

Market Insider

Stage 0
Laporan Project



Latar Belakang Masalah

- Sebuah perusahaan melaksanakan kampanye pemasaran, di mana kampanye pemasaran terakhirnya berhasil meraih respons sebesar 14.91% dari 2240 pelanggan **(Problem)**. Merespons permasalahan ini, tujuan utama perusahaan adalah meningkatkan jumlah pelanggan yang merespons kampanye pemasaran berikutnya **(Goal)**.
- Dalam usaha meningkatkan efektivitas kampanye pemasaran, kami akan mengembangkan model Machine Learning sebagai Pendekatan Analitis **(Objective)**. Model ini bertujuan memprediksi respons pelanggan berdasarkan data historis, memungkinkan penyesuaian strategi pemasaran secara real-time, dan mengoptimalkan pengeluaran biaya marketing.
- Fokus kami adalah menargetkan pelanggan potensial konversi tinggi, dengan **Conversion Rate (CR)** sebagai **metrik bisnis** utama untuk mengukur keberhasilan pencapaian objektif kami. Dengan menggabungkan analisis pola-pola dalam data historis dan pendekatan analitis, kami yakin model ini akan memberikan wawasan mendalam untuk optimalisasi strategi pemasaran dan hasil kampanye yang lebih baik.

Pembagian Tugas

Pada proses penggeraan Project ini, kami membagi team kami menjadi 2 mini team, yaitu:

1. Team Data Analyst:

- a. Achmad Hilman Shadiqin
- b. Riyan Maula
- c. Nabilah Astiarini

2. Team Data Scientist:

- a. Andreawan Sofian
- b. Figo Akmal Munir
- c. Dzakwan Darussalam

Market Insider

Stage 1
Laporan Project



1. Descriptive Statistics

data.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 29 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   ID               2240 non-null    int64  
 1   Year_Birth        2240 non-null    int64  
 2   Education         2240 non-null    object  
 3   Marital_Status    2240 non-null    object  
 4   Income            2216 non-null    float64 
 5   Kidhome          2240 non-null    int64  
 6   Teenhome          2240 non-null    int64  
 7   Dt_Customer       2240 non-null    object  
 8   Recency           2240 non-null    int64  
 9   MntWines          2240 non-null    int64  
 10  MntFruits         2240 non-null    int64  
 11  MntMeatProducts   2240 non-null    int64  
 12  MntFishProducts   2240 non-null    int64  
 13  MntSweetProducts  2240 non-null    int64  
 14  MntGoldProds      2240 non-null    int64  
 15  NumDealsPurchases 2240 non-null    int64  
 16  NumWebPurchases   2240 non-null    int64  
 17  NumCatalogPurchases 2240 non-null    int64  
 18  NumStorePurchases 2240 non-null    int64  
 19  NumWebVisitsMonth 2240 non-null    int64  
 20  AcceptedCmp3       2240 non-null    int64  
 21  AcceptedCmp4       2240 non-null    int64  
 22  AcceptedCmp5       2240 non-null    int64  
 23  AcceptedCmp1       2240 non-null    int64  
 24  AcceptedCmp2       2240 non-null    int64  
 25  Complain          2240 non-null    int64  
 26  Z_CostContact     2240 non-null    int64  
 27  Z_Revenue          2240 non-null    int64  
 28  Response           2240 non-null    int64  
dtypes: float64(1), int64(25), object(3)
memory usage: 507.6+ KB
```

data.nunique()

ID	2240
Year_Birth	59
Education	5
Marital_Status	8
Income	1974
Kidhome	3
Teenhome	3
Dt_Customer	663
Recency	100
MntWines	776
MntFruits	158
MntMeatProducts	558
MntFishProducts	182
MntSweetProducts	177
MntGoldProds	213
NumDealsPurchases	15
NumWebPurchases	15
NumCatalogPurchases	14
NumStorePurchases	14
NumWebVisitsMonth	16
AcceptedCmp3	2
AcceptedCmp4	2
AcceptedCmp5	2
AcceptedCmp1	2
AcceptedCmp2	2
Complain	2
Z_CostContact	1
Z_Revenue	1
Response	2

dtype: int64

data.duplicated()

	duplicated	rows	percentage
0		182	8.12
	Education	Marital_Status	Income
1501	2n Cycle	Divorced	31163.0
1107	2n Cycle	Married	15315.0
1171	2n Cycle	Married	20130.0
383	2n Cycle	Married	35688.0
2015	2n Cycle	Married	37284.0
...
1218	PhD	Together	68682.0
2062	PhD	Together	70038.0
1630	PhD	Together	73059.0
1959	PhD	Widow	56551.0
2217	PhD	Widow	82032.0

182 rows × 3 columns

1. Descriptive Statistics

data.describe()

a. Numerical Features

	count	mean	std	min	10%	20%	30%	40%	50%	60%	70%	80%	90%	max
Year_Birth	2240.0	1968.805804	11.984069	1893.0	1952.0	1957.0	1962.0	1966.0	1970.0	1973.0	1976.0	1979.0	1984.0	1996.0
Income	2216.0	52247.251354	25173.076661	1730.0	24117.5	32011.0	38198.5	44529.0	51381.5	58482.0	65247.5	71819.0	79844.0	666666.0
Kidhome	2240.0	0.444196	0.538398	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	2.0
Teenhome	2240.0	0.506250	0.544538	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	2.0
Recency	2240.0	49.109375	28.962453	0.0	9.0	19.0	29.0	39.0	49.0	59.0	69.0	79.0	89.0	99.0
MntWines	2240.0	303.935714	336.597393	0.0	6.0	16.0	34.0	81.0	173.5	284.4	418.6	581.2	822.1	1493.0
MntFruits	2240.0	26.302232	39.773434	0.0	0.0	1.0	2.0	4.0	8.0	15.0	25.0	44.0	83.0	199.0
MntMeatProducts	2240.0	166.950000	225.715373	0.0	7.0	12.0	20.0	35.0	67.0	108.4	177.0	298.4	499.0	1725.0
MntFishProducts	2240.0	37.525446	54.628979	0.0	0.0	2.0	3.0	7.0	12.0	20.0	37.0	65.0	120.0	259.0
MntSweetProducts	2240.0	27.062946	41.280498	0.0	0.0	1.0	2.0	5.0	8.0	14.0	26.0	44.2	89.0	263.0
MntGoldProds	2240.0	44.021875	52.167439	0.0	3.0	6.0	11.0	17.0	24.0	34.0	46.0	73.0	122.0	362.0
NumDealsPurchases	2240.0	2.325000	1.932238	0.0	1.0	1.0	1.0	1.0	2.0	2.0	3.0	3.0	5.0	15.0
NumWebPurchases	2240.0	4.084821	2.778714	0.0	1.0	2.0	2.0	3.0	4.0	4.0	5.0	6.0	8.0	27.0
NumCatalogPurchases	2240.0	2.662054	2.923101	0.0	0.0	0.0	1.0	1.0	2.0	2.0	4.0	5.0	7.0	28.0
NumStorePurchases	2240.0	5.790179	3.250958	0.0	2.0	3.0	3.0	4.0	5.0	6.0	7.0	9.0	11.0	13.0
NumWebVisitsMonth	2240.0	5.316518	2.426645	0.0	2.0	3.0	4.0	5.0	6.0	6.0	7.0	7.0	8.0	20.0
AcceptedCmp1	2240.0	0.064286	0.245316	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
AcceptedCmp2	2240.0	0.013393	0.114976	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
AcceptedCmp3	2240.0	0.072768	0.259813	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
AcceptedCmp4	2240.0	0.074554	0.262728	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
AcceptedCmp5	2240.0	0.072768	0.259813	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Complain	2240.0	0.009375	0.096391	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Response	2240.0	0.149107	0.356274	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0

b. Categorical Features

	count	unique	top	freq
Education	2240	5	Graduation	1127
Marital_Status	2240	8	Married	864

c. Date time Features

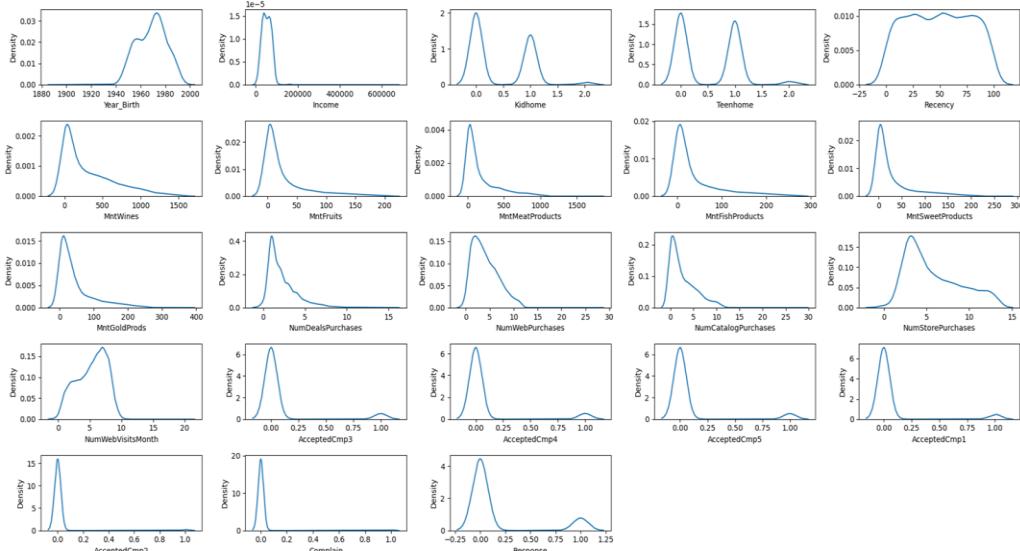
	count	unique	top	freq	first	last
Dt_Customer	2240	663	2012-08-31	12	2012-07-30	2014-06-29

1. Descriptive Statistics

1. **Apakah ada kolom dengan tipe data kurang sesuai, atau nama kolom dan isinya kurang sesuai?**
Feature **Dt_Customer** merupakan tanggal registrasi pelanggan dengan tipe data object. Tipe data ini tidak sesuai, sehingga perlu diubah menjadi tipe Date Time.
1. **Apakah ada kolom yang memiliki nilai kosong? Jika ada, apa saja?**
Feature **Income** mempunyai nilai kosong karena hanya berjumlah 2216 rows.
1. **Apakah ada kolom yang memiliki nilai summary agak aneh? (min/mean/median/max/unique/top/freq)**
 - Feature **ID** merupakan identifikasi pelanggan. Berdasarkan analisis jumlah nilai unik pada feature tersebut, diketahui bahwa jumlahnya sama dengan jumlah baris dataset (2240), sehingga tidak memungkinkan untuk mengamati riwayat perjalanan pelanggan.
 - Feature **Z_CostContact** dan **Z_Revenue** hanya memiliki satu data unik, maka keduanya tidak akan memberikan hasil analisis yang signifikan.
 - Feature **Dt_Customer**, pelanggan paling terakhir melakukan registrasi di 29 Juni 2014, maka dengan asumsi saat ini adalah tahun 2014, ada keanehan pada Feature **Year Birth** dimana tahun lahir tertua ada di tahun 1893 atau usia pelanggan 121 tahun. Hal ini merupakan hal yang kurang masuk akal. Diduga terdapat kesalahan input tahun lahir oleh pelanggan/kesalahan pencatatan oleh sistem.
 - Feature **Income** memiliki keanehan karena memiliki nilai maksimum mencapai ratusan ribu (666.666), sedangkan nilai ukuran pemasaran dan penyebarannya hanya mencapai puluhan ribu. Diduga nilai ini merupakan outlier yang disebabkan karena kesalahan input atau pencatatan oleh sistem.
 - Feature **MntFishProducts**, **MntFruits**, **MntGoldProds**, **MntMeatProducts**, **MntSweetProducts**, **MntWines** memiliki keanehan dilihat dari nilai maksimumnya yang jauh dari nilai ukuran pemasaran atau ukuran penyebaran lainnya. Sehingga diduga terdapat nilai outlier pada feature tersebut.
 - Feature **Marital Status** memiliki keanehan karena memiliki 8 nilai unik. Maka pada tahap selanjutnya akan dianalisis lebih dalam setiap nilai pada feature ini.

2. Univariate Analysis

a. Numerical Features



Berdasarkan kdeplot di samping, diketahui distribusi dari feature dan target sebagai berikut:

1. Distribusi Normal:

Recency

1. Distribusi Left-Skewed (median > mean):

- Year_Birth

1. Distribusi Right-Skewed (mean > median)

- Income

- MntFishProducts, MntFruits, MntGoldProd, MntMeatProducts, MntSweetProducts, MntWines

- NumCatalogPurchases, NumDealsPurchases, NumStorePurchases, NumWebPurchases, NumWebVisitsMonth

- AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5 (didominasi dengan value 0)

- Responsee

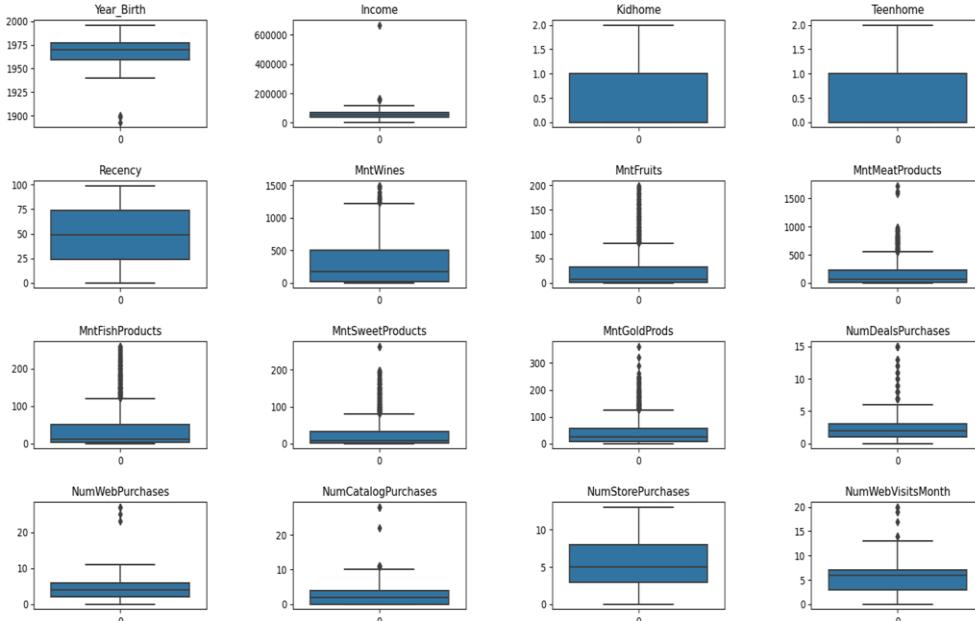
1. Distribusi Bimodal

- Kidhome

- Teenhome

2. Univariate Analysis

a. Numerical Features

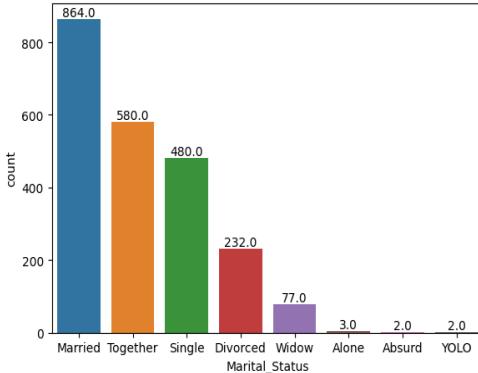
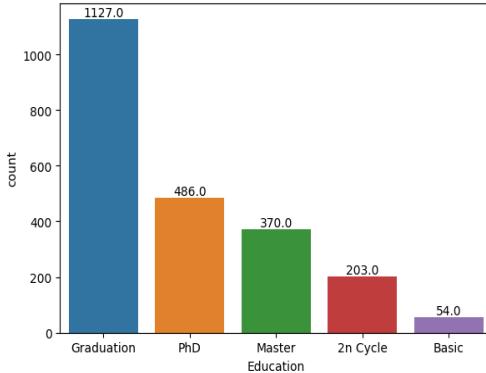


Berdasarkan boxplot di samping, dapat dilihat feature yang mengandung nilai outlier sebagai berikut:

1. Year_Birth memiliki outlier yaitu pada nilai minimumnya (di bawah tahun 1900)
2. Income memiliki nilai outlier yang cukup ekstrim yaitu pada nilai maksimumnya.
3. MntFishProducts, MntFruits, MntGoldProd, MntMeatProducts, MntSweetProducts, MntWines memiliki nilai outlier yang cukup banyak yaitu pada bagian atas boxplot (nilai tinggi)
4. NumCatalogPurchases, NumDealsPurchases, NumStorePurchases, NumWebPurchases, NumWebVisitsMonth memiliki nilai outlier yang cukup banyak yaitu pada bagian atas boxplot (nilai tinggi)

2. Univariate Analysis

b. Categorical Features

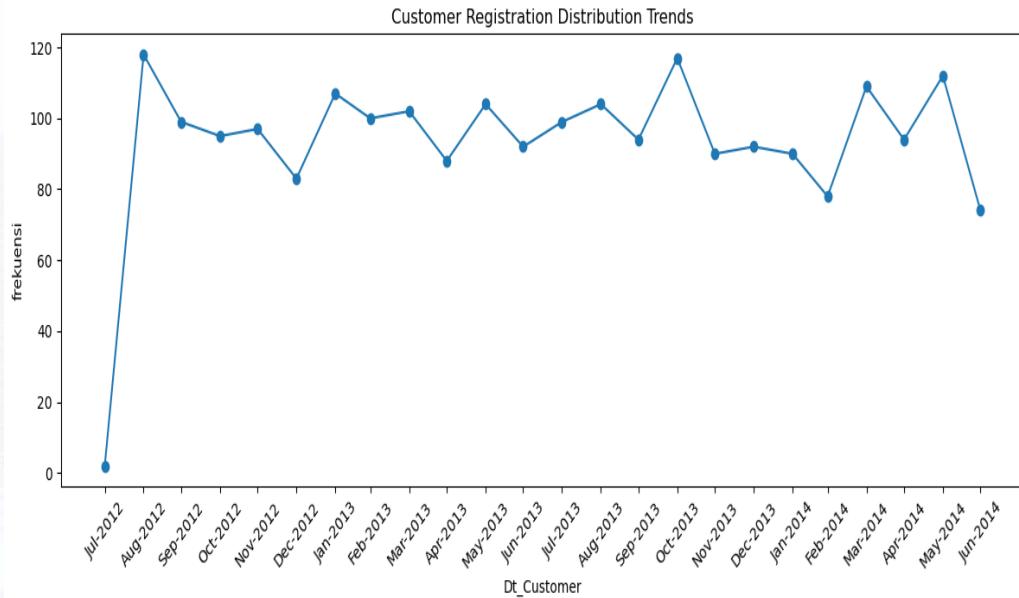


Berdasarkan grafik di samping, diketahui bahwa:

1. Pelanggan yang mempunyai level pendidikan Graduation adalah yang terbanyak, dilanjutkan dengan PnD, Master, 2n Cycle, dan Basic.
2. Pelanggan dengan status perkawinannya Married adalah yang terbanyak, dilanjutkan dengan Together, Single, Divorced, Widow, Absurd, dan YOLO.
3. Level Pendidikan Basic dan 2n Cycle perlu dipahami lebih lanjut definisinya.
4. Status Perkawinan Alone, Absurd, dan YOLO dianggap sebagai outlier yang tidak terdefinisi.

2. Univariate Analysis

c. Date Time Features

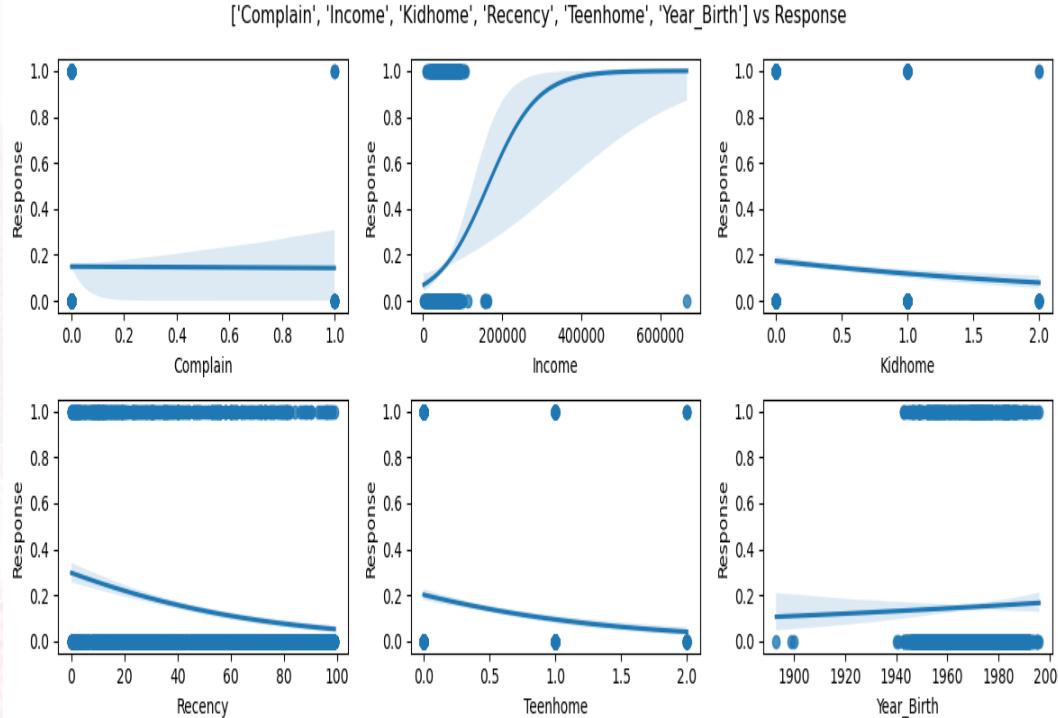


Berdasarkan grafik di samping, diketahui bahwa:

1. Sepanjang Juli-2012 hingga Juni-2014, jumlah customer yang mendaftar terbesar terjadi pada Agustus-2012 dan Oktober-2013 mencapai 120 customer baru.
2. Sedangkan Juli-2012 dan Juni-2014 adalah yang terendah.

3. Multivariate Analysis

a. Numerical Features



Berdasarkan Regression Plot disamping:

1.Income:

Korelasi positif yang kuat dengan Response, menunjukkan bahwa semakin tinggi pendapatan, semakin tinggi kemungkinan pelanggan merespon.

2.Jumlah Anak (Teenhome, Kidhome):

Korelasi negatif dengan Response, menunjukkan bahwa semakin sedikit jumlah anak, semakin tinggi kemungkinan pelanggan merespon.

3.Recency:

Korelasi negatif dengan Response, menunjukkan bahwa semakin lama sejak pelanggan terakhir berinteraksi, semakin tinggi kemungkinan pelanggan merespon.

4.Year_Birth:

Korelasi negatif dengan Response, menunjukkan bahwa semakin muda usia pelanggan, semakin tinggi kemungkinan pelanggan merespon.

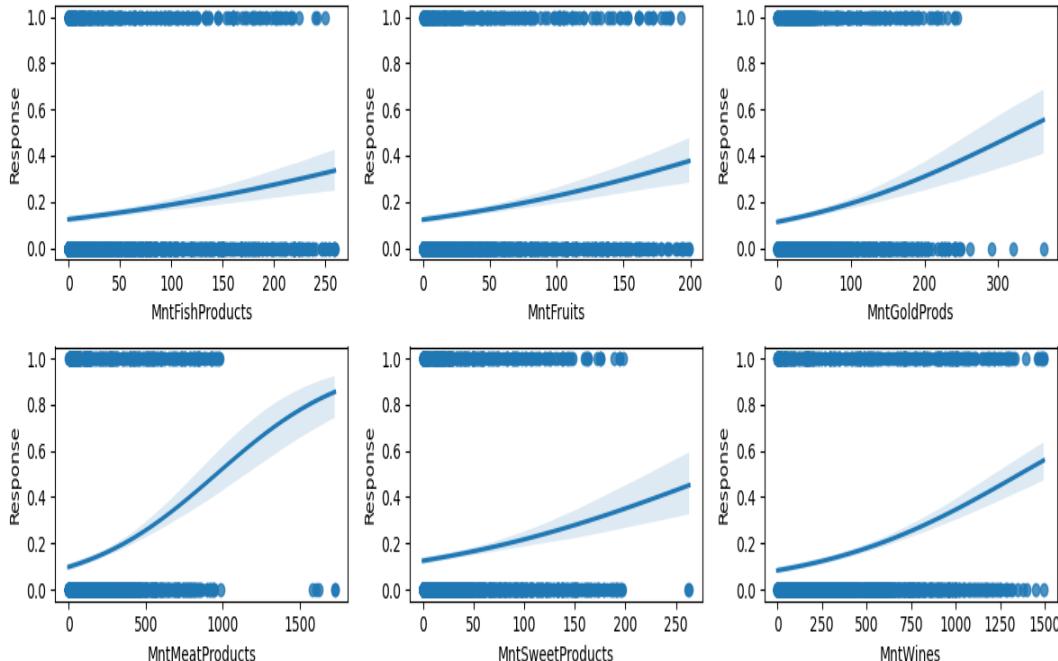
5.Complain:

Tidak terlihat adanya korelasi yang signifikan terhadap Response, ditandai dengan garis regresi yang cenderung lurus.

3. Multivariate Analysis

a. Numerical Features

[MntFishProducts, 'MntFruits', 'MntGoldProds', 'MntMeatProducts', 'MntSweetProducts', 'MntWines'] vs Response

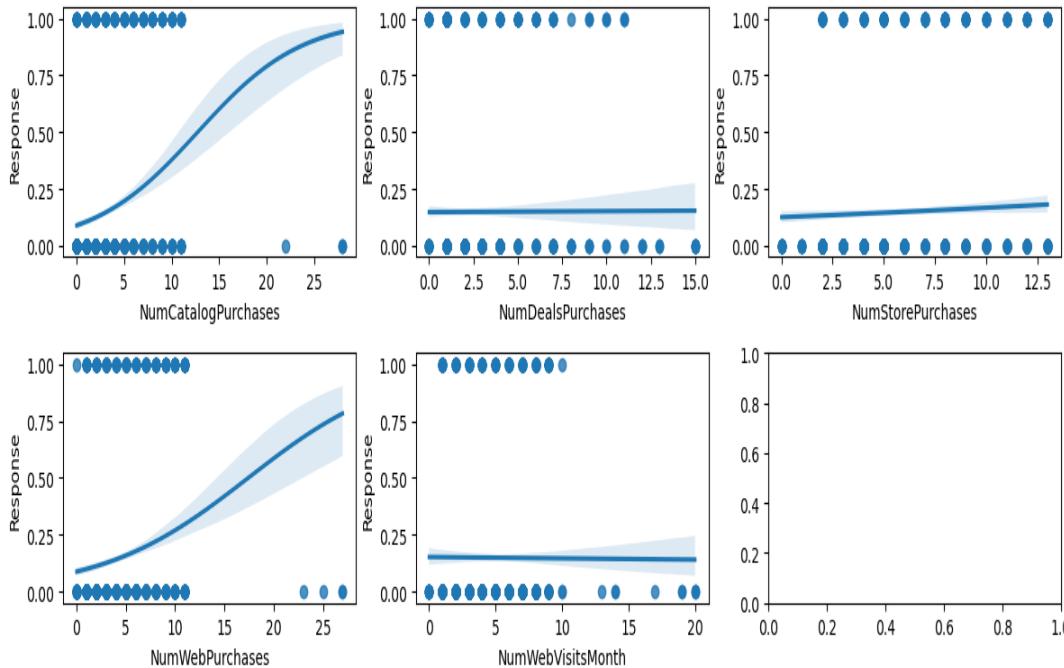


Berdasarkan regression plot di samping, diketahui bahwa semua feature amount spent product selama 2 tahun memiliki korelasi positif terhadap response dimana spending untuk produk Meat menunjukkan pola korelasi yang lebih kuat dibandingkan dengan produk lainnya.

3. Multivariate Analysis

a. Numerical Features

['NumCatalogPurchases', 'NumDealsPurchases', 'NumStorePurchases', 'NumWebPurchases', 'NumWebVisitsMonth'] vs Response

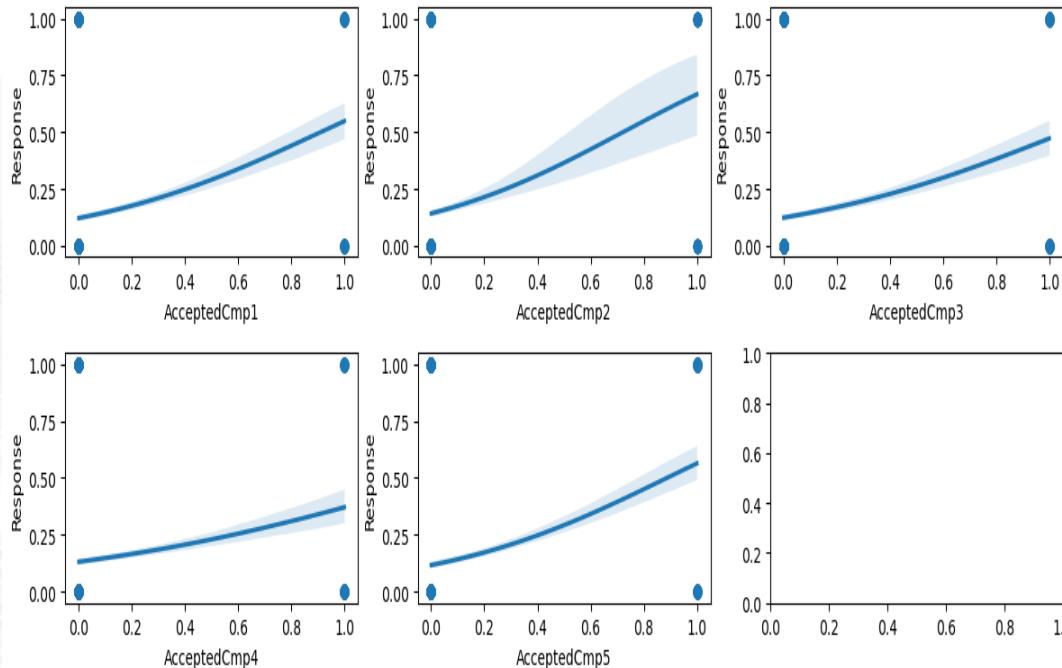


Berdasarkan regression plot di samping, diketahui bahwa pembelian yang memiliki korelasi yang kuat terhadap Response adalah pembelian melalui Katalog dan Web. Sementara pembelian dengan diskon (Deals) atau melalui Store tidak menunjukkan adanya korelasi yang signifikan terhadap Response.

3. Multivariate Analysis

a. Numerical Features

['AcceptedCmp1', 'AcceptedCmp2', 'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5'] vs Response

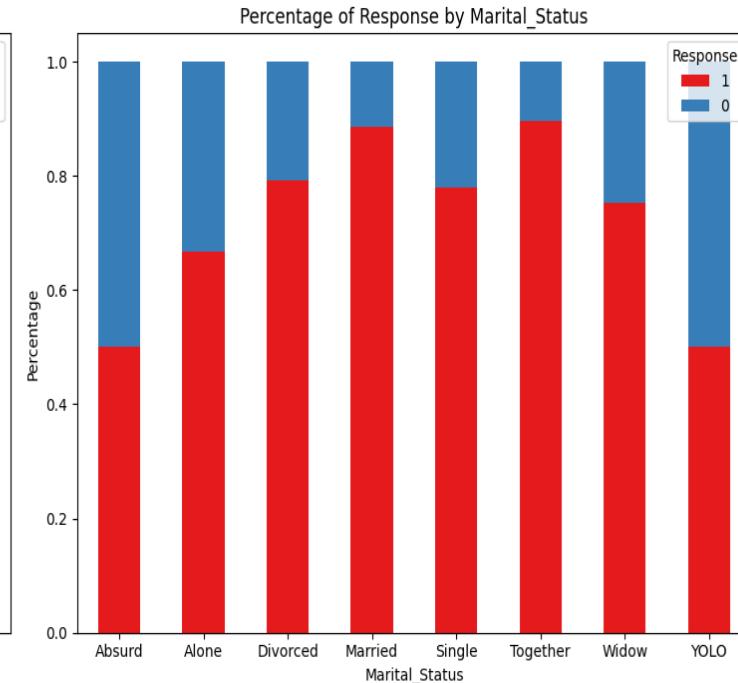
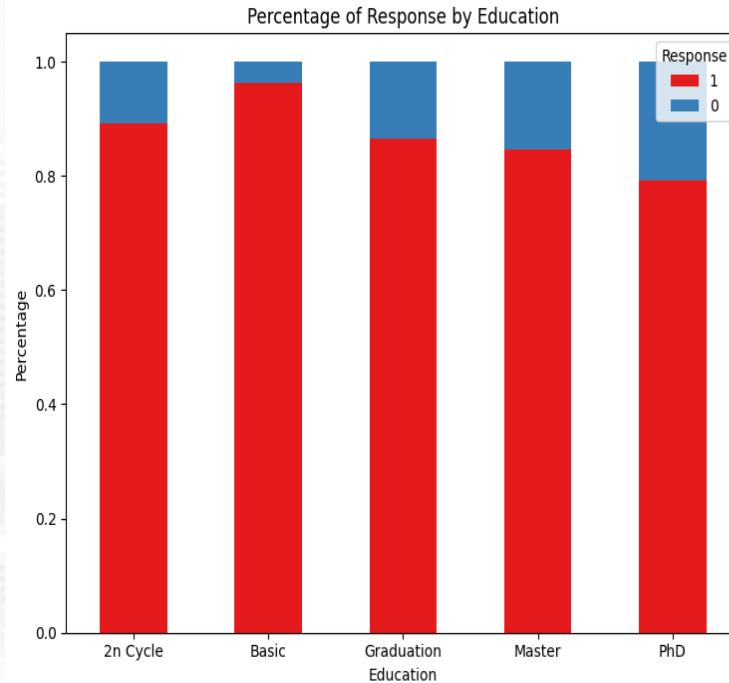


Berdasarkan regression plot di samping, diketahui bahwa kelima feature campaign memiliki korelasi positif terhadap Response.

Untuk membuktikan hasil dari semua Regression Plot, akan dibuat pula Correlation Heatmap untuk mengukur besar korelasi semua feature terhadap Response.

3. Multivariate Analysis

b. Categorical Features



3. Multivariate Analysis

b. Categorical Features

Insight Response Campaign Berdasarkan Tingkat Pendidikan:

- Tingkat Response Tinggi pada Tingkat Pendidikan Tinggi:**
Pelanggan dengan tingkat pendidikan PhD (20.78%) memiliki tingkat Response Campaign tertinggi, diikuti oleh Master (15.41%) dan Graduation (13.48%). Ini menunjukkan bahwa tingkat pendidikan tinggi berkorelasi positif dengan kemungkinan merespon Campaign.
- Varian Response pada Tingkat Pendidikan Rendah:**
Tingkat Response lebih bervariasi pada tingkat pendidikan rendah, di mana Basic (3.70%) menunjukkan Response lebih rendah dibandingkan dengan 2n Cycle (10.84%) dan Graduation (13.49%). Hal ini mungkin menunjukkan bahwa pelanggan dengan tingkat pendidikan lebih rendah cenderung memiliki tingkat Response yang lebih rendah terhadap Campaign.
- Pentingnya Pendidikan dalam Pengaruh Response:**
Pelanggan dengan tingkat pendidikan tinggi, seperti PhD dan Master, cenderung lebih Responsif terhadap Campaign. Ini dapat menjadi informasi kunci dalam merancang Campaign yang lebih efektif dan menargetkan kelompok pelanggan yang lebih cenderung merespon berdasarkan tingkat pendidikan mereka.

Insight Response Campaign Berdasarkan Status Perkawinan:

- Perbedaan Signifikan pada Tingkat Response:**
Terdapat perbedaan signifikan dalam tingkat Response Campaign antara status perkawinan. Pelanggan yang "Married" memiliki tingkat Response yang lebih rendah (11.34%) dibandingkan dengan pelanggan yang "Divorced" (20.69%), "Single" (22.08%), dan "Together" (10.35%).
- Tingkat Response Tinggi pada Status Perkawinan "Single" dan "Divorced":**
Pelanggan yang berstatus "Single" dan "Divorced" menunjukkan tingkat Response Campaign yang lebih tinggi, masing-masing sebesar 22.08% dan 20.69%. Ini menandakan bahwa status perkawinan ini dapat menjadi faktor penting dalam menentukan Response positif terhadap Campaign.
- Pentingnya Personalisasi Campaign untuk Setiap Status Perkawinan:**
Perusahaan dapat mempertimbangkan strategi pemasaran yang lebih personal dan disesuaikan dengan masing-masing status perkawinan. Status perkawinan seperti "Married" mungkin memerlukan pendekatan yang berbeda untuk meningkatkan Response Campaign.

3. Multivariate Analysis

c. Date Time Features

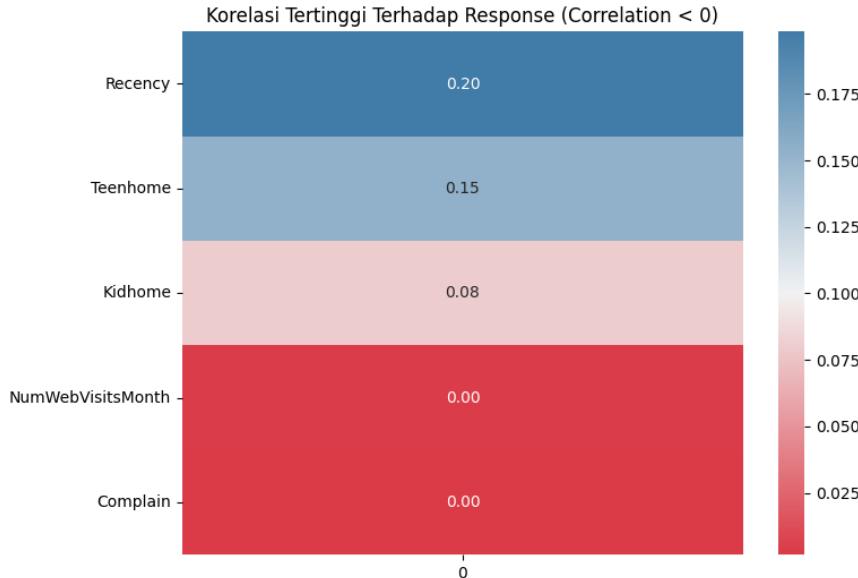
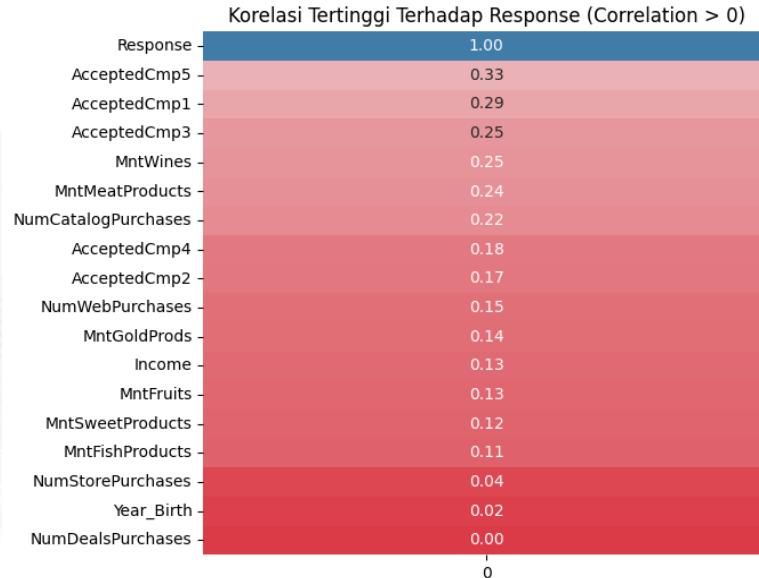


Berdasarkan grafik di samping, diketahui bahwa:

1. Dalam rentang waktu Juli 2012 hingga Juni 2014, lebih banyak pelanggan yang mendaftar namun tidak merespon campaign dibandingkan dengan yang merespon campaign.
2. Jumlah pelanggan yang mendaftar dan merespon campaign mencapai puncaknya pada Agustus 2012 - September 2012, yaitu kurang lebih 30 pelanggan per bulan, namun cenderung menurun hingga Juni 2014.

3. Multivariate Analysis

Berikut ini merupakan correlation heatmap untuk menganalisis korelasi antara feature dengan target (Responsee)



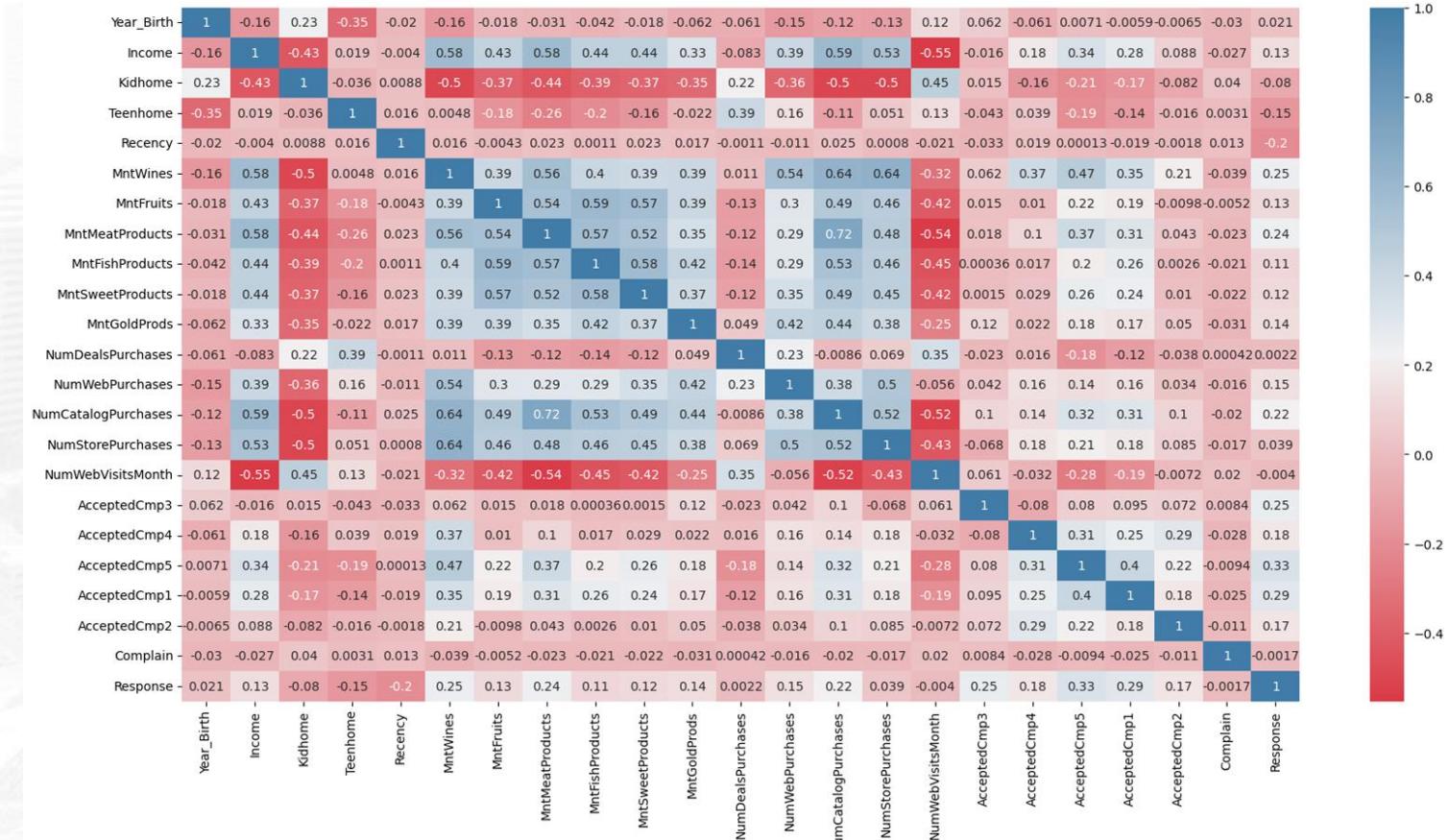
Dari hasil korelasi di atas, beberapa poin penting dapat diidentifikasi:

Feature yang termasuk dalam campaign (AcceptedCmp1, AcceptedCmp2, dst.), feature yang termasuk dalam produk spending seperti (MntWines, MntMeatProducts, dst.), feature yang termasuk pada channel penjualan seperti (NumCatalogPurchases, NumWebPurchases, dst.) mempunyai korelasi positif terhadap target.

Sedangkan hanya beberapa feature yang mempunyai nilai korelasi negatif terhadap target yaitu Recency dan jumlah anak (Teenhome dan Kidhome).

3. Multivariate Analysis

Berikut ini merupakan correlation heatmap untuk menganalisis korelasi antara feature dan target serta korelasi antar feature.



3. Multivariate Analysis

Dari hasil korelasi di atas, beberapa poin penting dapat diidentifikasi:

1. Korelasi Antara Feature:

Korelasi antara beberapa pasang feature cukup tinggi, yang bisa menunjukkan adanya multicollinearity. Contohnya, "MntMeatProducts" dengan "NumCatalogPurchases" (0.72), "MntWines" dengan "NumCatalogPurchases" dan "NumStorePurchases" (0.64), "Income" dan beberapa feature pengeluaran makanan ("MntWines", "MntFruits", "MntMeatProducts")

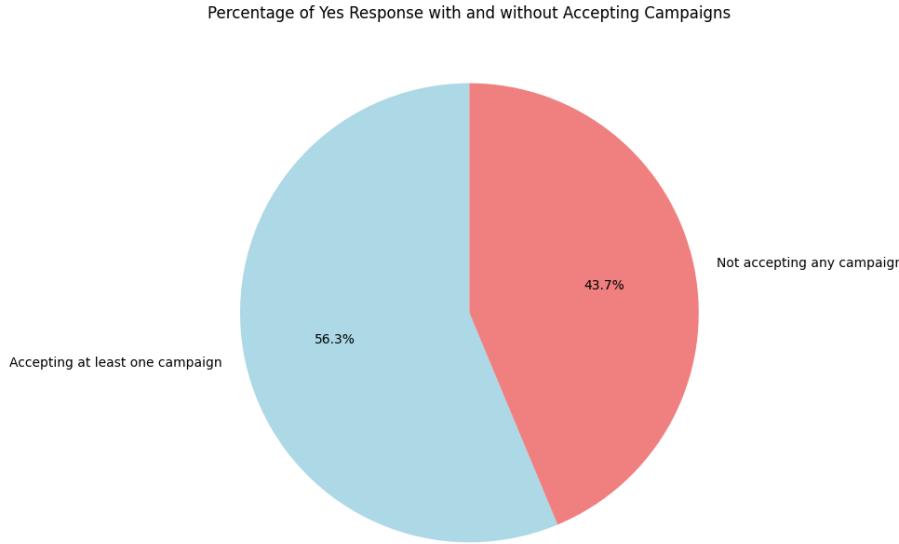
1. Rekomendasi untuk Proses Modeling:

Kami akan melakukan analisis lebih lanjut untuk memahami apakah feature-feature yang berkorelasi tinggi tersebut memang memiliki signifikansi dalam konteks domain atau jika ada feature yang dapat dihilangkan.

Penerapan PCA bisa menjadi solusi untuk mengatasi multicollinearity dan mereduksi dimensi. Namun, sebelumnya, kami akan memastikan untuk melakukan standarisasi atau normalisasi pada data.

Kami akan juga menganalisa feature-feature dengan korelasi tinggi terhadap target ("Response") karena mungkin memiliki kontribusi yang signifikan pada model.

4. Business Insight



Dari grafik disamping, kita dapat mengambil beberapa insight yang dapat berguna dalam konteks bisnis:

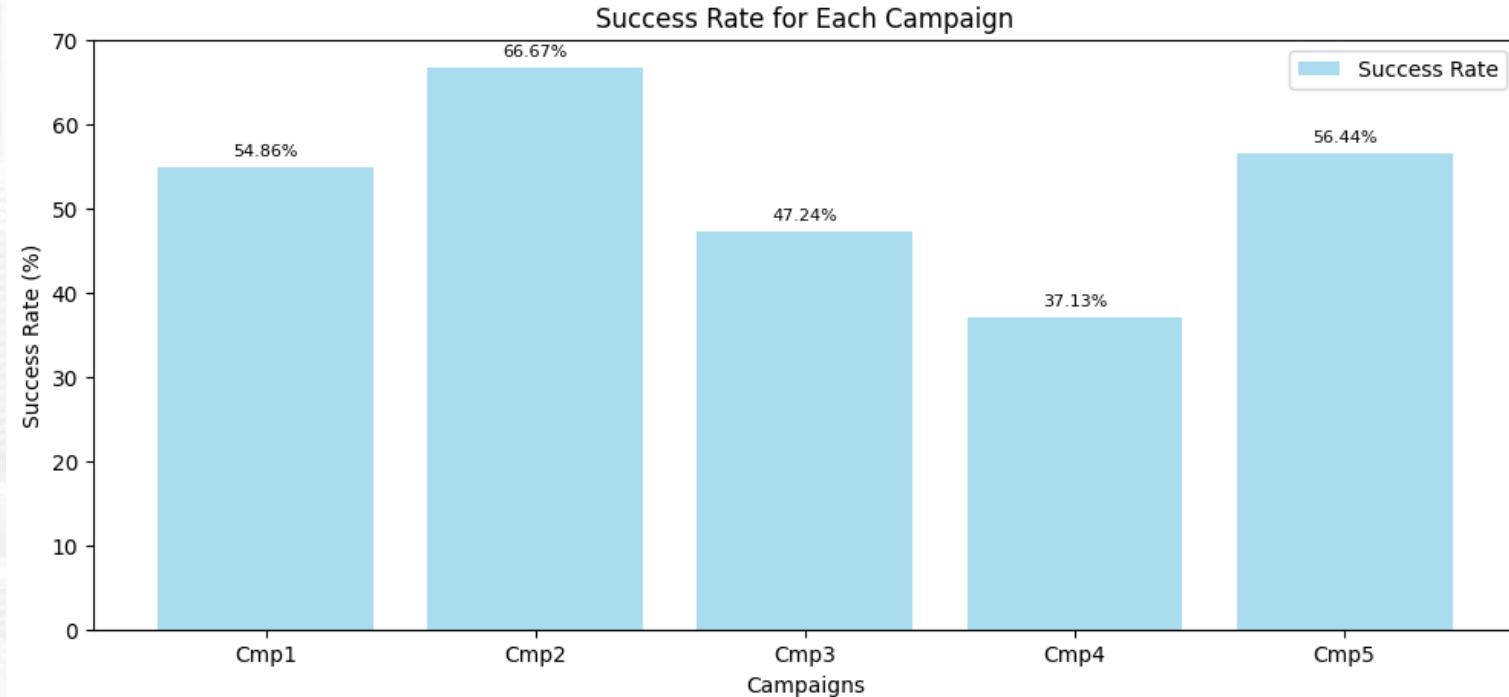
1. Pentingnya Campaign

Lebih dari setengah dari responden yang memberikan Yes Response telah menerima setidaknya satu Campaign (56.29%). Hal ini menunjukkan bahwa Campaign memainkan peran penting dalam meraih Response positif dari pelanggan.

2. Peluang Meningkatkan Kesuksesan Campaign

Terdapat potensi untuk meningkatkan kesuksesan Campaign karena masih ada sekitar 43.71% responden yang memberikan Yes Response tanpa menerima Campaign apa pun. Analisis lebih lanjut dapat dilakukan untuk memahami alasan di balik keputusan ini dan untuk memperbaiki strategi Campaign agar lebih menarik bagi pelanggan.

4. Business Insight



4. Business Insight

Dari grafik diatas, kita dapat mengambil beberapa insight sebagai berikut:

1. Efektivitas Pada Campaign Tertentu

Pemahaman tentang keberhasilan Campaign tertentu dapat membantu bisnis mengidentifikasi strategi yang efektif dan jenis Campaign yang lebih disukai oleh pelanggan. Misalnya, **Campaign 2 (Cmp2)** memiliki tingkat kesuksesan yang tinggi (**66.67%**), sehingga strategi dari Campaign 2 dapat dijadikan acuan untuk strategi Campaign di masa mendatang.

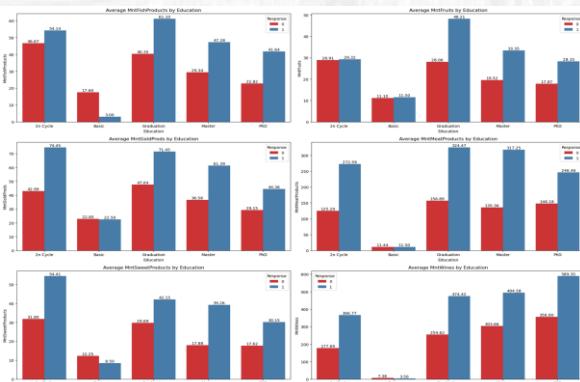
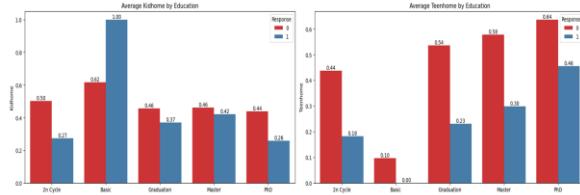
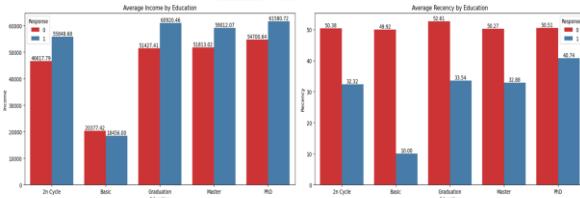
2. Potensi Peningkatan Respon Keseluruhan

Meskipun total Success Rate pada seluruh Campaign dan responden saat ini adalah 14.91%, peluang untuk meningkatkannya terbuka lebar. Analisis lebih lanjut, termasuk penargetan yang lebih baik, pesan Campaign yang lebih efektif, dan pemahaman lebih mendalam tentang preferensi pelanggan, dapat membantu meningkatkan tingkat respon keseluruhan.

3. Segmentasi Pelanggan:

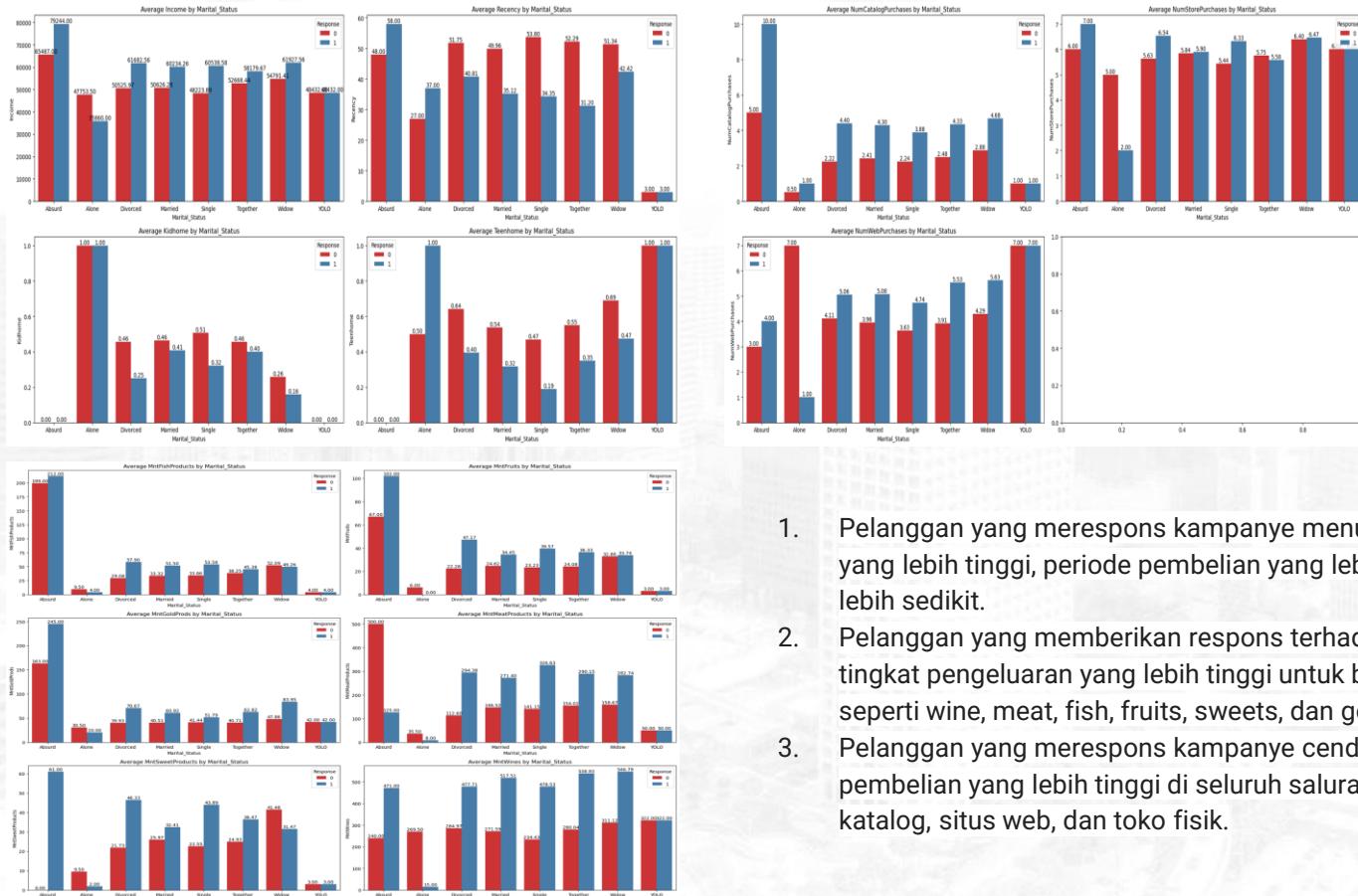
Penting untuk melakukan segmentasi pelanggan berdasarkan preferensi dan kebiasaan mereka terkait acceptance pada Campaign. Dengan memahami kelompok pelanggan yang merespon Campaign dengan baik dan kelompok yang tidak, untuk kedepannya dapat menyusun strategi yang lebih terarah dan personalisasi Campaign sesuai dengan setiap segmen.

4. Business Insight



1. Pelanggan yang memberikan respons terhadap kampanye memiliki rata-rata pendapatan yang lebih tinggi, periode pembelian yang lebih baru, dan jumlah anak yang lebih sedikit.
2. Pelanggan yang merespons kampanye menunjukkan tingkat pengeluaran yang lebih tinggi untuk berbagai kategori produk seperti wine, meat, fish, fruits, sweets, dan gold.
3. Pelanggan yang merespons kampanye cenderung memiliki rata-rata nilai pembelian yang lebih tinggi di seluruh saluran pembelian, termasuk melalui katalog, situs web, dan toko fisik.

4. Business Insight



- Pelanggan yang merespons kampanye menunjukkan rata-rata pendapatan yang lebih tinggi, periode pembelian yang lebih baru, dan jumlah anak yang lebih sedikit.
- Pelanggan yang memberikan respons terhadap kampanye menunjukkan tingkat pengeluaran yang lebih tinggi untuk berbagai kategori produk seperti wine, meat, fish, fruits, sweets, dan gold.
- Pelanggan yang merespons kampanye cenderung memiliki rata-rata nilai pembelian yang lebih tinggi di seluruh saluran pembelian, termasuk melalui katalog, situs web, dan toko fisik.

Summary - Business Insight

Berdasarkan analisis mendalam terhadap data Campaign dan karakteristik pelanggan, kami menyarankan beberapa langkah strategis untuk meningkatkan efektivitas Campaign dan memaksimalkan keuntungan bisnis:

1. Segmentasi Pelanggan Berdasarkan Respons Campaign:

Melakukan segmentasi pelanggan berdasarkan respons Campaign dapat membantu dalam menyesuaikan strategi pemasaran. Fokuskan upaya pada kelompok pelanggan yang telah menunjukkan respons positif, seperti tingkat pendidikan Graduation, PhD, dan Master, serta status pernikahan Single, Married, dan Divorced.

2. Personalisasi Pesan dan Penawaran:

Personalisasi pesan dan penawaran Campaign untuk setiap kelompok pelanggan yang telah diidentifikasi dapat meningkatkan keterlibatan. Berdasarkan karakteristik unik dari setiap kelompok, buatlah pesan yang relevan dan tawarkan insentif yang sesuai dengan preferensi mereka.

3. Penargetan Tingkat Pendidikan Tinggi:

Tingkat pendidikan tinggi seperti Graduation, PhD, dan Master memiliki potensi besar untuk respons Campaign. Fokuskan penawaran khusus, informasi produk, dan keuntungan tambahan pada kelompok ini untuk memaksimalkan partisipasi.

4. Optimalkan Pengeluaran Pelanggan yang Merespon:

Pelanggan yang merespon Campaign memiliki kecenderungan pengeluaran yang lebih tinggi pada berbagai kategori produk. Optimalisasi persediaan dan promosi pada produk-produk yang paling diminati oleh kelompok pelanggan ini dapat meningkatkan nilai transaksi.

5. Memperkuat Campaign dengan Data Pembelian dan Channel:

Analisis menunjukkan bahwa pelanggan yang merespon Campaign memiliki rata-rata pembelian yang lebih tinggi di berbagai saluran seperti catalog, web, dan toko fisik. Penguatan Campaign dengan peningkatan ketersediaan produk melalui saluran ini dapat meningkatkan aksesibilitas produk bagi pelanggan.

6. Monitoring secara Realtime dan Analisis Reaksi Pelanggan:

Melakukan monitoring secara real time terhadap respons pelanggan contohnya dengan menggunakan BI Tools untuk Dashboarding, dan melakukan analisis lebih lanjut terhadap perubahan tren dan preferensi. Keterlibatan yang berkelanjutan dan penyesuaian cepat terhadap dinamika pasar dapat menjadi kunci kesuksesan jangka panjang.

5. GIT

<https://github.com/hilmanman92/market-insider/tree/master>

market-insider Public

master had recent pushes 9 minutes ago

Compare & pull request

master 2 branches 0 tags

Go to file Add file <> Code

This branch is 3 commits ahead, 1 commit behind main.

Contribute

hilmanman92 penambahan summary pada Ddesc, Statistics dan summary EDA pada RE... 37019d 10 minutes ago 3 commits

README.md penambahan summary pada Ddesc, Statistics dan summary EDA pada R... 10 minutes ago

market insider - marketing campaign... penambahan summary pada desc, statistics, eda, dan business insights 19 minutes ago

marketing_campaign.csv add project to github 13 hours ago

README.md

Summary Descriptive Statistics

- Apakah ada kolom dengan tipe data kurang sesuai, atau nama kolom dan isinya kurang sesuai?
 - Feature Dt_Customer merupakan tanggal registrasi pelanggan dengan tipe data object. Tipe data ini tidak sesuai, sehingga perlu diubah menjadi tipe Date Time.
- Apakah ada kolom yang memiliki nilai kosong? Jika ada, apa saja?
 - Feature Income mempunyai nilai kosong karena hanya berjumlah 2216 rows.
- Apakah ada kolom yang memiliki nilai summary agak aneh? (min/mean/median/max/unique/top/freq)

About

Final Project of Market Insider

Readme Activity 0 stars 1 watching 0 forks

Releases

No releases published Create a new release

Packages

No packages published Publish your first package

Market Insider

Stage 2
Laporan Project



1. Data Pre-Processing

A. Data Splitting

```
● ● ●  
1 # split train and test set  
2  
3 # memisahkan antara training dan test set  
4 from sklearn.model_selection import train_test_split  
5 data_train, data_test = train_test_split(data, test_size=0.2, random_state=42)  
6  
7 # menampilkan shape dari train dan test set  
8 print(f'data_train: {data_train.shape}, data_test: {data_test.shape}')
```

```
data_train: (1792, 26), data_test: (448, 26)
```

Setelah dilakukan data split maka data kita telah memiliki partisi untuk test dan training dimana data_train 1792 sample dengan 26 feature dan data_test 448 sample dengan 26 feature juga.

1. Data Pre-Processing

B. Handling Missing Values

Pada data_train Terdapat 19 missing values (1.06%) dimensi sebelum (1792, 26) setelah Drop missing values – dimensi sesudah (1773, 26)

Pada data_test Terdapat 5 missing values (1.12%) dimensi sebelum (448, 26) setelah Drop missing values dimensi sesudah (443, 26)

C. Handling Duplicate Values

Pada data_train Terdapat 116 duplicate values (6.54%) dimensi sebelum (1773, 26) setelah Drop duplicate values – dimensi sesudah (1657,26)

Pada data_test Terdapat 6 duplicate values (1.35%) dimensi sebelum (443, 26) setelah Drop missing values dimensi sesudah (437, 26)

```
● ● ●  
Data Train Shape : (1792, 26)  
Data Test Shape : (448, 26)  
#1.Identification Missing Values  
    column missing values percentage  
    0 Income          19   1.06  
    column missing values percentage  
    0 Income          5    1.12  
#After Missing Value Handing  
Data Train Shape : (1773, 26)  
Data Test Shape : (443, 26)  
#2.Identification Duplicated Rows  
    duplicated rows percentage  
    0                      116  6.54  
    duplicated rows percentage  
    0                      6   1.35  
#After Duplicated Value Handing  
Data Train Shape : (1657, 26)  
Data Test Shape : (437, 26)  
( )
```

1. Data Pre-Processing

D. Handling Outliers

Menggunakan Z-Score threshold std = 3, pada column Income dan Year_Birth

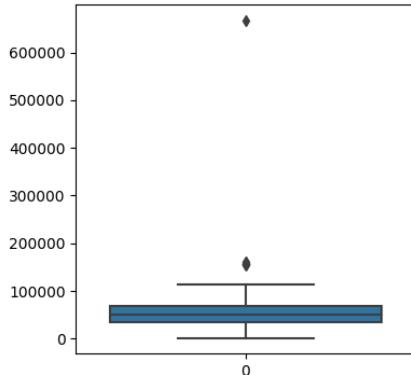
Column Income

Before	1657.00
After	1650.00
Outliers	7.00
% Outliers	0.42

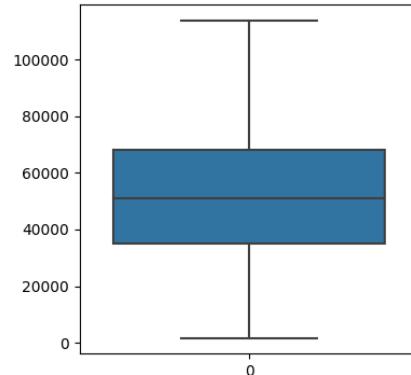
Column Year_Birth

Before	1650.00
After	1649.00
Outliers	1.00
% Outliers	0.06

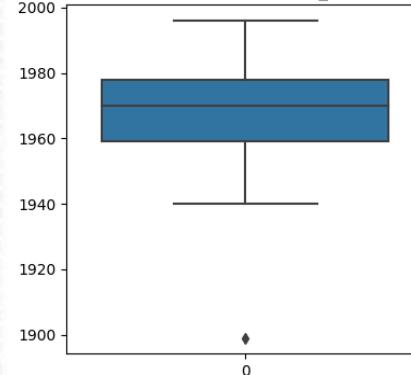
Before Outliers in Income



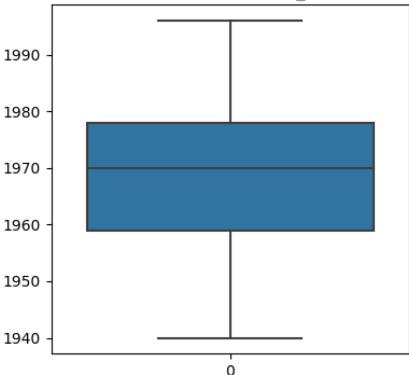
After Outliers in Income



Before Outliers in Year_Birth



After Outliers in Year_Birth



2. Feature Engineering

A. Feature Extractions

```
● ● ●  
1 # membuat feature baru berdasarkan status hubungan  
2 marital = {  
3     'Single': 'Not in relationship',  
4     'Together': 'In relationship',  
5     'Married': 'In relationship',  
6     'Divorced': 'Not in relationship',  
7     'Widow': 'Not in relationship',  
8     'Alone': 'Not in relationship',  
9     'Absurd': 'Not in relationship',  
10    'YOLO': 'Not in relationship'  
11 }  
12 data['Relationship_Status'] = data['Marital_Status'].map(marital)  
13  
14 # membuat feature baru total_children dari penjumlahan feature kidhome dan teenhome  
15 data['Total_Children'] = data['Kidhome'] + data['Teenhome']  
16  
17 # membuat feature baru berdasarkan jumlah anggota keluarga  
18 def fam_size(x):  
19     if x['Relationship_Status'] == 'Not in relationship':  
20         result = 1 + x['Teenhome'] + x['Kidhome']  
21     elif x['Relationship_Status'] == 'In relationship':  
22         result = 2 + x['Teenhome'] + x['Kidhome']  
23     return result  
24 data['Family_Size'] = data.apply(fam_size, axis=1)
```

```
● ● ●  
1 # membuat feature baru berdasarkan tanggal bergabung dan diasumsikan data dikumpulkan pada awal juli 2014  
2 data['Customer_Lifespan'] = (pd.to_datetime('2014-07-01') - data['Dt_Customer']).dt.days  
3  
4 # ekstraksi feature Datetime menjadi feature baru  
5 data['Year'] = data['Dt_Customer'].dt.year  
6 data.drop(['Dt_Customer'],axis=1,inplace=True)  
7  
8 # membuat feature baru total purchase, total spending, dan total offers  
9 data['Total_Purchase'] = data.apply(lambda x: x[purchase_cols[:-1]].sum(), axis=1)  
10 data['Total_Spending'] = data.apply(lambda x: x[spending_cols].sum(), axis=1)  
11 data['Total_Offers'] = data.apply(lambda x: x[campaign_cols[:-1]].sum(), axis=1)
```

Beberapa feature baru yang diextract dari feature-feature sebelumnya seperti Relationship_Status, Family_Size, Customer_Lifespan, Year, Total_Purchase, Total_Spending, Total_Offers, dan lainnya. (lihat full di source code)

2. Feature Engineering

A. Feature Extractions - RFM

```

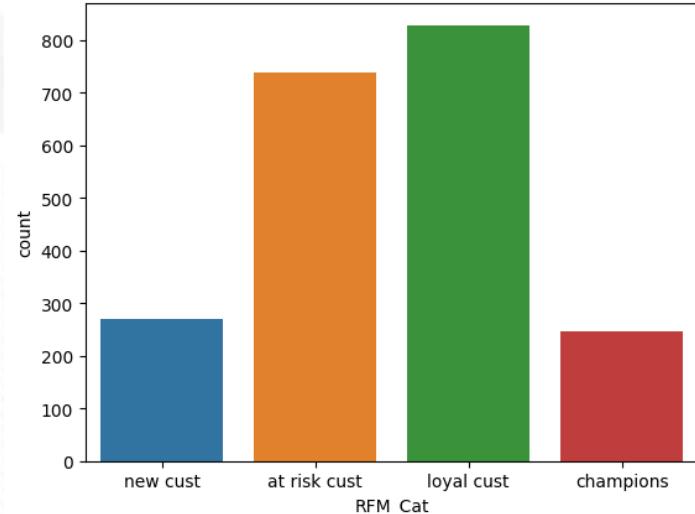
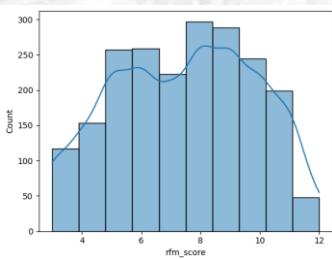
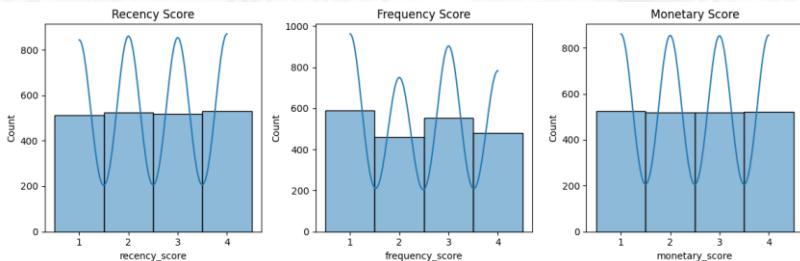
● ● ●

1 # membuat feature baru category rfm score (champions, loyal, at risk, new)
2 rfm = pd.DataFrame()
3 rfm['Recency'] = data['Recency']
4 rfm['Frequency'] = data['Total_Purchase']
5 rfm['Monetary'] = data['Total_Spending']
6 rfm.head()

● ● ●

1 # kalkulasi score berdasarkan quantile masing-masing feature
2 rfm['recency_score'] = pd.qcut(rfm['Recency'], q=[0, 0.25, 0.5, 0.75, 1], labels=[4, 3, 2, 1])
3 rfm['frequency_score'] = pd.qcut(rfm['Frequency'], q=[0, 0.25, 0.5, 0.75, 1], labels=[1, 2, 3, 4])
4 rfm['monetary_score'] = pd.qcut(rfm['Monetary'], q=[0, 0.25, 0.5, 0.75, 1], labels=[1, 2, 3, 4])

```



Melakukan segmentasi customer dengan menggunakan metode rfm. Customer dibagi menjadi 4 segment, yaitu champions, loyal, at risk, dan new customer. Berdasarkan count plot diketahui bahwa loyal customer paling banyak, champions adalah yang paling sedikit pada dataset. Lalu hasilnya digabungkan dengan dataset utama.

2. Feature Engineering

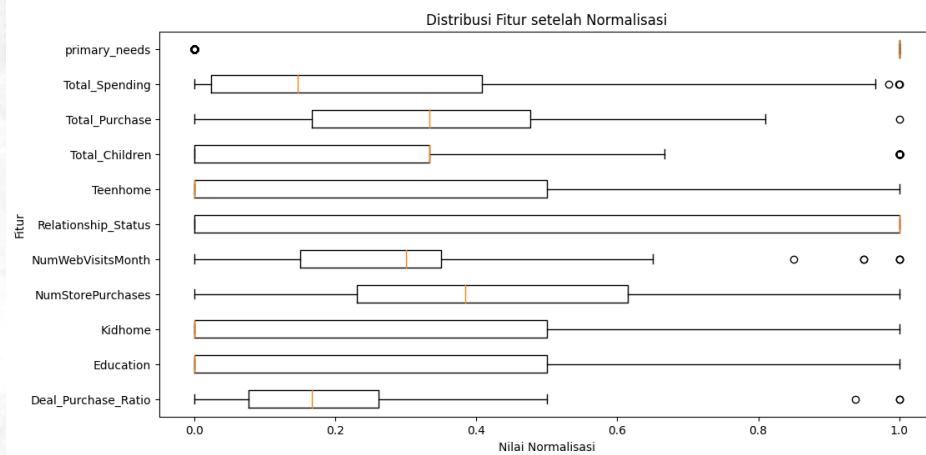
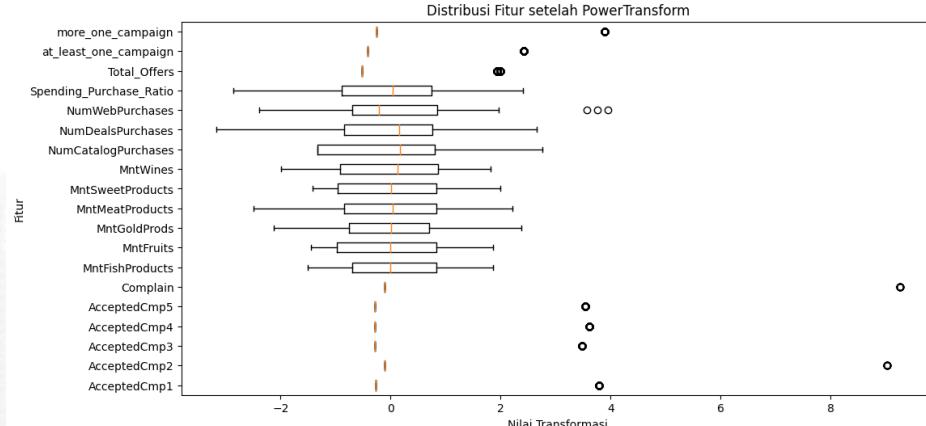
B. Feature Encoding

```
● ● ●  
  
# encoding education  
edu = {'Graduation': 1, 'Master': 2, 'PhD': 3}  
oe_edu = OrdinalEncoder(categories=[list(edu.keys())])  
data['Education'] = oe_edu.fit_transform(data[['Education']])  
  
# encoding marital_status  
marital_mapping = {'Single': 1, 'Married': 2, 'Divorced': 3}  
data['Marital_Status'] = data['Marital_Status'].map(marital_mapping)  
  
# encoding relationship_status  
rel_mapping = {'Not in relationship': 0,  
               'In relationship': 1}  
data['Relationship_Status'] = data['Relationship_Status'].map(  
    rel_mapping)  
  
# encoding primary_needs  
pr_mapping = {'primary_needs': 0,  
              'secondary_needs': 1}  
data['primary_needs'] = data['primary_needs'].map(pr_mapping)  
  
# encoding rfm_cat  
rfm_mapping = {'new cust': 1, 'at risk cust': 2,  
               'loyal cust': 3, 'champions': 4}  
oe_rfm = OrdinalEncoder(categories=[list(rfm_mapping.keys())])  
data['RFM_Cat'] = oe_rfm.fit_transform(data[['RFM_Cat']])
```

Menggunakan Ordinal encoder untuk mengubah categorical feature menjadi nilai ordinal. Beberapa feature yang dilakukan encode seperti education, marital_status, relationship_status, primary_needs, dan rfm_cat

2. Feature Engineering

C. Feature Transformation



Pada dataset ini melakukan 2 jenis transformasi yaitu normalisasi dan log transformasi/power transformasi.

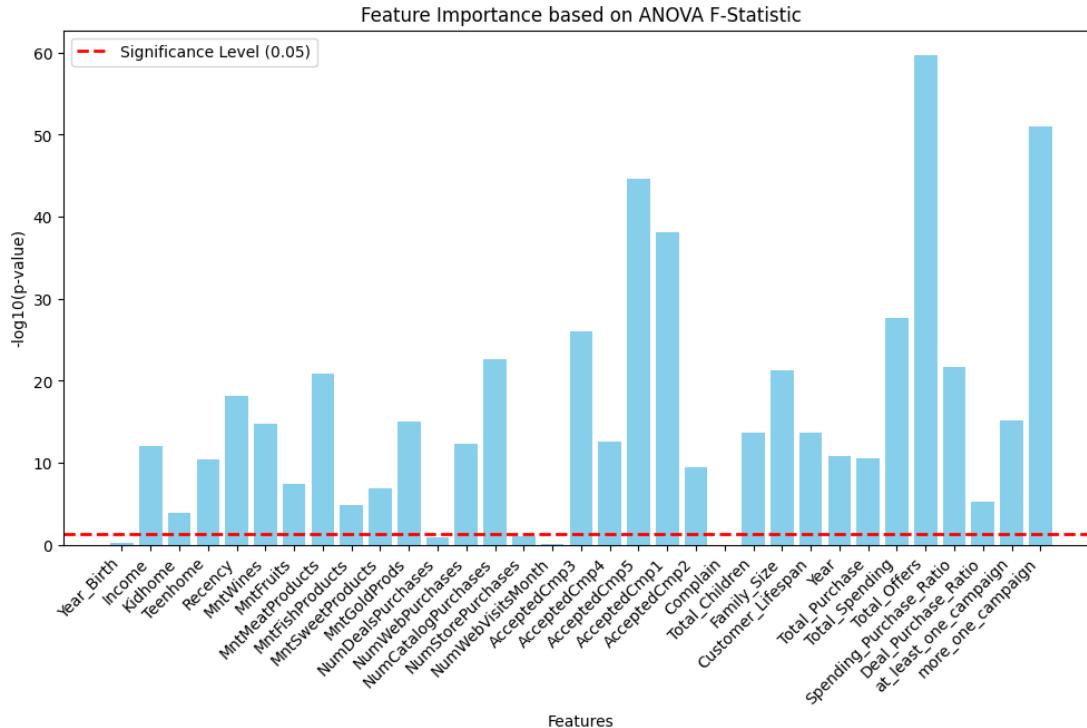
Normalisasi digunakan pada saat: $\text{mean} < \text{median} < \text{mode}$.

Transformasi digunakan pada saat: nilai $\text{skew_val} \leq -1$ atau $\text{skew_val} \geq 1$.

Library yang digunakan adalah MinMaxScaler untuk normalisasi dan PowerTransformer untuk yang transformasi log.

2. Feature Engineering

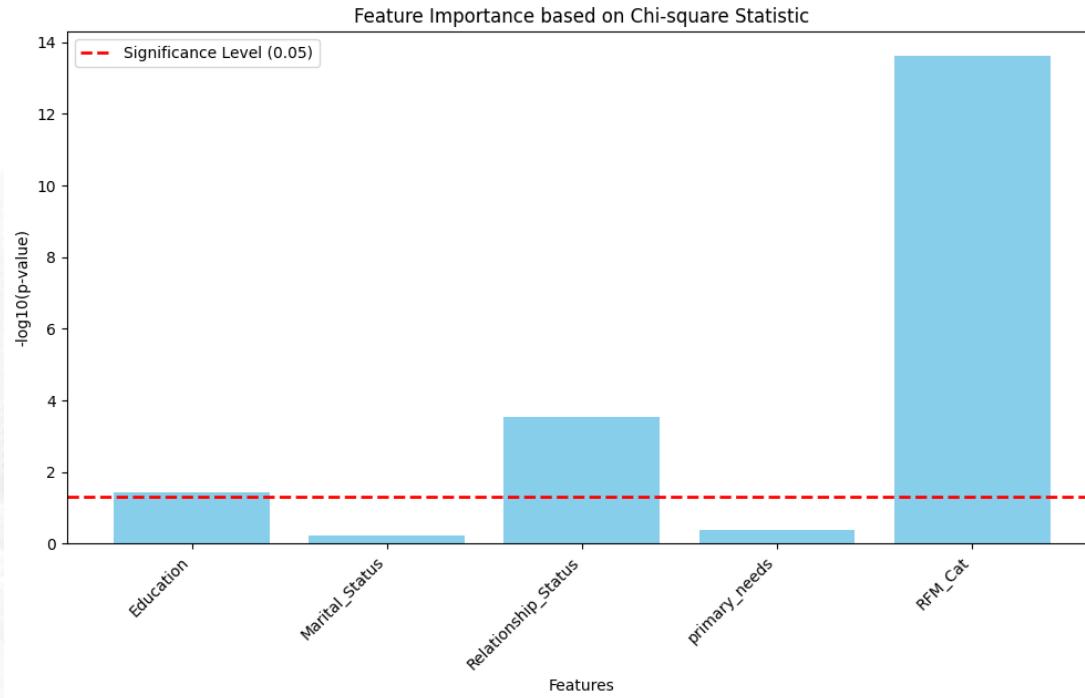
D. Feature Selection - ANOVA



Melakukan feature selection pada numerical feature dengan menggunakan metode anova, dan hasilnya didapatkan bahwa Year Birth, Num Deals Purchase, Num Web Visits Month, Num Store Purchases, Complain, tidak mampu melewati threshold p value < 0.05.

2. Feature Engineering

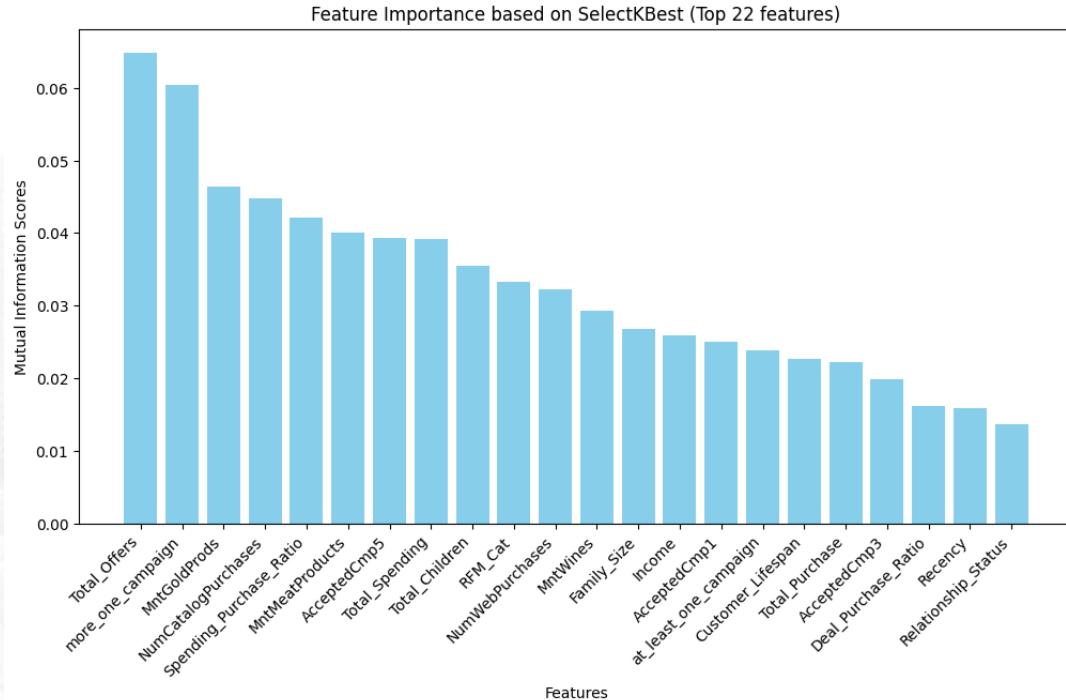
D. Feature Selection - Chi Square



Melakukan feature selection pada categorical feature dengan menggunakan metode chi-square, dan hasilnya didapatkan bahwa Marital Status dan primary needs tidak mampu melewati threshold p value < 0.05.

2. Feature Engineering

D. Feature Selection - Mutual Info Classif



Setelah menggabungkan feature selection dari metode anova dan chi2, selanjutnya menggunakan kbest dan mutual info classification. Dan didapatkan jumlah k yang terbaik adalah 22 feature. Dimana total offers merupakan feature dengan score tertinggi, sedangkan relationship status adalah yang terendah. Selanjutkan melakukan pengecekan multicollinearity menggunakan metode VIF.

2. Feature Engineering

D. Feature Selection - VIF (Redundancy Analysis)

Avg. VIF Score: 490540.08822727663

	Feature	VIF
0	Income	6.072868e+00
1	Kidhome	3.017963e+00
2	MntWines	1.288884e+01
3	MntMeatProducts	1.041846e+01
4	MntGoldProds	2.594825e+00
5	NumWebPurchases	3.143823e+00
6	NumCatalogPurchases	5.211826e+00
7	AcceptedCmp3	1.266840e+01
8	AcceptedCmp4	1.280350e+01
9	AcceptedCmp5	1.110481e+01
10	AcceptedCmp1	9.829085e+00
11	Total_Children	1.088710e+02
12	Family_Size	5.526401e+02
13	Customer_Lifespan	1.217927e+00
14	Total_Spending	9.014238e+00
15	Total_Offers	5.126044e+06
16	Spending_Purchase_Ratio	2.008915e+01
17	Deal_Purchase_Ratio	3.564921e+00
18	at_least_one_campaign	3.810943e+06
19	more_one_campaign	1.854052e+06
20	Relationship_Status	5.471427e+01
21	RFM_Cat	2.922084e+00

Avg. VIF Score: 3.4913412803840322

	Feature	VIF
0	Recency	3.042867
1	MntWines	5.538016
2	MntMeatProducts	5.603630
3	MntGoldProds	2.004252
4	NumWebPurchases	2.505467
5	NumCatalogPurchases	5.007321
6	AcceptedCmp3	1.842999
7	AcceptedCmp5	1.743514
8	AcceptedCmp1	1.526316
9	Total_Children	3.785204
10	Customer_Lifespan	4.307159
11	Total_Offers	3.193541
12	Relationship_Status	2.818932
13	RFM_Cat	5.959560

Avg. VIF Score: 3.568806872225212

	Feature	VIF
0	Recency	3.054826
1	MntWines	5.980497
2	MntMeatProducts	6.564739
3	MntGoldProds	2.090503
4	NumWebPurchases	2.521329
5	NumCatalogPurchases	5.060327
6	AcceptedCmp3	2.870500
7	AcceptedCmp5	1.830245
8	AcceptedCmp1	1.622926
9	Total_Children	3.900620
10	Customer_Lifespan	4.309492
11	Total_Offers	6.041355
12	Relationship_Status	2.825440
13	RFM_Cat	6.032342
14	AcceptedCmp2	1.124760
15	AcceptedCmp4	2.481471
16	MntFruits	2.358346

Pada Metode VIF, kami melakukan secara manual untuk mengurangi atau menambah feature dimana setiap feature tidak boleh melebihi threshold yakni >10 . Dan mencari nilai rata-rata optimal 2-5. Maka didapatkan 17 features, yang diantaranya terdapat features yang dihasilkan dari extraction seperti RFM_Cat, Total_Offers, Relationship_Status, dan Total_Children.

2. Feature Engineering

E. Feature Imbalance

```
● ● ●  
1 # melakukan imbalance handling pada target  
2 from imblearn.over_sampling import SMOTE  
3  
4 # menampilkan jumlah kelas sebelum oversampling  
5 print("Jumlah kelas sebelum oversampling:")  
6 print("Kelas 0:", sum(y_train == 0))  
7 print("Kelas 1:", sum(y_train == 1))  
8  
9 # melakukan oversampling dengan SMOTE  
10 smote = SMOTE(random_state=42)  
11 X_resampled, y_resampled = smote.fit_resample(X_train, y_train)  
12  
13 # menampilkan jumlah kelas setelah oversampling  
14 print("\nJumlah kelas setelah oversampling:")  
15 print("Kelas 0:", sum(y_resampled == 0))  
16 print("Kelas 1:", sum(y_resampled == 1))
```

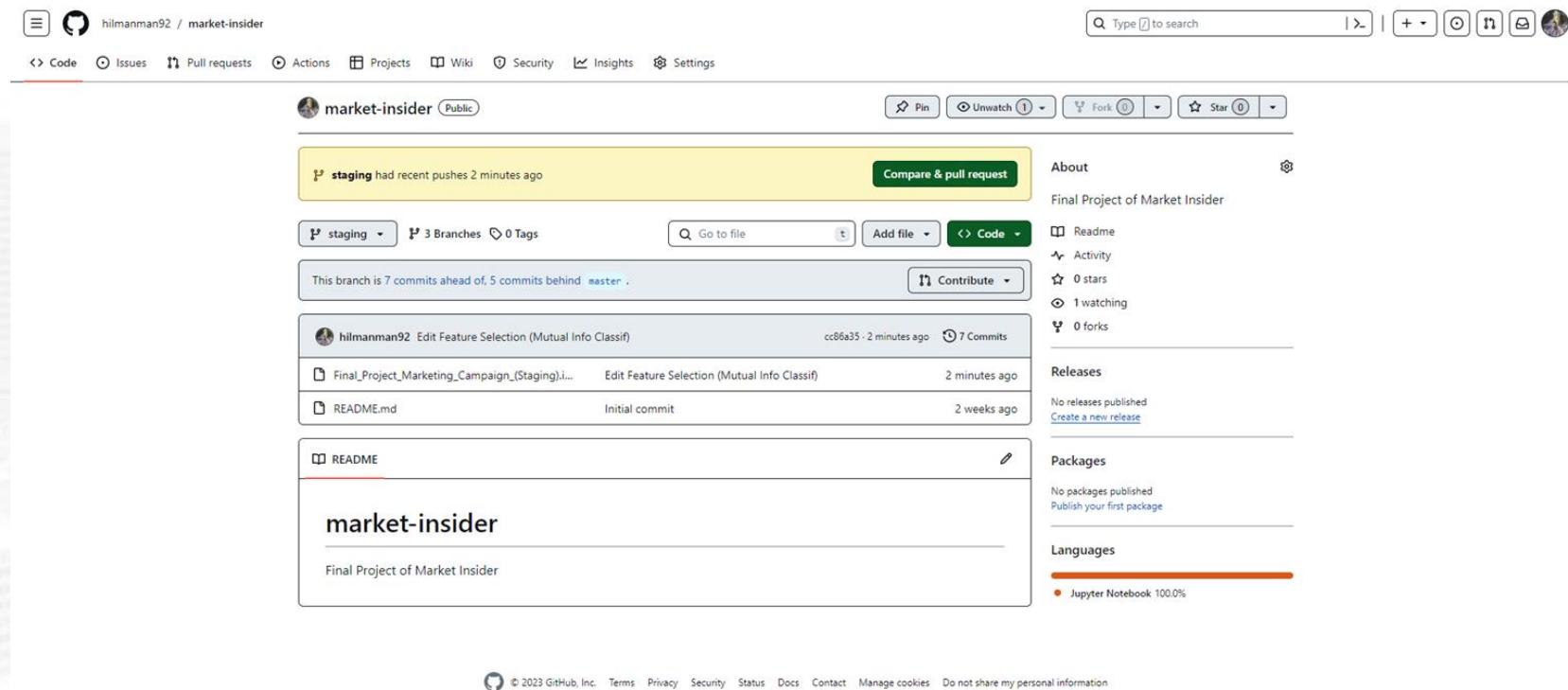
Menggunakan metode oversampling pada library SMOTE. Dimana feature target dengan values 1, awalnya berjumlah 251 menjadi 1397.

```
Jumlah kelas sebelum oversampling:  
Kelas 0: 1397  
Kelas 1: 251
```

```
Jumlah kelas setelah oversampling:  
Kelas 0: 1397  
Kelas 1: 1397
```

5. GIT

<https://github.com/hilmanman92/market-insider/tree/staging>



The screenshot shows a GitHub repository page for the project "market-insider". The repository is public and has a single branch named "staging". The "staging" branch is ahead of the "master" branch by 7 commits and behind it by 5 commits. The repository was last updated 2 minutes ago. The README file contains the text "market-insider" and "Final Project of Market Insider". The repository has 0 stars, 1 watching, and 0 forks. It has no releases published.

market-insider Public

staging had recent pushes 2 minutes ago

Compare & pull request

staging 3 Branches 0 Tags

Go to file Add file Code

This branch is 7 commits ahead of master . Contribute

hilmanman92 Edit Feature Selection (Mutual Info Classif) cc86a35 - 2 minutes ago 7 Commits

Final_Project_Marketing_Campaign_(Staging).ipynb Edit Feature Selection (Mutual Info Classif) 2 minutes ago

README.md Initial commit 2 weeks ago

README

market-insider

Final Project of Market Insider

About

Final Project of Market Insider

Readme Activity 0 stars 1 watching 0 forks

Releases

No releases published Create a new release

Packages

No packages published Publish your first package

Languages

Jupyter Notebook 100.0%

© 2023 GitHub, Inc. Terms Privacy Security Status Docs Contact Manage cookies Do not share my personal information

Market Insider

Stage 3
Laporan Project



1. Modelling

A. Data Splitting

```
# mendefinisikan X (features) dan y (target), dimana X dan y hasil dari imbalance handling sebelumnya
X_train = X_resampled
y_train = y_resampled
```

Data telah di split antara data training dan data testing di tahap Pre-Processing. Sebelum melakukan pemodelan, akan didefinisikan kembali data train yang sudah dilakukan imbalance handling yaitu oversampling dengan teknik SMOTE dikarenakan ingin meningkatkan jumlah kuantitas suatu label.

1. Modelling

B. Modelling

Support Vector Machine (SVM), Adaboost, dan XGBoost akan menjadi algoritma yang digunakan. Pemilihan ketiga model ini disebabkan oleh rendahnya tingkat kesalahan dan tingginya tingkat presisi yang telah teruji berdasarkan percobaan dengan 8 model.

	CV_Precision	Precision_Train	Precision_Test	Diff	Diff (%)
LogisticRegression	0.802285	0.805656	0.844000	-0.038343	-4.759285
KNeighborsClassifier	0.753420	0.812624	0.802228	0.010396	1.279259
DecisionTreeClassifier	0.850591	0.996430	0.818996	0.177434	17.806926
SVC	0.799256	0.856226	0.836876	0.019350	2.259904
RandomForestClassifier	0.665800	0.712551	0.780197	-0.067646	-9.493512
GaussianNB	0.810698	0.738874	0.802251	-0.063377	-8.577558
XGBClassifier	0.847233	0.864785	0.827637	0.037148	4.295622
AdaBoostClassifier	0.872463	0.879967	0.822993	0.056974	6.474576

1. Modelling

C. Model Evaluation

Metrics Evaluation yang digunakan adalah Precision dan Accuracy. Precision digunakan untuk meminimalkan False Positive karena pada kasus ini, kami perlu melakukan cost efficiency pada budget marketing. Setelah itu, kami juga mempertimbangkan nilai Accuracy sebagai parameter sekunder pada dua model yang menghasilkan model terbaik.

Berdasarkan hasil pemodelan dari ketiga model dibawah ini, model Adaboost menunjukkan nilai Precision dan Accuracy yang lebih tinggi dibandingkan dengan dua model lainnya. Namun, perbedaan nilai prediksi (diff) model Adaboost pada data training dan data test memiliki selisih yang lebih besar dibandingkan dengan dua model lainnya. Dengan demikian, model Adaboost memiliki performa yang lebih baik dalam mengklasifikasikan data dengan akurasi dan ketepatan yang lebih tinggi, tetapi memiliki perbedaan yang lebih besar dalam memprediksi data dibandingkan dengan dua model lainnya. Nilai diff tersebut akan diperkecil melalui hyperparameter tuning.

	CV Precision	Precision_Train	Precision_Test	Diff	Diff (%)
SVC	0.799378	0.856226	0.836876	0.019350	2.259904
XGBClassifier	0.847233	0.864785	0.827637	0.037148	4.295622
AdaBoostClassifier	0.872463	0.879967	0.822993	0.056974	6.474576
	CV Accuracy	Accuracy_Train	Accuracy_Test	Diff	Diff (%)
SVC	0.832502	0.852899	0.743119	0.109780	12.871371
XGBClassifier	0.840023	0.864710	0.807339	0.057371	6.634668
AdaBoostClassifier	0.861142	0.879742	0.816514	0.063229	7.187166

1. Modelling

D. Hyperparameter Tuning

Untuk menemukan nilai optimal untuk parameter yang digunakan pada model, akan dilakukan hyperparameter tuning. Dengan melakukan langkah ini, diharapkan dapat meningkatkan performa model dan menghasilkan model yang best-fit. Metode yang digunakan adalah Grid Search pada masing-masing model yang sudah dipilih.

```
# Define grid search
grid = dict(learning_rate=learning_rate, n_estimators=n_estimators, algorithm=algorithm)
grid_search = GridSearchCV(estimator=model, param_grid=grid, n_jobs=-1, cv=5, scoring='precision', error_score=0, return_
grid_result = grid_search.fit(X_train, y_train)
```

Metode Grid search dipilih karena Grid Search dapat mengeksplorasi seluruh kombinasi hyperparameter yang telah ditentukan dalam parameter yang didefinisikan sebelumnya. Ini memastikan bahwa kita tidak akan melewatkkan setiap kemungkinan konfigurasi hyperparameter yang mungkin menghasilkan model yang optimal.

1. Modelling

E. Pemilihan model terbaik

Setelah melakukan hyperparameter tuning, berikut ini adalah hasil metrics evaluation dari SVM (CV Precision > 0.8 dan Diff < 10%). Model ini menghasilkan nilai Precision dan Accuracy lebih rendah dibanding model lainnya.

Precision SVM

Model	CV Precision	Precision_Train	Precision_Test	Diff	Diff (%)	Parameters
Parameter Set						
param_16	SVC	0.874959	0.844516	0.807559	0.036957	4.376101
param_13	SVC	0.874557	0.840083	0.810913	0.029170	3.472247
param_10	SVC	0.873749	0.837537	0.814110	0.023427	2.797151

Accuracy SVM

Model	CV Accuracy	Accuracy_Train	Accuracy_Test	Diff	Diff (%)	Parameters
Parameter Set						
param_5	SVC	0.834291	0.854331	0.740826	0.113505	13.285841
param_2	SVC	0.832502	0.852899	0.743119	0.109780	12.871371
param_28	SVC	0.817103	0.849320	0.802752	0.046568	5.482937

1. Modelling

E. Pemilihan model terbaik

Setelah melakukan hyperparameter tuning pada model XGBoost, berikut ini adalah hasil metrics evaluationnya (CV Precision > 0.8 dan Diff < 10%). Untuk model ini, kami melakukan perhitungan dengan max depth=1 untuk menghindari overfit.

Precision XGBoost

Parameter Set	Model	CV Precision	Precision_Train	Precision_Test	Diff	Diff (%)	Parameters
param_140	XGBClassifier	0.912467	0.922357	0.818055	0.104302	11.308190	{'colsample_bytree': 0.8, 'gamma': 1, 'learnin...
param_143	XGBClassifier	0.912467	0.922357	0.818055	0.104302	11.308190	{'colsample_bytree': 0.8, 'gamma': 1, 'learnin...
param_1147	XGBClassifier	0.912229	0.924031	0.835093	0.088938	9.625034	{'colsample_bytree': 1.0, 'gamma': 2, 'learnin...

Accuracy XGBoost

Parameter Set	Model	CV Accuracy	Accuracy_Train	Accuracy_Test	Diff	Diff (%)	Parameters
param_1003	XGBClassifier	0.879048	0.924123	0.850917	0.073206	7.921638	{'colsample_bytree': 1.0, 'gamma': 1, 'learnin...
param_1006	XGBClassifier	0.879048	0.924123	0.850917	0.073206	7.921638	{'colsample_bytree': 1.0, 'gamma': 1, 'learnin...
param_247	XGBClassifier	0.878325	0.918754	0.837156	0.081599	8.881427	{'colsample_bytree': 0.8, 'gamma': 2, 'learnin...

1. Modelling

E. Pemilihan model terbaik

Setelah melakukan hyperparameter tuning pada model Adaboost, berikut ini adalah hasil metrics evaluationnya (CV Precision > 0.8 dan Diff < 10%).

Precision Adaboost

	Model	CV Precision	Precision_Train	Precision_Test	Diff	Diff (%)	Parameters
Parameter Set							
param_56	AdaBoostClassifier	0.908271	0.918675	0.843316	0.075359	8.203003	{"algorithm": "SAMME.R", "learning_rate": 1.0,...}
param_55	AdaBoostClassifier	0.903253	0.910859	0.831003	0.079856	8.767125	{"algorithm": "SAMME.R", "learning_rate": 1.0,...}
param_52	AdaBoostClassifier	0.900182	0.912634	0.822355	0.090279	9.892101	{"algorithm": "SAMME.R", "learning_rate": 0.8,...}

Accuracy Adaboost

	Model	CV Accuracy	Accuracy_Train	Accuracy_Test	Diff	Diff (%)	Parameters
Parameter Set							
param_55	AdaBoostClassifier	0.872244	0.910523	0.834862	0.075660	8.309532	{"algorithm": "SAMME.R", "learning_rate": 1.0,...}
param_56	AdaBoostClassifier	0.871889	0.918397	0.850917	0.067479	7.347494	{"algorithm": "SAMME.R", "learning_rate": 1.0,...}
param_52	AdaBoostClassifier	0.870808	0.912312	0.830275	0.082037	8.992193	{"algorithm": "SAMME.R", "learning_rate": 0.8,...}

1. Modelling

E. Pemilihan model terbaik

Dapat dilihat pada gambar dibawah, bahwa model XGB memiliki tingkat presisi yang paling tinggi dibandingkan SVC dan AdaBoost. Namun, XGB memiliki Diff yang paling tinggi sehingga potensi overfit nya paling tinggi. Sehingga, kami memutuskan untuk memilih **Adaboost Classifier dengan param_56** karena model ini menunjukkan nilai Precision dan Accuracy yang relatif moderate (tidak paling rendah presisinya dan tidak paling tinggi Diff-nya).

	Model	CV	Precision	Precision_Train	Precision_Test	Diff	Diff (%)	Parameters
Parameter Set								
param_16	SVC	0.874959		0.844516	0.807559	0.036957	4.376101	{'C': 3, 'gamma': 'auto', 'kernel': 'poly'}
param_13	SVC	0.874557		0.840083	0.810913	0.029170	3.472247	{'C': 3, 'gamma': 'scale', 'kernel': 'poly'}
param_10	SVC	0.873749		0.837537	0.814110	0.023427	2.797151	{'C': 2, 'gamma': 'auto', 'kernel': 'poly'}

	Model	CV	Precision	Precision_Train	Precision_Test	Diff	Diff (%)	Parameters
Parameter Set								
param_140	XGBClassifier	0.912467		0.922357	0.818055	0.104302	11.308190	{'colsample_bytree': 0.8, 'gamma': 1, 'learnin...}
param_143	XGBClassifier	0.912467		0.922357	0.818055	0.104302	11.308190	{'colsample_bytree': 0.8, 'gamma': 1, 'learnin...}
param_1147	XGBClassifier	0.912229		0.924031	0.835093	0.088938	9.625034	{'colsample_bytree': 1.0, 'gamma': 2, 'learnin...}

	Model	CV	Precision	Precision_Train	Precision_Test	Diff	Diff (%)	Parameters
Parameter Set								
param_56	AdaBoostClassifier	0.908271		0.918675	0.843316	0.075359	8.203003	{'algorithm': 'SAMME.R', 'learning_rate': 1.0,...}
param_55	AdaBoostClassifier	0.903253		0.910859	0.831003	0.079856	8.767125	{'algorithm': 'SAMME.R', 'learning_rate': 1.0,...}
param_52	AdaBoostClassifier	0.900182		0.912634	0.822355	0.090279	9.892101	{'algorithm': 'SAMME.R', 'learning_rate': 0.8,...}

1. Modelling

E. Pemilihan model terbaik

Setelah melakukan hyperparameter tuning pada model AdaBoost, didapatkan hasil presisi yang lebih baik dibandingkan sebelum hyperparameter tuning. Dapat dilihat bahwa terjadi kenaikan presisi sebesar 3.6% dari 87.2% menjadi 90.8%. Kenaikan tersebut diimbangi oleh suatu tradeoff yaitu peningkatan yang terjadi pada nilai selisih antara nilai presisi pada data train dan data test yaitu sekitar 1.7% dari 6.5% menjadi 8.2%. Selisih tersebut masih dapat ditoleransi karena masih dibawah 10% untuk dikategorikan sebagai best-fit.

Sebelum tuning	Model	CV Precision	Precision_Train	Precision_Test	Diff	
					Diff	Diff (%)
	AdaBoostClassifier	0.872463	0.879967	0.822993	0.056974	6.474576
+1.7% Diff	+3.6% Precision					
Setelah tuning	Parameter Set	Model	CV Precision	Precision_Train	Precision_Test	Diff
	param_56	AdaBoostClassifier	0.908271	0.918675	0.843316	0.075359 8.203003

Berikut pertimbangan dalam pemilihan model ini:

1. Modelling

E. Pemilihan model terbaik

CV Precision yang Tinggi:

param_56 - CV Precision: 0.908271: Model ini memiliki nilai CV Precision yang tinggi selama proses cross validation. Ini menunjukkan bahwa model memiliki kemampuan yang baik untuk memberikan prediksi positif yang benar pada data yang belum pernah dilihat sebelumnya.

Perbedaan (Diff) yang Relatif Kecil:

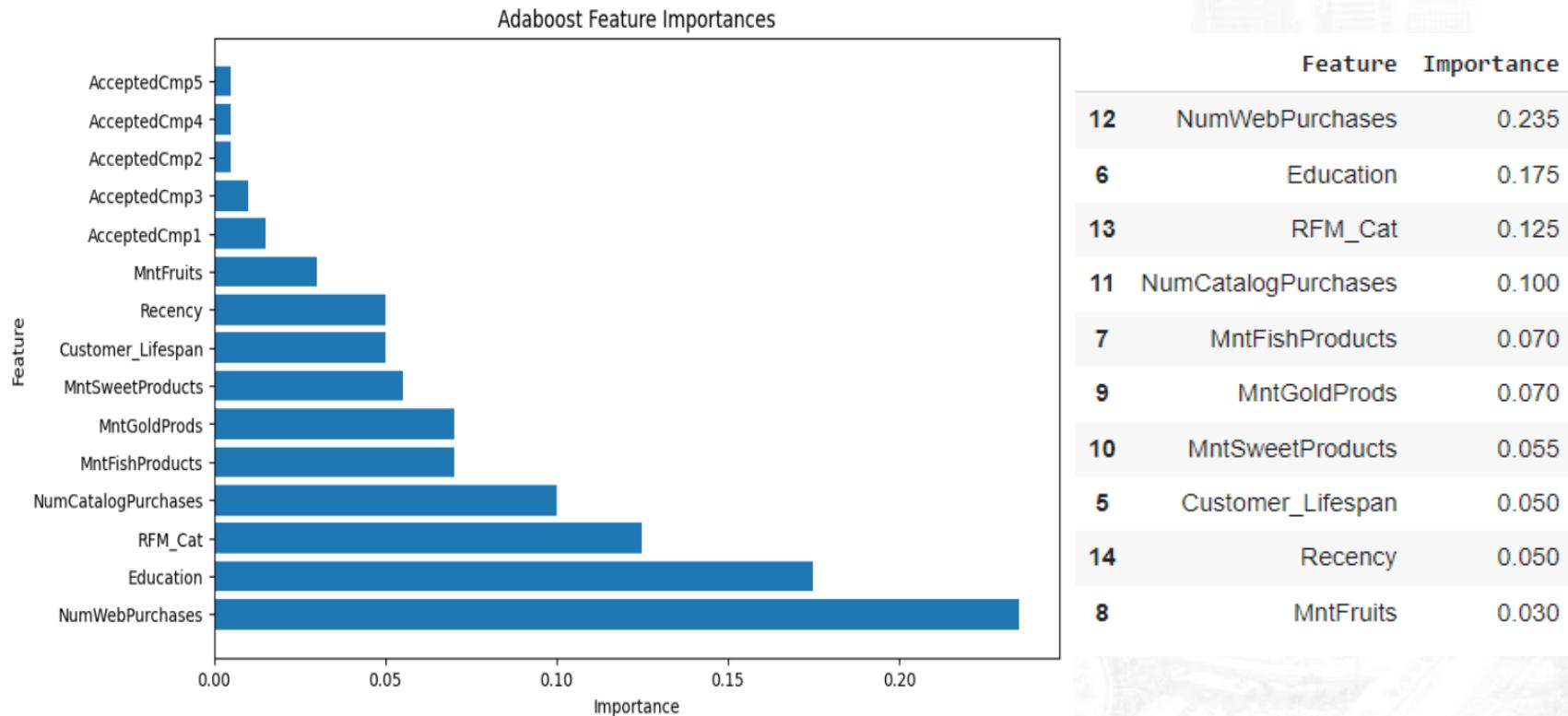
param_56 - Diff: 0.075359 (8.2%): Meskipun terdapat perbedaan antara Precision pada data train dan data test, perbedaannya relatif kecil (8.2%). Hal ini menunjukkan bahwa model ini cenderung tidak overfitting secara signifikan pada data train dan masih dapat menggeneralisasi dengan baik pada data test.

Precision_Train dan Precision_Test yang Tinggi:

Model ini juga memiliki nilai Precision yang tinggi pada data Train dan data test, menunjukkan kemampuan baik pada keduanya. Dengan mempertimbangkan kombinasi CV Precision yang tinggi, perbedaan yang relatif kecil antara data train dan test, serta nilai Precision yang tinggi pada kedua dataset.

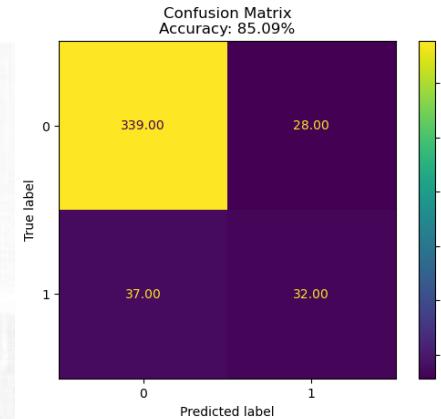
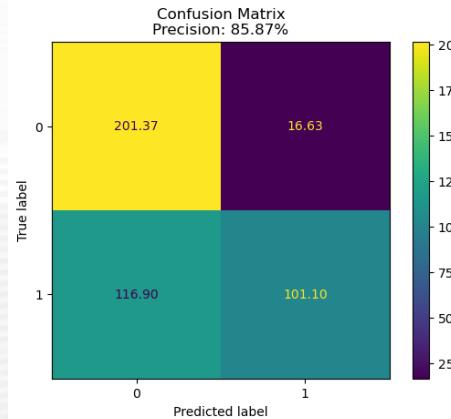
2. Feature Importance

A. Grafik hasil feature importance dari model Adaboost Classifier



2. Feature Importance

A. Confusion matrix dari model Adaboost Classifier



Classification Report on Test Data				
	precision	recall	f1-score	support
0	0.90	0.92	0.91	367
1	0.53	0.46	0.50	69
accuracy			0.85	436
macro avg	0.72	0.69	0.70	436
weighted avg	0.84	0.85	0.85	436

Classification Report on Train Data				
	precision	recall	f1-score	support
0	0.91	0.93	0.92	1397
1	0.93	0.91	0.92	1397
accuracy			0.92	2794
macro avg	0.92	0.92	0.92	2794
weighted avg	0.92	0.92	0.92	2794

Perhitungan confusion matrix pada precision menggunakan average weighted, karena jumlah label pada data test tidak seimbang.

2. Feature Importance

A. Prediksi Peningkatan Response Rate dan ROI Rate

total customer: 2240
total response: 334
rate accept: 14.91%
total cost: 2240
total revenue: 1670
total profit: -570
roi rate before pemodelan: -34.13%

total customer: 118
total response: 101
rate accept: 85.59%
total cost: 118
total revenue: 505
total profit: 387
roi rate after pemodelan: 76.63%

Setelah melalui proses pemodelan dan evaluasi, maka didapatkan hasil adanya peningkatan response rate dari 14.91% menjadi 85.59%.

Menambahkan asumsi untuk cost per customer/campaign adalah \$1 dan revenue per customer/campaign adalah \$5, maka didapatkan peningkatan roi rate dari -34.13% menjadi 76.63%.

2. Feature Importance

B. Business Insight & Recommendation

Dari hasil feature importance, berikut adalah insight yang didapat dan rekomendasi untuk meningkatkan efektivitas campaign dan memaksimalkan keuntungan bisnis:

a. Number Purchase on Website:

Business Insight: Penjualan melalui website merupakan faktor terpenting dalam marketing campaign.

Business Recommendation: Tingkatkan fokus pada pengoptimalan website, meningkatkan pengalaman pembelian online, gunakan data segmentasi pelanggan untuk menawarkan insentif atau penawaran khusus melalui platform web, dan perkuat strategi pemasaran online, untuk menawarkan insentif atau penawaran khusus sesuai dengan preferensi masing-masing kelompok pelanggan.

a. Education:

Business Insight: Tingkat pendidikan Phd adalah tingkat pendidikan dengan response rate tertinggi.

Business Recommendation: Pertimbangkan untuk mengadakan kolaborasi atau acara yang menghadirkan konten berkualitas tinggi, yang dapat menarik perhatian pelanggan Ph.D. dan memberikan nilai tambah dalam konteks pendidikan atau pengetahuan.

a. RFM_Cat:

Business Insight: Kategori RFM memainkan peran penting dalam respons pelanggan. Champions dan Loyal Customer mempunyai response rate tertinggi.

Business Recommendation: Tingkatkan layanan pelanggan untuk pelanggan dalam kategori Champions dan Loyal Customer, Kembangkan inovasi produk yang dapat memenuhi harapan tinggi dari pelanggan dalam kategori ini, Tawarkan produk atau layanan tambahan yang relevan dengan preferensi mereka.

2. Feature Importance

B. Business Insight & Recommendation

d. Number Purchase on Catalog:

Business Insight: Pembelian melalui katalog juga memiliki dampak yang signifikan

Business Recommendation: Pastikan bahwa produk yang ditampilkan di katalog sesuai dengan kebutuhan dan preferensi target pelanggan, sediakan penawaran eksklusif atau diskon khusus yang hanya dapat diakses melalui pembelian melalui katalog, dan pastikan bahwa inventaris katalog selalu diperbarui dan mencerminkan stok aktual.

e. Amount Spent on Fish, Gold, Sweet, dan Fruit in Last 2 Years:

Berdasarkan tingkat pendidikan customer, customer yang merespon campaign cenderung memiliki pengeluaran yang lebih besar baik pada produk ikan, buah, emas, dan sweets (permen dan coklat).

Business Recommendation: Buat penawaran bundle atau paket khusus dan berikan diskon bagi pelanggan yang membeli paket ini. Sediakan informasi yang mendalam tentang kualitas dan sumber produk.

f. Customer Lifespan & Recency:

Business Insight: Rata-rata Customer lifespan berdasarkan tingkat pendidikan maupun segmentasi pelanggan yang lebih tinggi (> 400 hari) dan rata-rata recency berdasarkan tingkat pendidikan yang lebih rendah (35 hari) cenderung merespon marketing campaign.

Business Recommendation: Fokuskan kampanye pemasaran pada pelanggan dengan recency rendah, perbarui program loyalitas untuk pelanggan dengan customer lifespan tinggi, sertakan kampanye edukasi yang memberikan nilai tambah bagi pelanggan dengan tingkat pendidikan tinggi, tawarkan konten yang berfokus pada kecerdasan kepada pelanggan dengan tingkat pendidikan tinggi.