 4 (Nov 20 - Nov 26)

Epic	Category	Schedule	Status	Owner	Participants
Plan Meeting	Meeting	Nov 20	Done	BY	BY LZ AL HY
2nd Presentation	Presentation	Nov 20	Done	BY	BY LZ AL
Model design	Machine Learning	Nov 20 - 21	Done	LZ	BY LZ AL
Model building 2	Machine Learning	Nov 22 - 24	Done	AL	BY LZ AL
Model evaluating	Machine Learning	Nov 25 - 26	Done	BY	BY LZ AL
Review Meeting	Meeting	Nov 26	Done	BY	BY LZ HY AL
+ Add Epic					
		Nov 20 - 26		BY +2	

week 5 (Nov 27 - Dec 3)

Epic	Category	Schedule	Status	Owner	Participants
Plan Meeting	Meeting	Nov 27	Done	BY	BY LZ AL HY
Model study	Machine Learning	Nov 27 - 29	Done	AL	BY LZ AL
Marketing strategy anal...	Strategic	Nov 30 - Dec 1	Done	LZ	BY LZ AL
Review Meeting	Meeting	Dec 3	Done	BY	BY LZ AL HY
+ Add Epic					
		Nov 27 - Dec 3		BY +2	

week 6 (Dec 4 - Dec 5)

Epic	Category	Schedule	Status	Owner	Participants
Final Powerpoint	Presentation	Dec 4 - 5	Done	LZ	BY LZ AL
Final Presentation	Presentation	Dec 5	In process	AL	BY LZ AL
Report design	Report	Dec 6 - 7	In process	AL	BY LZ AL
Report making	Report	Dec 8 - 9	In process	BY	BY LZ AL
+ Add Epic					
		Dec 4 - 9		LZ +2	

Q&A

THANK YOU