MARKETING CAMPAIGN ANALYSIS

Final Project

Rakamin Data Science Bootcamp
Batch 28

BY THE TEN GENERALIST

THE TEN GENERALIST

Data Scientist Team at
Ten Eleven, Inc (Retail Company)

Dhiaz Raflianza
(Mentor)

Aminudin

M. Nafiul Ahkam

M. Afif Hibban

M. Malik

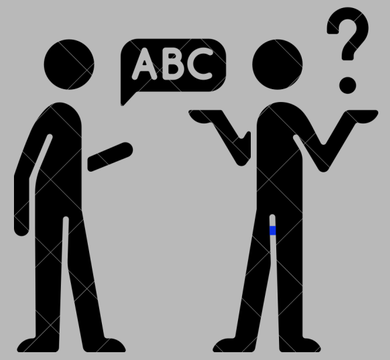
Suci Share Putri

Ramadhani Yovita H

Suci Rahmadiani

OUTLINE

01



Business Understanding

02

Exploratory Data Analysis

03

Data Preprocessing

04

Modelling & Evaluation

05

Business Recommendation

CHAPTER 1

BUSINESS UNDERSTANDING

Apa permasalahan yang sedang dialami perusahaan Ten Eleven?

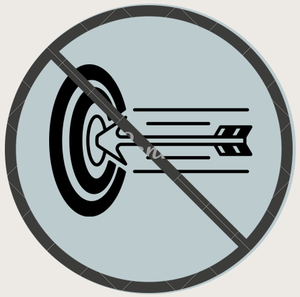
1

PROBLEM STATEMENT

2

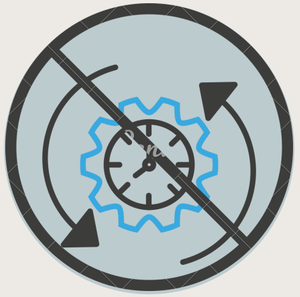
GOALS, OBJECTIVE, AND METRICS

PROBLEM STATEMENT



Lack of Accuracy

Marketing campaign yang dilaksanakan hanya mendapat respon sebesar 14,91%

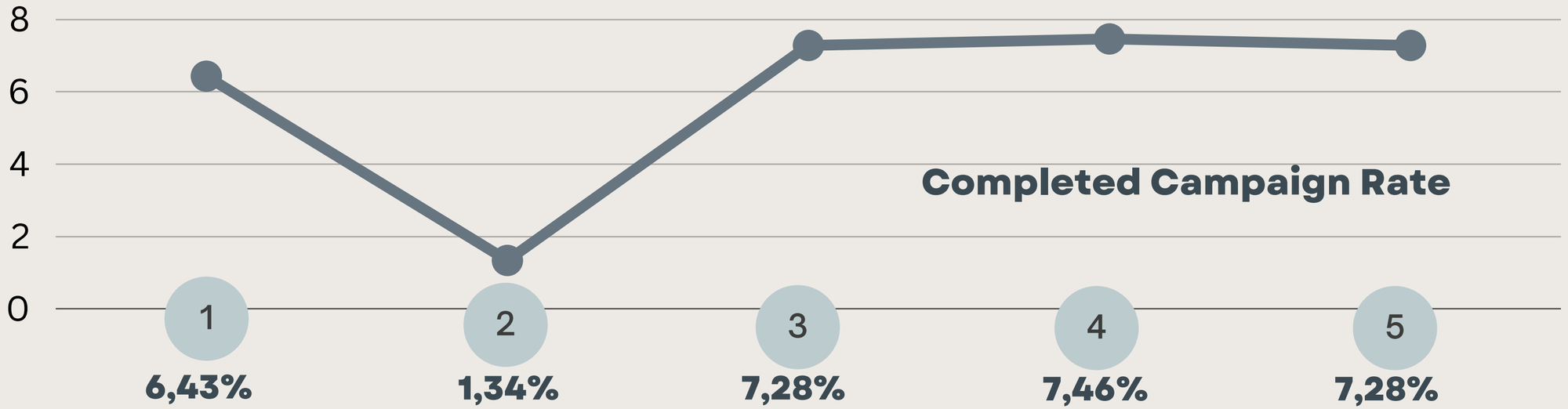


Unefficient Budget

Budget marketing belum digunakan secara efisien.

Loss Profit

Profit yang didapatkan dari campaign belum optimal.



GOALS, OBJECTIVE, AND METRICS

GOALS

Mengoptimalkan profit dengan membuat campaign yang lebih tepat sasaran

OBJECTIVE

Membuat model machine learning untuk memprediksi customer yang kemungkinan besar akan menerima promosi tertentu sehingga menjadikan campaign lebih tepat sasaran dan mendapatkan profit yang optimal.

BUSINESS METRICS

PROFIT PERCENTAGE

$$\frac{(\text{Revenue} - \text{cost})}{\text{cost}} \times 100\%$$

CHAPTER 2

EXPLORATORY DATA ANALYSIS

Bagaimana persebaran data marketing campaign pada perusahaan Ten Eleven dan insight apa yang bisa disimpulkan?

1. Descriptive Statistic
2. Univariate Analysis
3. Multivariate Analysis
4. Business Insight

DESCRIPTIVE STATISTIC

Data Shape
2240 Baris
29 Kolom

Duplicated Data
Tidak Ada

Cek Kesesuaian
Judul Kolom dan isinya sudah sesuai

Missing Data
24 row di kolom income

Konfigurasi Kolom

FEATURES

25 Numerik

- ID
- Year_Birth
- Income
- Complain
- Kidhome
- Teenhome

- MntWines
- MntFruits
- MntMeatProducts
- MntFishProducts
- MntSweetProducts
- MntGoldProducts

- NumDealsPurchases
- NumWebPurchases
- NumCatalogPurchases
- NumStorePurchases
- NumWebVisitMonth
- Z_CostContact
- Z_Revenue

- Recency
- AcceptedCmp1
- AcceptedCmp2
- AcceptedCmp3
- AcceptedCmp4
- AcceptedCmp5

3 Kategorikal

- Education
- Marital Status
- Dt_Customer

LABEL

1 Numerik

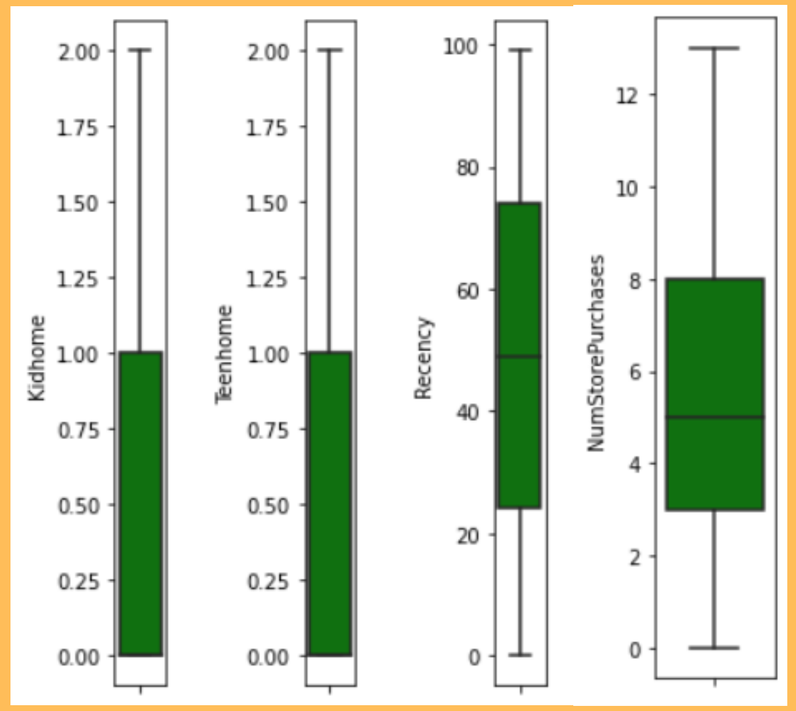
RESPONSE

Reaksi terhadap campaign terakhir

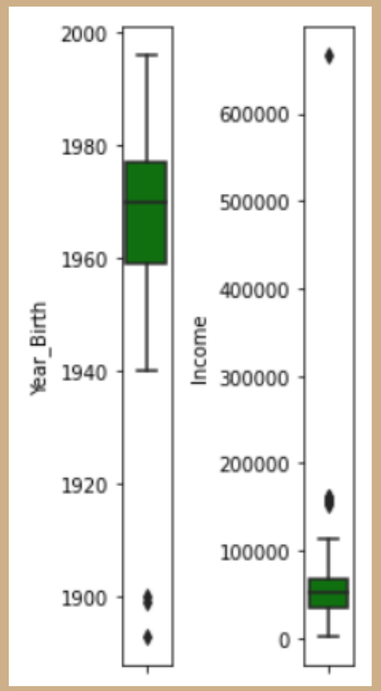
0: ignore
1: accept

UNIVARIATE ANALYSIS

Distribusi Kolom Numerik



4 Kolom tanpa Outlier



2 Kolom dengan Outlier Eksterim

19 Kolom Numerik Lainnya memiliki Outlier namun tidak terlalu ekstrem

UNIVARIATE ANALYSIS

Distribusi Kolom Categorical

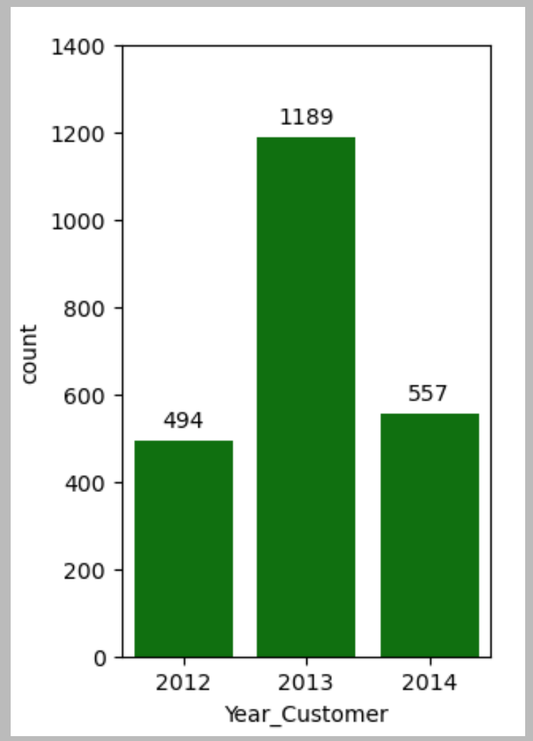
Penyederhanaan Kolom
Education

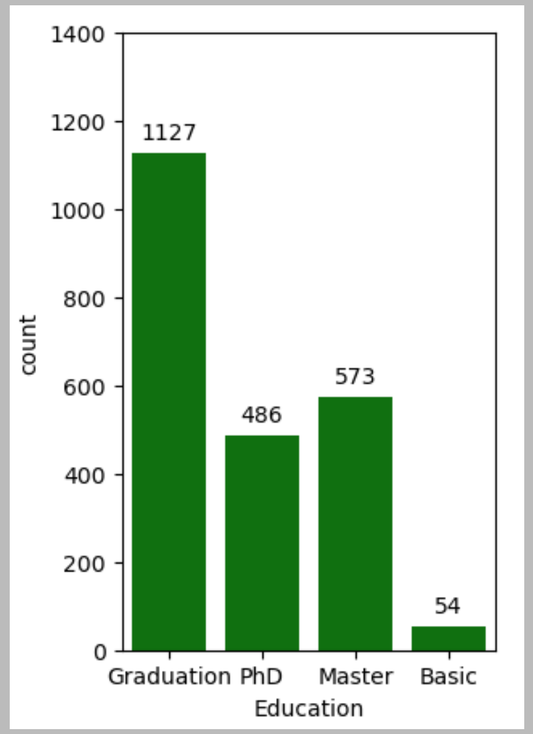
Graduation	→	Graduation
Master	→	Master
2n cycle	→	Master
Basic	→	Basic
PhD	→	PhD

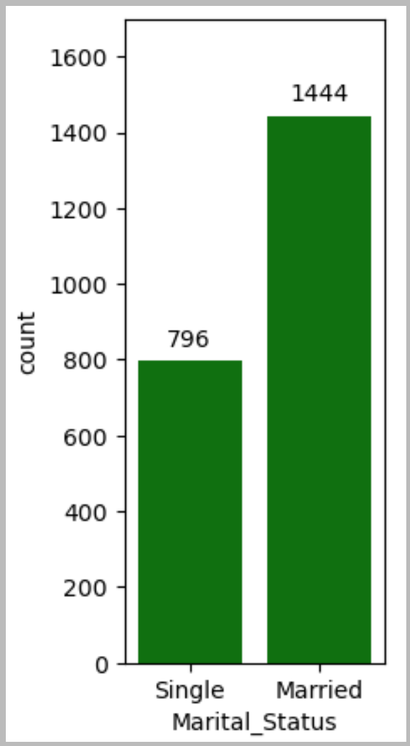
Penyederhanaan Kolom
Marital_Status

Single, Alone, Absurd, Divorced, YOLO, Widow, → **Single**

Together, Married → **Married**



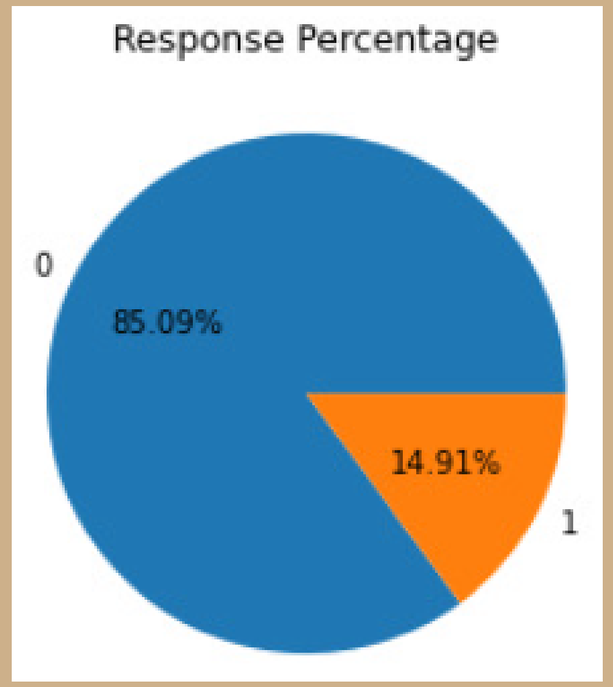




Mayoritas customer berlangganan mulai tahun 2013, merupakan lulusan S1, dan sudah menikah

UNIVARIATE ANALYSIS

Persentase Response

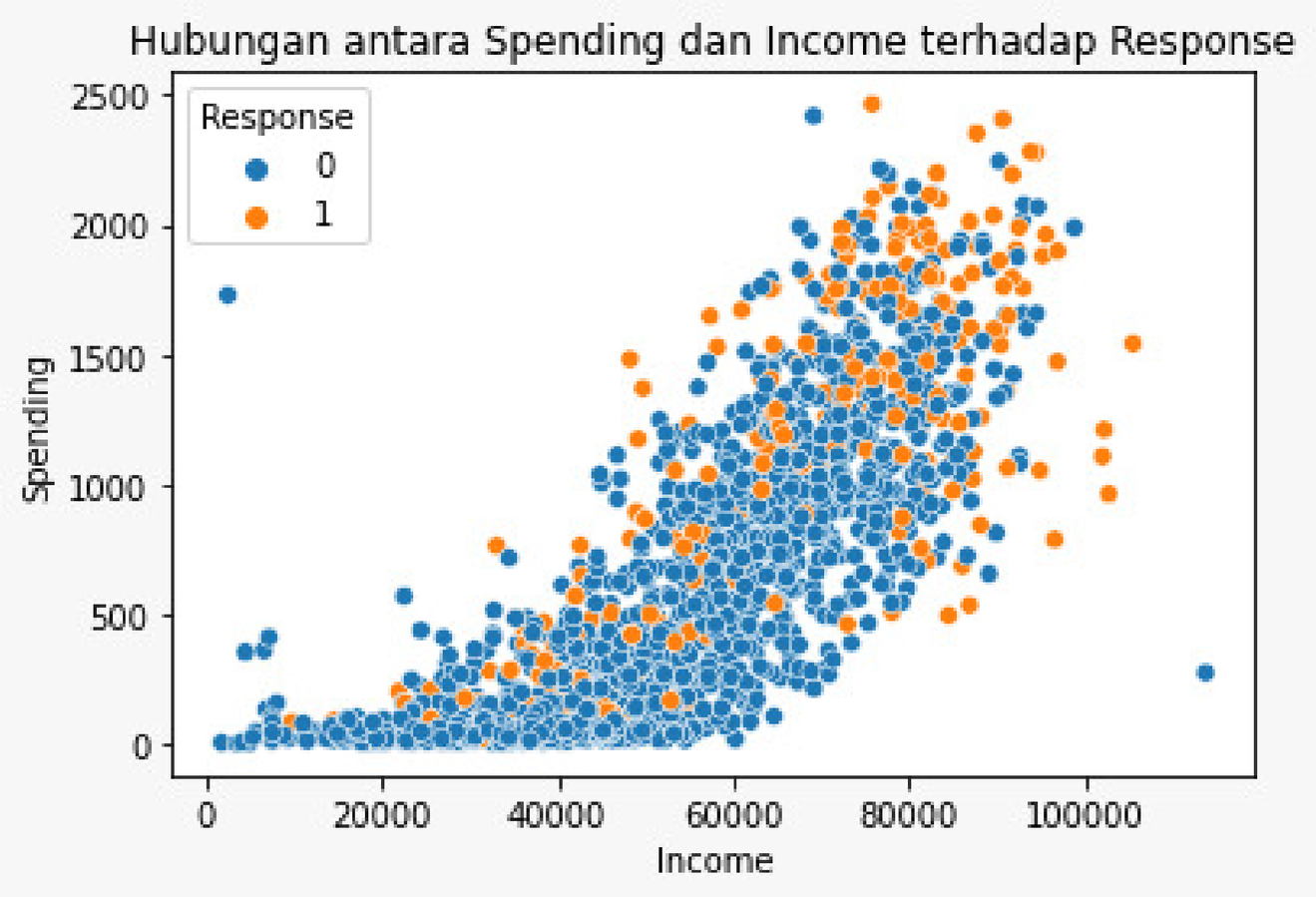


Pelanggan yang meresponse campaign hanya 14,9%

MODERATE IMBALANCE

MULTIVARIATE ANALYSIS

Response - Income & Spending

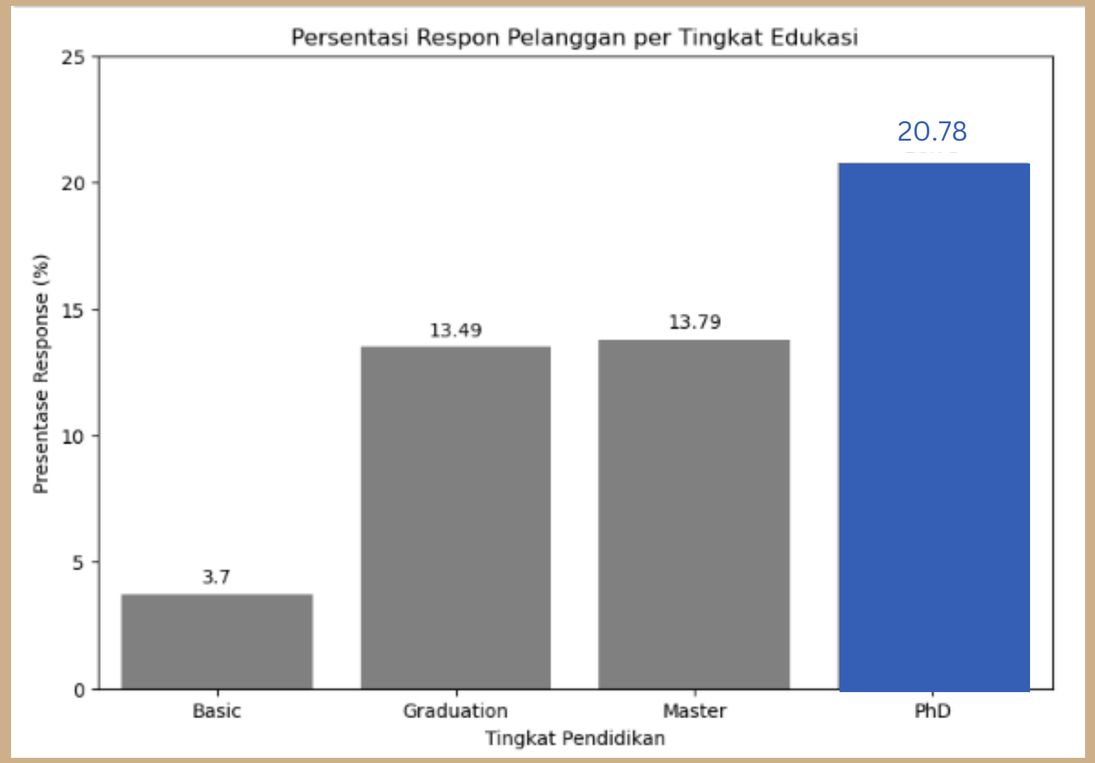


Nilai Spending =
MntWines + MntFruits + MntMeatProducts + MntFishProducts + MntSweetProducts + MntGoldProducts

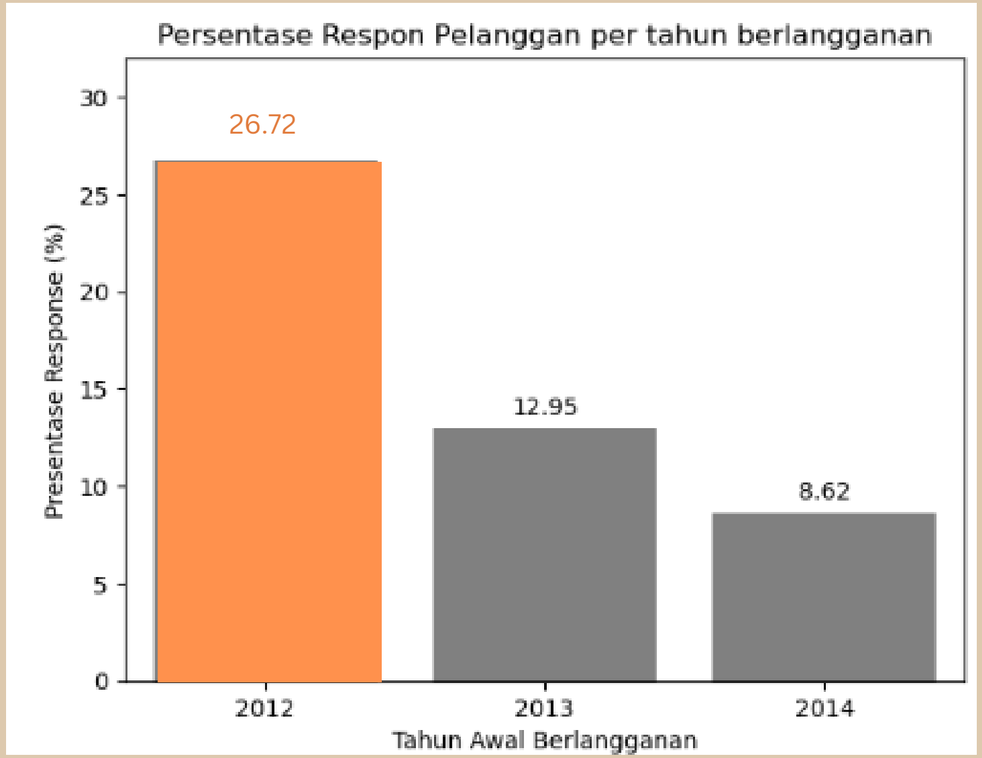
Semakin besar income dan spending, semakin besar kemungkinan customer merespons campaign

MULTIVARIATE ANALYSIS

Response - Tingkat Pendidikan & Response - Tahun Awal Berbelanja



Semakin tinggi level pendidikan, semakin besar potensi untuk merespons



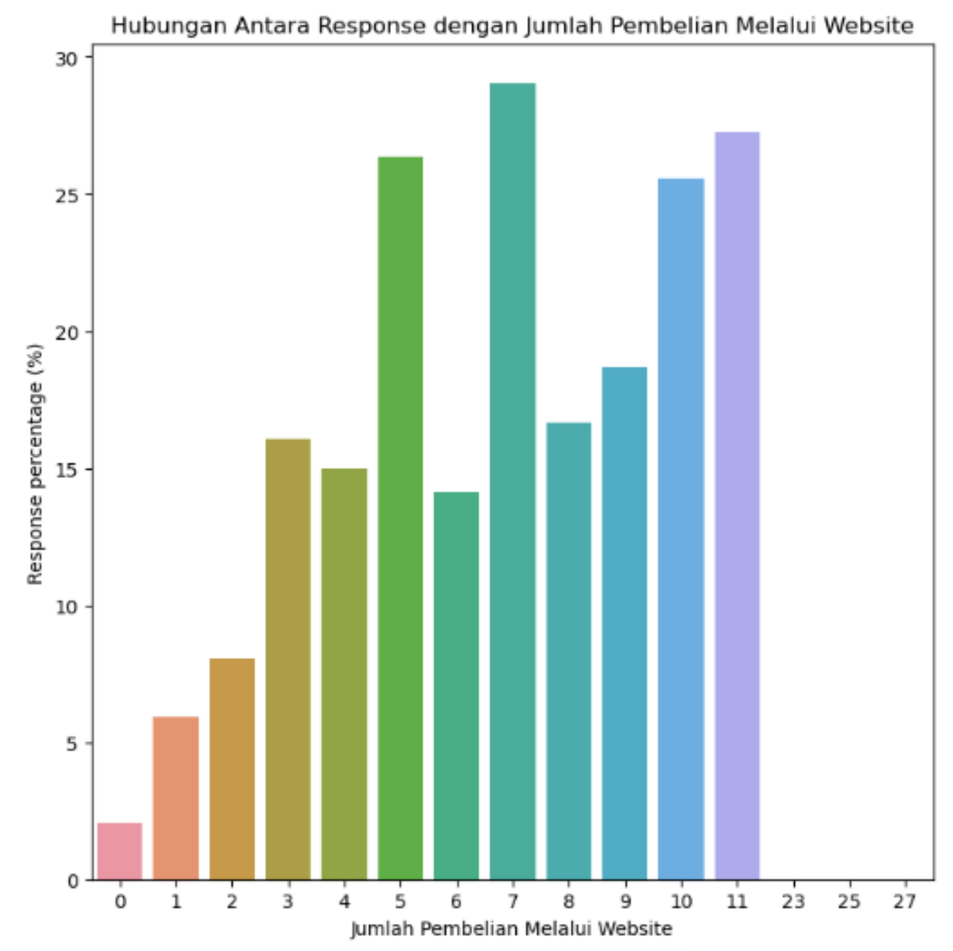
Semakin lama berlangganan, semakin besar potensi response

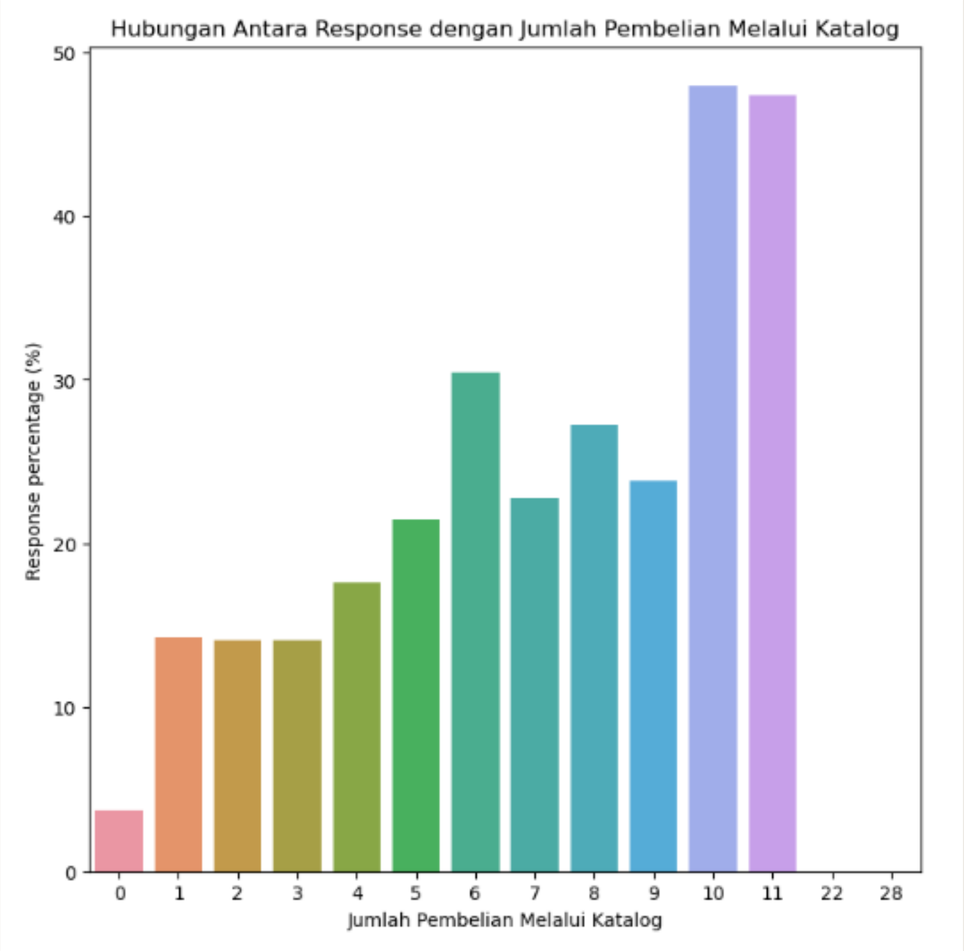
*** Notes :**

Percentage Response = presentase customer yang nilai response nya 1 dibagi total customer

MULTIVARIATE ANALYSIS

Response - Pembelian Melalui Website & Response - Pembelian Melalui Katalog

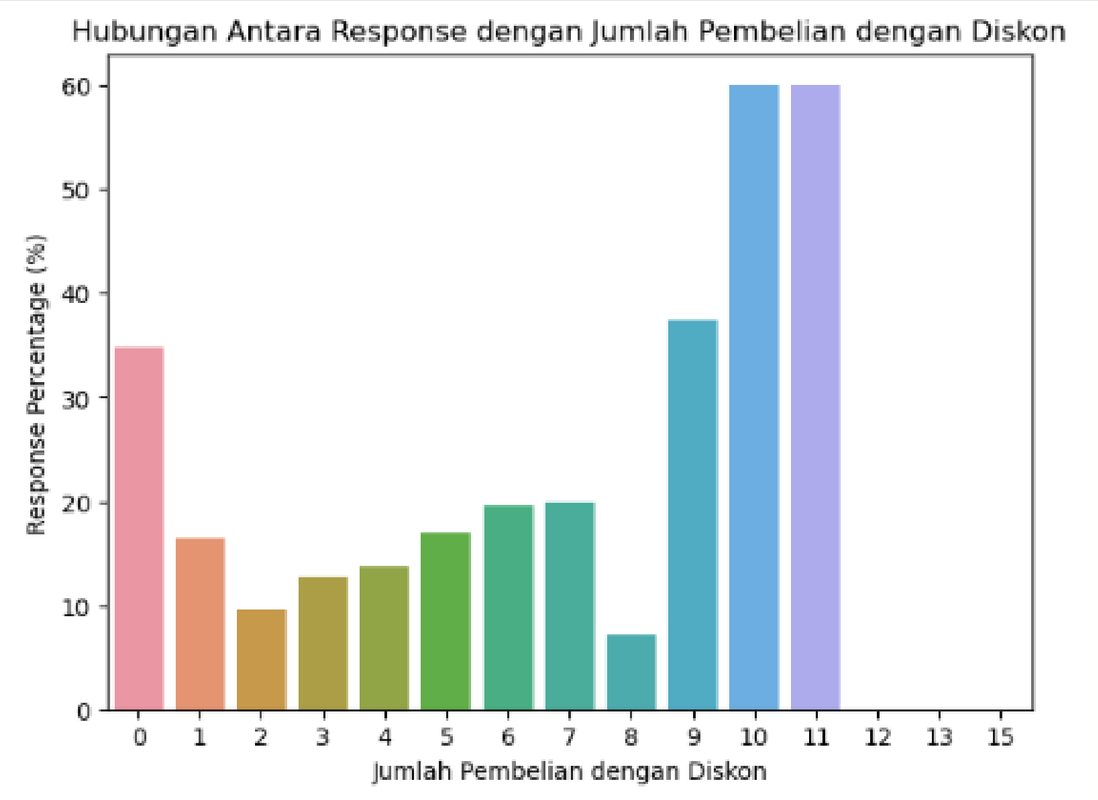




semakin sering customer belanja melalui catalog dan web, maka semakin besar potensi customer merespon campaign

MULTIVARIATE ANALYSIS

Response - Pembelian Dengan Diskon



- Semakin sering customer belanja dengan diskon, maka semakin besar peluang mereka merespon campaign
- **Terdapat pengecualian** pada customer yang tidak menerima diskon (Jumlah pembelian dengan diskon = 0).
- Customer yang tidak pernah menerima diskon namun tetap melakukan pembelian dapat diasumsikan sebagai customer loyal
- Customer yang menerima diskon 10 kali dan 11 kali memiliki kemungkinan lebih dari 50% untuk membeli

BUSINESS INSIGHT

Berdasarkan Multivariate Analysis antara response dengan beberapa fitur,

Peningkatan presentase customer yang meresponse campaign berbanding lurus dengan peningkatan :

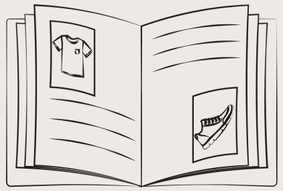
Spending

Lama berlangganan

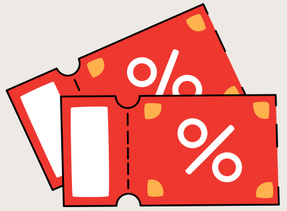
Tingkat Pendidikan

Pembelian melalui website

Income



Pembelian melalui katalog



Pembelian dengan diskon

CHAPTER 3

DATA PREPROCESSING

Proses persiapan dataset sebelum modelling

DATA PRE-PROCESSING (CONT'D)

1 Handling Missing Value

Baris kosong kolom`Income` dihilangkan karena hanya mencakup 1.07% dari jumlah data (lebih kecil dari 10%)

2 Data Type Conversion

Dt_Customer di convert dari **object type** ke **datetime type**

3 Feature Extraction

Menambahkan Feature '**kidsorteen**','**Spending**','**Year_customer**
'**campaign_result**'

4 Feature Encoding

Label Encoding:
Year_customer','marital_status'
One Hot Encoding:
'Generation', 'Education'

5 Feature Selection

Melakukan drop kolom:

- 'ID' ,
- 'MntWines,'MntMeatProducts','MntFish-Products','MntSweetProducts','MntGoldProds'
- Kidhome, Teenhome
- AcceptedCmp1,AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5
- Z_CostContact, Z_Revenue

DATA PRE-PROCESSING

6 Data Spliting

Train: 80% (1732)
Test : 20% (444)

7 Drop Outliers

Menghapus 3% data Outliers menggunakan **Z-Score**

dari **1772 baris data**
menjadi **1732 baris data**

8 Feature Transformation

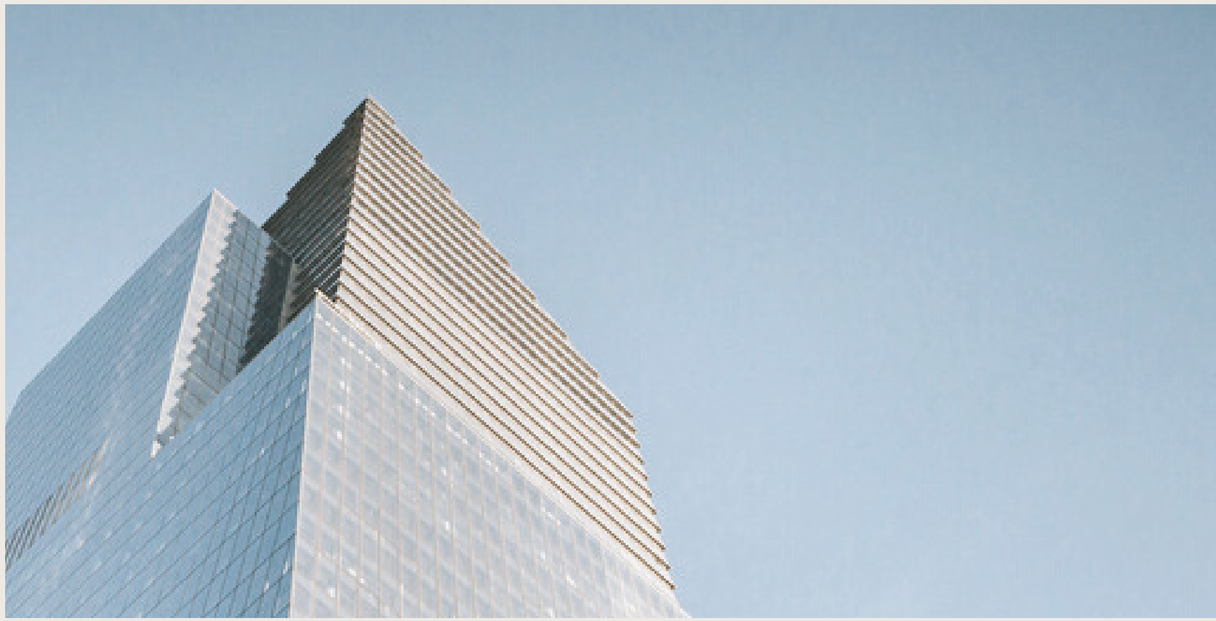
Melakukan **Logistic Transformation** pada data yang terindikasi ***'skewed ekstrim '*** dan menormalisasi menggunakan **MinMaxScaler**

9 Handle Imbalance Class

Melakukan **oversampling** menggunakan **RandomOversampling**

CHAPTER 4

MODELING



1 Model Result

2 Evaluation

3 Feature Importance

STAGE 3

MODELING

Algoritma

- Logistic Regression
- K-Nearest Neighbor
- Random Forest
- Decision Tree
- AdaBoost
- XGBoost

Metode Score Evaluasi

- Accuracy
- Precission
- Recall
- F-1 Score

Result

Test Set Model

	1 Accuracy	2 Precission	3 Recall	4 F-1 Score
Logistic Regression	0.88	0.59	0.35	0.44
K-Nearest Neighbor	0.87	0.59	0.21	0.31
Random Forest	0.89	0.69	0.40	0.51
Decision Tree	0.84	0.44	0.40	0.42
AdaBoost	0.88	0.57	0.52	0.54
XGBoost	0.89	0.68	0.45	0.54

Result

Test Set Model Hyperparameter

	1 Accuracy	2 Precission	3 Recall	4 F-1 Score
Logistic Regression	0.80	0.39	0.84	0.54
K-Nearest Neighbor	0.77	0.32	0.58	0.41
Random Forest	0.86	0.49	0.68	0.57
Decision Tree	0.82	0.37	0.42	0.39
AdaBoost	0.81	0.40	0.74	0.52
XGBoost	0.84	0.47	0.79	0.59

Recall terbesar

EVALUATION (CONFUSION MATRIX)

Precission

False Positive : Model memprediksi customer **response**, aktual **tidak**

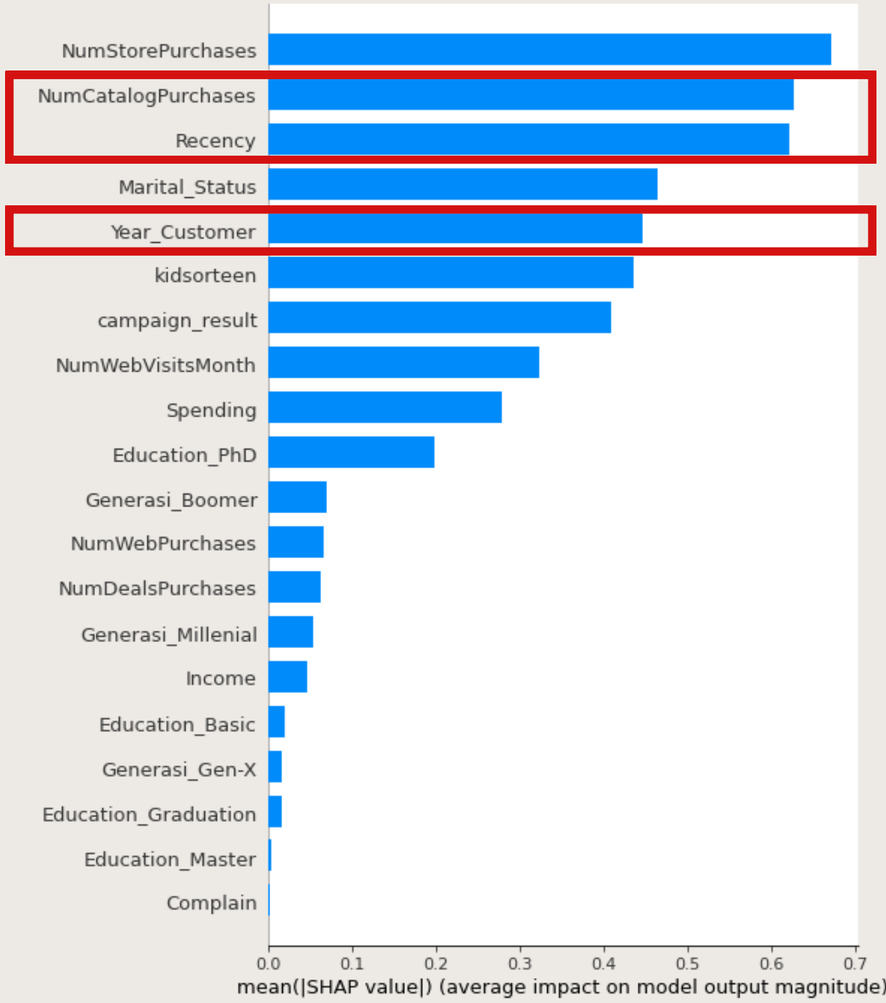
Impact : Cost campaign meningkat

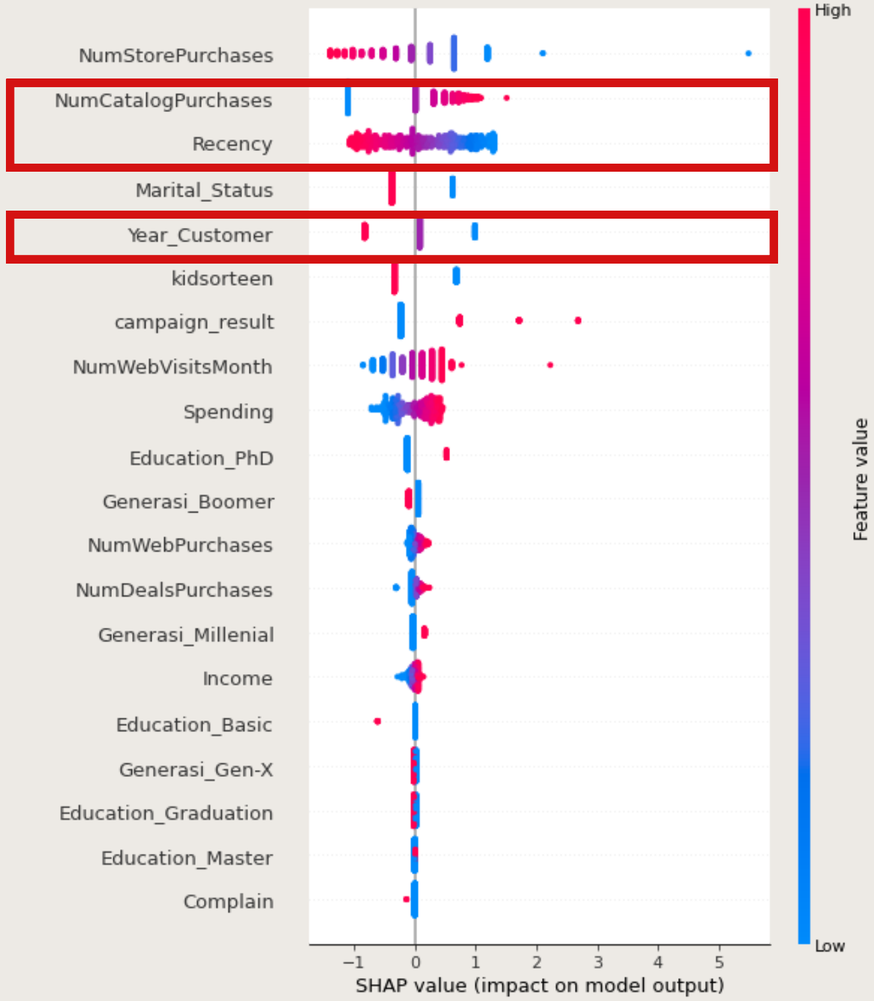
Recall

False Negative : Model memprediksi customer **tidak response**, aktual **response**

Impact : Loss Potential Revenue

LOGISTIC REGRESSION FEATURE IMPORTANCE

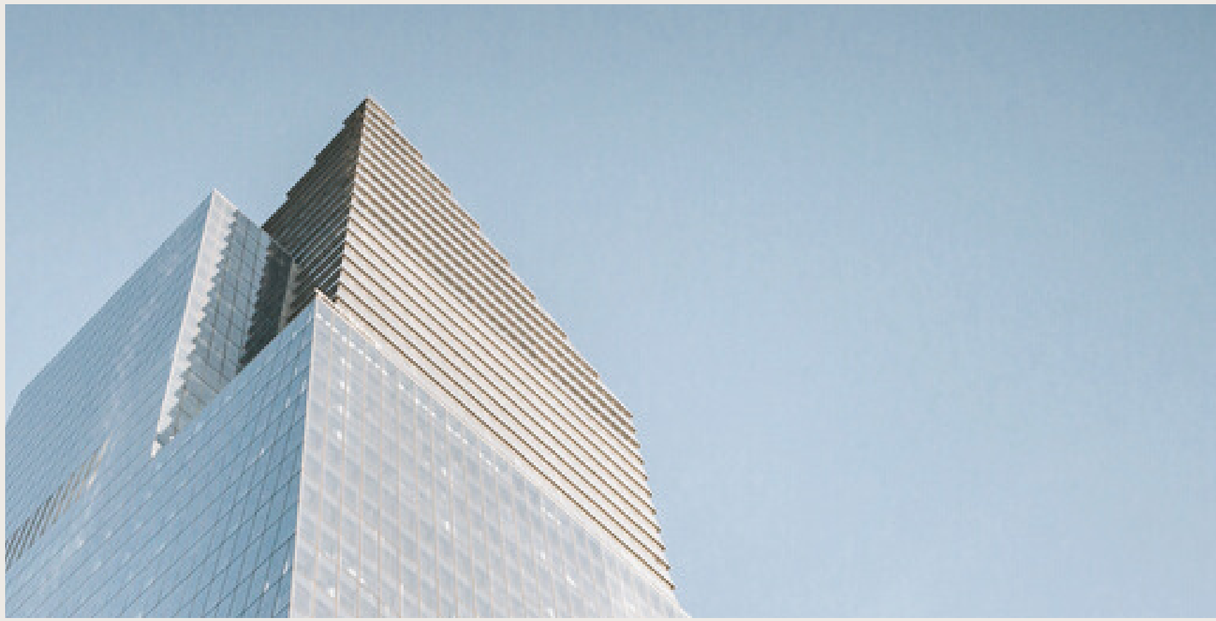




Feature Importance :

- NumCatalogPurchases
- Recency
- Year Customer

BUSINESS RECOMMENDATION & SIMULATION



1

Business Recommendation

2

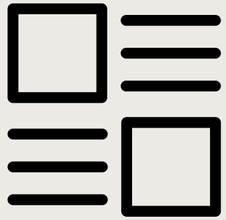
Business Simulation

BUSINESS RECOMMENDATION

Loyalty Program

CAC is more expensive than keeping the current customer

Memberikan voucher dan rekomendasi produk untuk customer lama yang tidak berbelanja pada rentang tertentu agar kembali berbelanja



Customer Experience via Catalog and Website shopping

Most customer response in line with : **Catalog & website shopping**

Optimalisasi rekomendasi produk berdasarkan *most buy* dan menawarkan **up-selling dan cross selling**

Voucher for minimum spent

Most **recent** customer have bigger probability on accepting campaign

Voucher / diskon yang dapat digunakan dalam periode terbatas agar customer kembali berbelanja, setelah serangkaian pembelanjaan

BUSINESS SIMULATION

- Campaign cost per customer 3 USD / customer
- Revenue dari campaign yang berhasil 11 USD / customer

	Target	Accepted	% Profit
Blind Marketing	444 Customers	62 customer (14 %)	- 48.8 %
Model Based	132 Customers	52 customer (39.39%)	44.44 %

Margin Profit 92.4%

PROFIT CALCULATION

	Blind Marketing	Model Based	Difference
Target	444	132	- 312
Cost	1332 USD	396 USD	- 936 USD
Potential Revenue	682 USD	572 USD	- 110 USD
Profit	-650 USD	176 USD	826 USD
Profit percentage	-48.80	44.44%	Decition

THANK YOU!

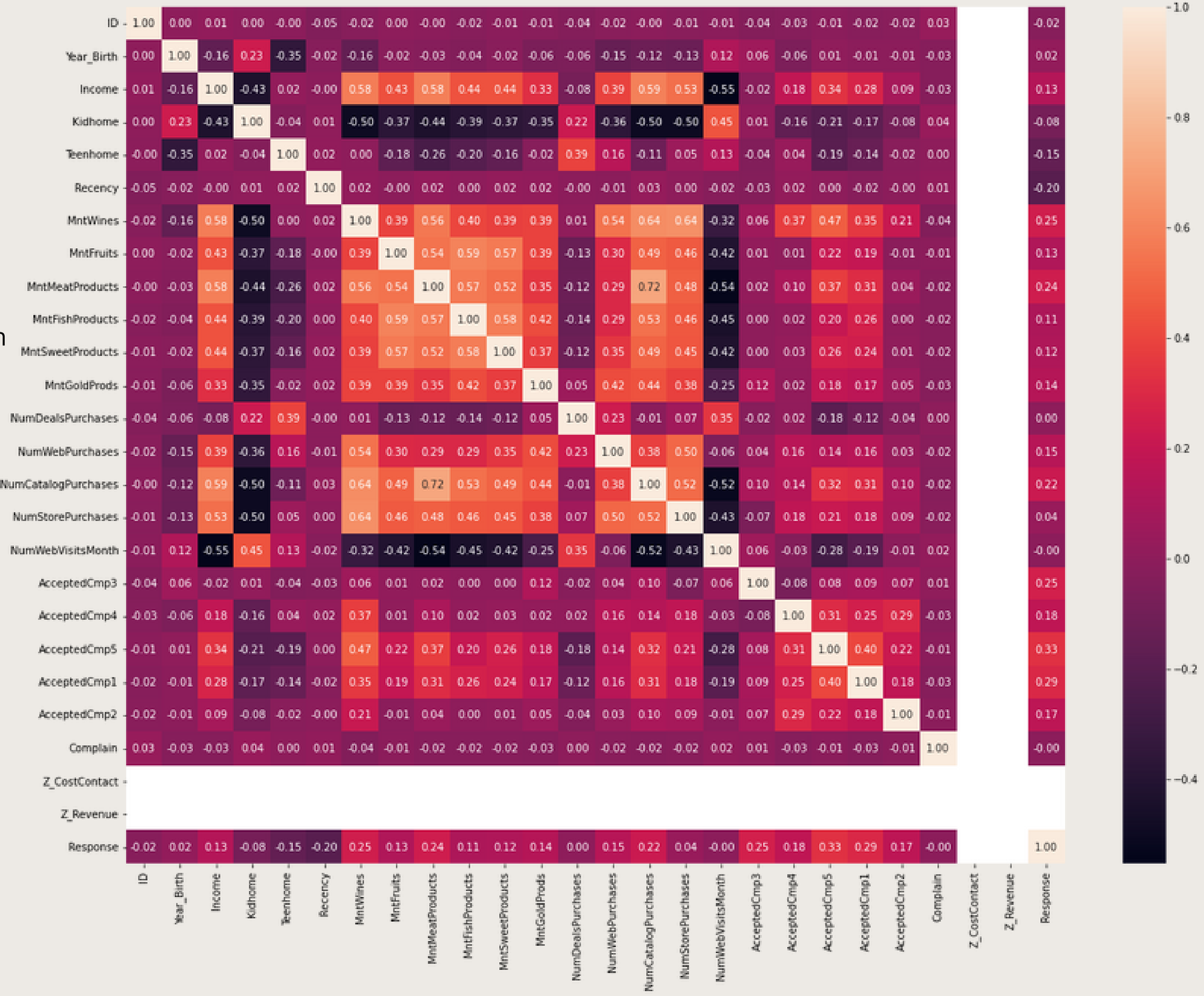
Appendix

Heatmap

Tidak ada korelasi linear yang kuat antara masing-masing feature dan target, karena nilai korelasi dibawah 0.5,sehingga feature-feature yang akan dipertahankan baru dapat diketahui pada stage pemilihan feature importance

- Customer yang memiliki anak cenderung memilih berbelanja menggunakan diskon. Pada heatmap fitur ini memiliki koefisien korelasi lebih tinggi dibanding dengan belanja melalui channel lain.

- Semakin besar income semakin banyak spending di tiap kategori. Pada heatmap, fitur-fitur ini memiliki koefisien korelasi berkisar di antara 0.33 - 0.58.



Hyperparameter

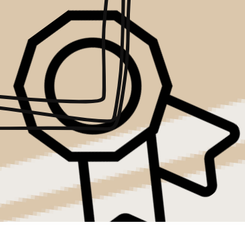
Precision Test

0.39

VS

Recall Test

0.84

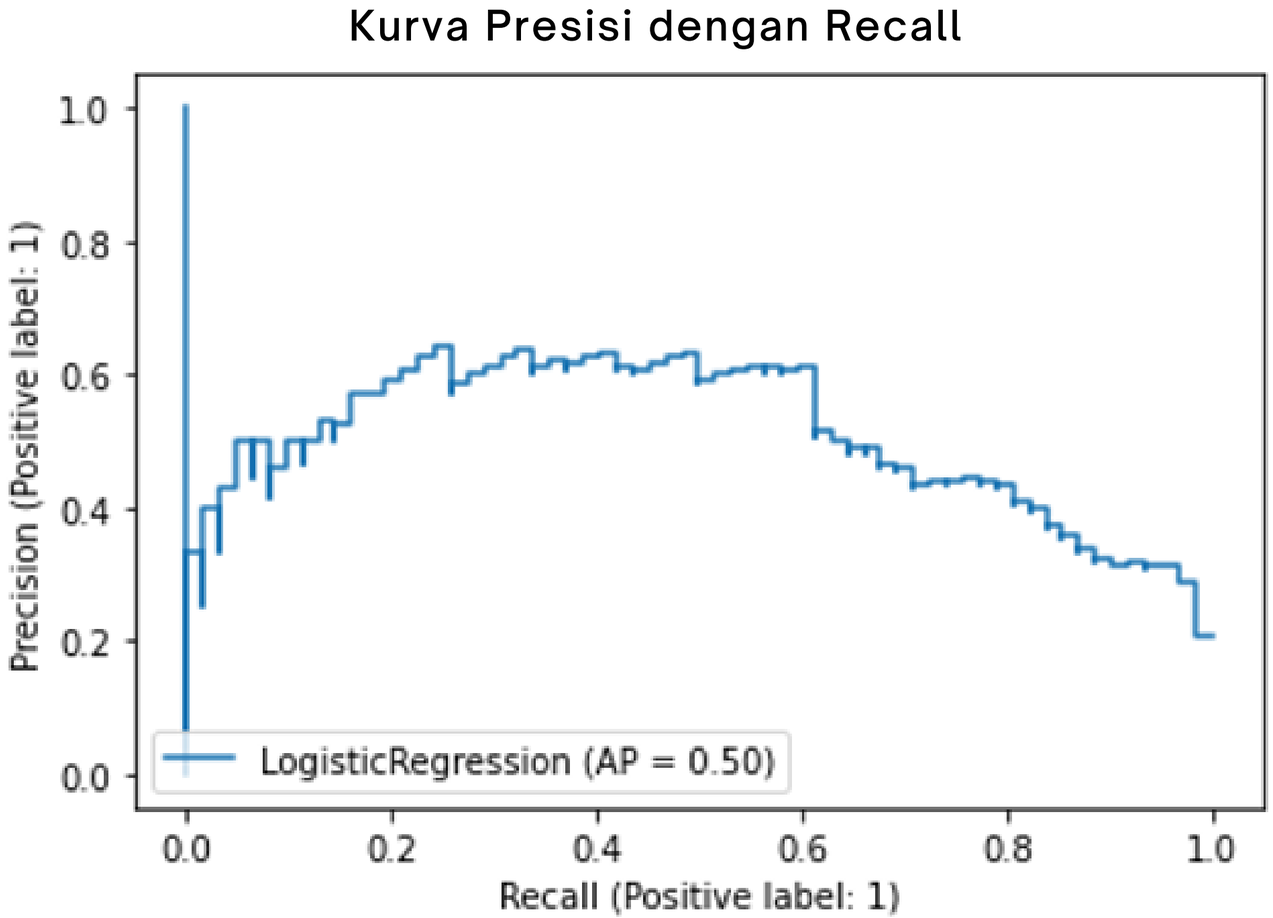


```
[ ]  lg_best = LogisticRegression(C= 0.41419999999999996, class_weight= 'balanced', dual= False, fit_intercept= True,
                                  intercept_scaling= 1, l1_ratio= None, max_iter= 1000, multi_class= 'auto', n_jobs= None,
                                  penalty= 'l2', random_state= 42, solver= 'lbfgs', tol= 0.0001, verbose= 0, warm_start= False)
     lg_best.fit(X_train, y_train)
```

Result of Hyperparameter

```
Confusion Matrix:
[[302  80]
 [ 10  52]]
Accuracy (Test Set): 0.80
Accuracy (Train Set): 0.82
Precision (Test Set): 0.39
Precision (Train Set: 0.45
Recall (Test Set): 0.84
Recall (Train Set): 0.83
F1-Score (Test Set): 0.54
F1-Score (Train Set): 0.58
roc_auc (test-proba): 0.88
roc_auc (train-proba): 0.91
recall (crossval train): 0.7575061248627185
recall (crossval test): 0.7750791497060153
```

Korelasi Presisi - Recall



Tidak terdapat nilai Recall -Precision yang optimal

Sebelum Hyperparameter Tuning

	Accuracy (Test)	Accuracy (Train)	Precission (Test)	Precission (Train)	Recall (Test)	Recall (Train)	F-1 Score (Test)	F1-Score (Train)
Logistic Regression	0.88	0.9	0.59	0.79	0.35	0.45	0.44	0.58
Decision Tree	0.84	0.99	0.44	1	0.4	0.95	0.42	0.97
Random Forest	0.89	0.99	0.69	0.98	0.4	0.97	0.51	0.98
K-Nearest Neighbors	0.87	0.89	0.59	0.85	0.21	0.37	0.31	0.52
AdaBoost	0.88	0.9	0.57	0.74	0.52	0.53	0.54	0.62
XGBoost	0.89	0.94	0.68	0.93	0.45	0.62	0.54	0.75

Setelah Hyperparameter Tuning

	Accuracy (Test)	Accuracy (Train)	Precission (Test)	Precission (Train)	Recall (Test)	Recall (Train)	F-1 Score (Test)	F1-Score (Train)
Logistic Regression	0.8	0.82	0.39	0.45	0.84	0.83	0.54	0.58
Decision Tree	0.82	0.99	0.37	1	0.42	0.95	0.39	0.97
Random Forest	0.86	0.96	0.49	0.81	0.68	0.97	0.57	0.88
K-Nearest Neighbors	0.77	0.89	0.32	0.57	0.58	0.98	0.41	0.72
AdaBoost	0.81	0.85	0.4	0.51	0.74	0.83	0.52	0.63
XGBoost	0.84	0.89	0.47	0.59	0.79	0.91	0.59	0.72