FINAL REPORT INCREASING CUSTOMER'S ACCEPTANCE IN MARKETING CAMPAIGN

Rakamin Data Science Batch 10

TEAM 1- GUIDO

Abdullah Ilman Fahmi
Deneva Widyaningtyas
Dimas Susanto
Ghaisani Anindya Ayuningtyas
Naufal Faherza Putra
Rahmi Ramadhani

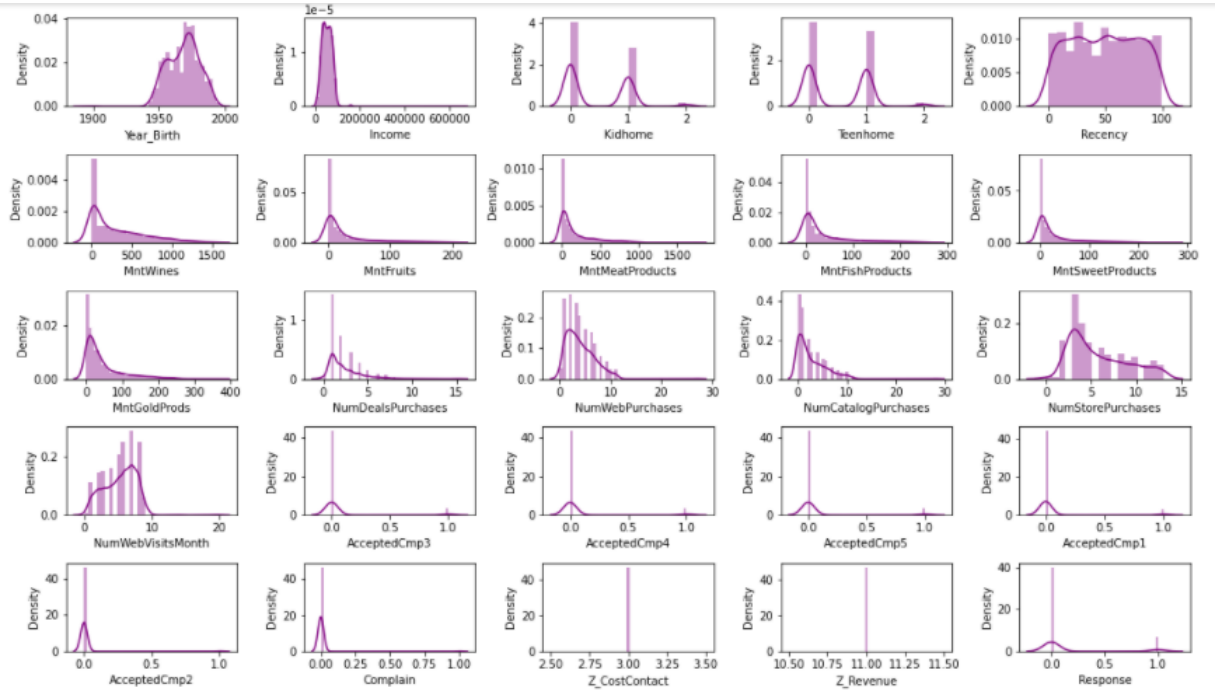
Problem Statement

Untuk menjaga *streak* kenaikan response rate, Guido's Market ditugaskan untuk mencari cara agar *response* untuk *campaign* selanjutnya terus meningkat. Di *campaign* terakhir, 15% pelanggan yang menerima *campaign* tersebut. Hal ini menyebabkan perbedaan sebesar $3046 antara *Cost* dan *Revenue*. Untuk menyelesaikan tugas tersebut, tim *Data Science Guido's Market* membuat model yang memberikan rekomendasi kriteria penting untuk menentukan pelanggan yang akan diberikan *campaign* selanjutnya. *Metric* yang akan menentukan kesuksesan dari tugas yang diberikan kepada tim *data science* adalah *response rate* pelanggan terhadap *campaign* selanjutnya.

Untuk dapat melaksanakan tugasnya, tim *data science* diberikan dataset yang berisikan informasi mengenai pelanggan, histori pembelian produk, dan respon pelanggan terhadap *campaign-campaign* sebelumnya. Dataset yang diberikan berisi kolom-kolom sebagai berikut:

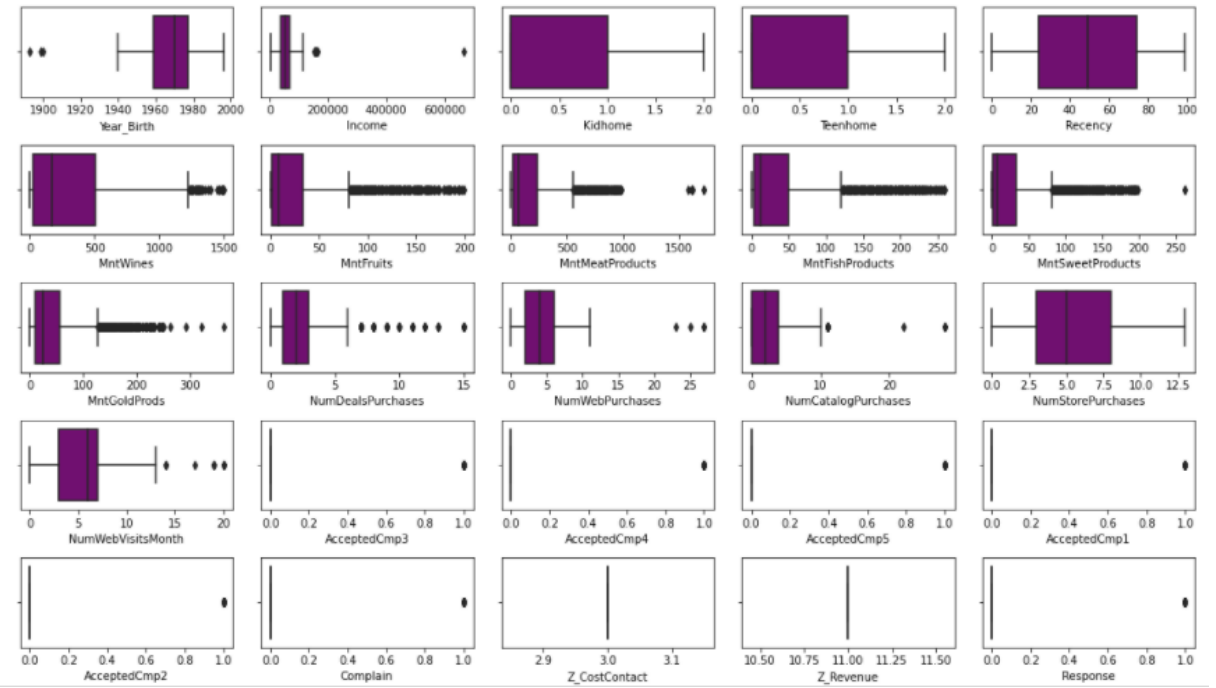
- ID: Nomor ID pelanggan.
- *Year_Birth*: Tahun lahir pelanggan.
- *Education*: Level edukasi pelanggan.
- *Marital_Status*: Status perkawinan pelanggan.
- *Income*: Pemasukkan pelanggan setiap tahunnya.
- *Kidhome*: Jumlah anak kecil yang ditanggung pelanggan.
- *Teenhome*: Jumlah remaja yang ditanggung pelanggan.
- *Dt_Customer*: Tanggal pelanggan mendaftar sebagai member Guido's Market
- *Recency*: Jumlah hari dari pembelian terakhir.
- *MntWines*: Jumlah yang dibelanjakan untuk produk Wine dalam dua tahun terakhir.
- *MntFruits*: Jumlah yang dibelanjakan untuk produk buah dalam dua tahun terakhir.
- *MntMeatProducts*: Jumlah yang dibelanjakan untuk produk daging dalam dua tahun terakhir.
- *MntFishProducts*: Jumlah yang dibelanjakan untuk produk ikan dalam dua tahun terakhir.
- *MntSweetProducts*: Jumlah yang dibelanjakan untuk produk permen dan cokelat dalam dua tahun terakhir.
- *MntGoldProds*: Jumlah yang dibelanjakan untuk produk emas dalam dua tahun terakhir.
- *NumDealsPurchases*: Jumlah pembelian menggunakan diskon.
- NumWebPurchases: Jumlah pembelian melalui website.
- NumCatalogPurchases: Jumlah pembelian melalui katalog.
- NumStorePurchases: Jumlah pembelian melalui toko.
- *NumWebVisitsMonth*: Jumlah kunjungan website dalam waktu satu bulan terakhir.
- *AcceptedCmp1*: Customer menerima *campaign* pertama atau tidak.
- *AcceptedCmp2*: Customer menerima *campaign* kedua atau tidak.
- *AcceptedCmp3*: Customer menerima *campaign* ketiga atau tidak
- *AcceptedCmp4*: Customer menerima *campaign* keempat atau tidak.
- *AcceptedCmp5*: Customer menerima *campaign* kelima atau tidak
- *Complain*: Jumlah komplain yang diberikan oleh pelanggan.
- *Z_CostContact*
- *Z_Revenue*
- *Response*: Customer menerima *campaign* terakhir atau tidak.

Distribusi Data



Figur 1: Distribusi Data

Dari figur 1 dapat dilihat bahwa distribusi data memiliki kecenderungan *positively / negatively skewed*. Hal ini kemungkinan besar disebabkan karena adanya nilai outlier. Untuk boxplot, hal yang paling dapat diperhatikan adalah keberadaan outlier.



Figur 2: Tampilan Boxplot dari Data

Outlier utama terlihat pada kolom *Year_Birth, Income, MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntFishProducts, MntGoldProductst, NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumWebVisitsMonth.*

Dari boxplot pada figur 2 terlihat data yang cenderung berdistribusi normal yaitu kolom *Recency, NumStorePurchases, Z_CostContact,* dan *Z_Revenue*. Pada kolom *Recency* dan *NumStorePurchases* memiliki keragaman data yang cenderung lebih besar daripada kolom *Z_CostContact* dan *Z_Revenue*, hal ini terlihat dari bentuk kotak boxplotnya.

Sedangkan data yang cenderung positively skewed adalah pada kolom *Income, Kidhome, Teenhome, MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntFishProducts, MntGoldProductst, NumDealsPurchases, NumWebPurchases, NumCatalogPurchases,* dan *NumWebVisitsMonth*. Sedangkan yang memiliki distribusi cenderung *negatively skewed* hanya kolom *Year_Birth.*

Pre-processing

- **Data Cleansing**

Dari dataset yang diberikan, ditemukan 24 *missing value* dalam kolom Income, 201 data duplikat, dan 4 data yang *invalid* pada kolom Year_Birth dan Income. Untuk mengisi *missing value* dalam kolom Income, *value* yang hilang diisi oleh rata-rata nilai Income dari setiap level edukasi pelanggan. Hal ini dilakukan karena nilai penghasilan dari setiap customer cenderung bisa diperkirakan dan tidak memiliki perbedaan angka yang jauh. Nilai rata-rata penghasilan ini juga diambil berdasarkan background pendidikan dari setiap customer. Selain itu, data duplikat ditemukan dengan cara memfilter secara manual kolom *Year_Birth*, Education, Marital_Status, dan Income yang memiliki *value* yang sama persis untuk kolom-kolom lainnya (Figur 1). Data duplikat tersebut di *drop* karena tidak masuk akal dan diasumsikan error. *Data cleansing* terakhir yang dilakukan adalah menghapus 3 data *invalid* pada kolom Year_Birth dan 1 data invalid pada kolom Income. Data *invalid* pada kolom Income memiliki value 666666, sementara pada kolom Year_Birth data yang *invalid* memiliki value 1893, 1899, dan 1900. Data-data tersebut diasumsikan sebagai data yang *invalid* karena tidak masuk akal dan angkanya berbeda jauh dengan data lainnya, selain itu tidak sesuai jika dikorelasikan dengan kolom lainnya.

ID	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases
3	3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3	3

Figur 3: Sebagian data yang memiliki *value* yang sama untuk semua kolom.

- **Feature Engineering**

Feature engineering pertama adalah melakukan label encoding, diantaranya yaitu:

➔ Pertama, mengubah kolom *Dt_Customer* menjadi data numerik yaitu kolom *'Year_Customer'* yang hanya berisi tahun pendaftaran customer

- ➔ Kedua, menjumlahkan kolom *Kidhome* dan *Teenhome* menjadi kolom *'Child'* yang berisi jumlah anak yang dimiliki customer
- ➔ Ketiga, menjumlahkan seluruh amount spent produk yang dibeli customer, yaitu kolom *MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts*, dan *MntGoldProducts* menjadi kolom *'Monetary'*, yaitu kolom total pembelian yang dilakukan customer selama 2 tahun terakhir
- ➔ Keempat, menjumlahkan seluruh pembelian customer dari berbagai platform yaitu *NumDealsPurchases, NumWebPurchases, NumCatalogPurchases*, dan *NumStorePurchases* menjadi kolom *'Frequency'*, yaitu seberapa banyak transaksi yang dilakukan oleh customer dilihat dari platform pembelian.
- ➔ Kelima, mengubah kolom *Marital_Status* menjadi dua kategori, yaitu Single dan Not Single. Kolom baru yang dihasilkan adalah *'MaritalStat'. Status Married* dan *Together* disatukan menjadi *Not Single,* sedangkan status pernikahan *Single, Divorce, Widow,* dan status lainnya menjadi Single. Hal ini dirasa penting dilakukan agar mempersingkat jenis status pernikahan menjadi dua tipe saja.
- ➔ Keenam, menjumlahkan kolom *AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5 menjadi kolom 'AcceptedCmp'* yang berisi total campaign yang diterima per customer.
- ➔ Ketujuh, menghasilkan feature baru yaitu '*TopCategory'* yang berisi category product yang paling banyak dibeli masing-masing customer.

Feature engineering kedua adalah melakukan one hot encoding pada kolom *MaritalStat* dan *Education*. Pada setiap kategori dalam kolom *MaritalStat* diubah menjadi fitur baru yaitu *'MaritalStat_NotSingle'* dan *'MaritalStat_Single'*. Sedangkan pada kolom *Education*, setiap kategori dibuat menjadi suatu kolom baru berdasarkan tiap pendidikannya, yaitu '*Education_2n Cycle','Education_Basic', 'Education_Graduation', 'Education_Master',* dan *'Education_PhD'.*

Feature engineering berupa *one hot encoding* lainnya yang dilakukan adalah membuat segmentasi customer dari penghasilan atau *Income*, dan membaginya menjadi tiga yaitu *High Income, Medium Income, dan Low Income*. Fitur baru yang terbentuk adalah 'Income_Category'. Ketiga kategori ini ditentukan agar dapat mengetahui latar belakang ekonomi dari customer, yang nantinya berguna untuk menentukan target campaign ke customer yang sesuai. Berikutnya adalah melakukan segmentasi kategori usia customer dan membaginya menjadi empat, yaitu *Adults, Middle Age, Pre-Elderly,* dan *Elderly,* yang ditentukan dari Year_Birth customer. Fitur baru yang terbentuk adalah *'Age_Category'*.

Feature engineering yang terakhir adalah menghapus fitur '*Marital_Status', 'Kidhome','Teenhome', 'Dt_Customer', 'Z_CostContact',* dan *'Z_Revenue'*, yang seluruhnya telah tergantikan dengan hasil *feature engineering* dan kemungkinan besar tidak digunakan lagi.

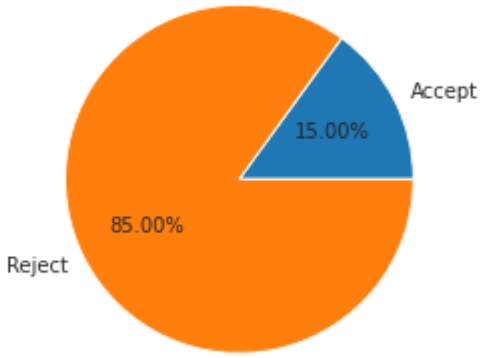
Seluruh *feature engineering* ini dibuat agar memudahkan dalam proses analisis data selanjutnya dan membantu dalam tahapan *modelling*, karena setiap valuenya dapat menjadi fitur baru yang berkaitan dengan kriteria penentuan target customer yang akan menerima *campaign* selanjutnya.

EDA & Insights

Beberapa hasil eksplorasi data yang ditemukan:

1. ***Response***

Fitur yang menjadi target kami adalah *Response*, yaitu data *customer* yang menerima (*Yes*) atau tidak menerima (*No*) pada *campaign* terakhir. Hasil eksplorasi data yang ditemukan adalah bahwa sebanyak 85% atau 1730 orang *customer* menolak *campaign* terakhir, sedangkan hanya sebanyak 15% atau 305 orang *customer* yang menerima *campaign* terakhir.

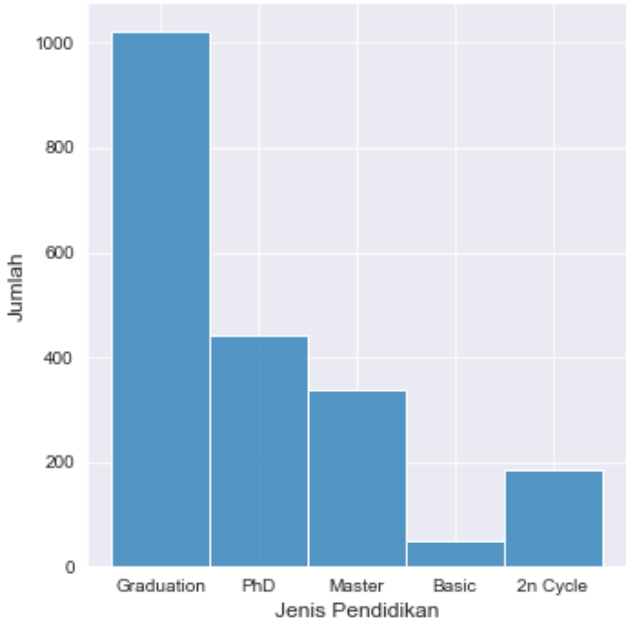


Figur 4: Persentase Customer yang memberikan Respon terhadap *Campaign* terakhir.

Rendahnya angka respon yang menerima *campaign* ini menyebabkan *Guido's Market* perlu menemukan cara untuk meningkatkan respon pelanggan pada *marketing campaign* selanjutnya.

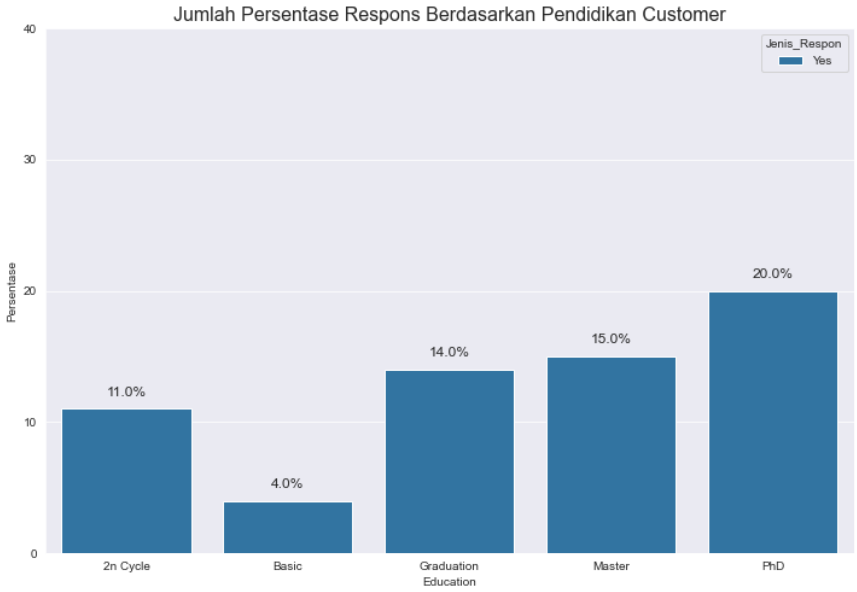
2. ***Education***

Jenis pendidikan dari *customer* terbagi menjadi 5, yaitu jenis pendidikan *Basic, Graduation, 2n Cycle, Master*, dan *PhD*. *Basic* yaitu setara dengan lulusan SD, *Graduation* merupakan lulusan S1, *2n Cycle* merupakan pendidikan setara dengan S2, *Master* merupakan lulusan S2, dan *PhD* merupakan pendidikan S3. Dari kelima jenis latar belakang pendidikan ini, jumlah yang paling banyak adalah latar belakang pendidikan S1 atau *Graduation* yaitu sebanyak 1024 orang, dan yang paling sedikit adalah *Basic* yaitu hanya 49 orang.



Figur 5 : Jumlah Jenis Pendidikan

Dari visualisasi diatas dapat diketahui bahwa *customer* dengan latar belakang jenis pendidikan tergolong sering melakukan transaksi di *Guido's Market*, hal ini dapat dikaitkan pula bahwa latar belakang pendidikan terbanyak yang dimiliki oleh penduduk di daerah tersebut adalah lulusan S1, dan hanya sebagian orang yang memilih untuk melanjutkan pendidikan ke jenjang S2 maupun S3. Selain itu berarti kebutuhan produk dari lulusan pendidikan S1 lebih banyak dibanding jenis pendidikan lainnya.



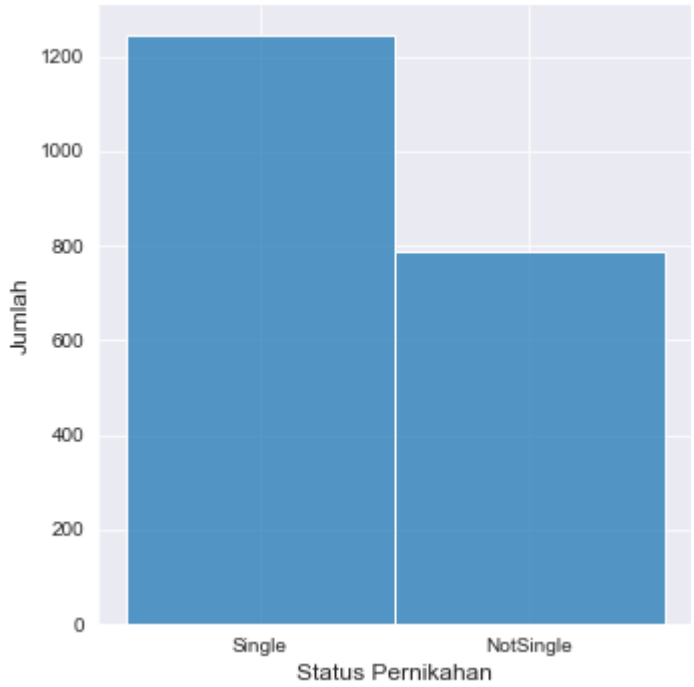
Figur 6 : Jumlah Persentase Respon *Yes* Berdasarkan Jenis Pendidikan *Customer*

Adapun bila dikaitkan dengan fitur target yaitu respon, terlihat bahwa *response rate* tertinggi berada pada latar belakang *PhD* yaitu sebesar 20%, dan yang terendah adalah *Basic* sebesar 4%. *Response rate* disini berarti peluang terbesar yang memberikan respon pada campaign adalah dari latar belakang pendidikan *PhD*.

Dari visualisasi ini dapat diambil kesimpulan bahwa peluang *customer* yang aktif merespon adalah dari latar belakang pendidikan S3 atau *PhD*, yang berarti jenis *customer* ini diprediksi akan selalu menerima tawaran yang diberikan. Walaupun jumlah keseluruhan *customer* nya tidak terlalu banyak, namun rata-rata orangnya mau memberikan respon karena memang aktif dalam belanja dan pencarian promosi. Sedangkan latar belakang pendidikan SD atau *Basic* paling rendah dalam memberikan respon, mungkin saja karena jumlahnya yang kurang banyak dan kurang aktif dalam berbelanja

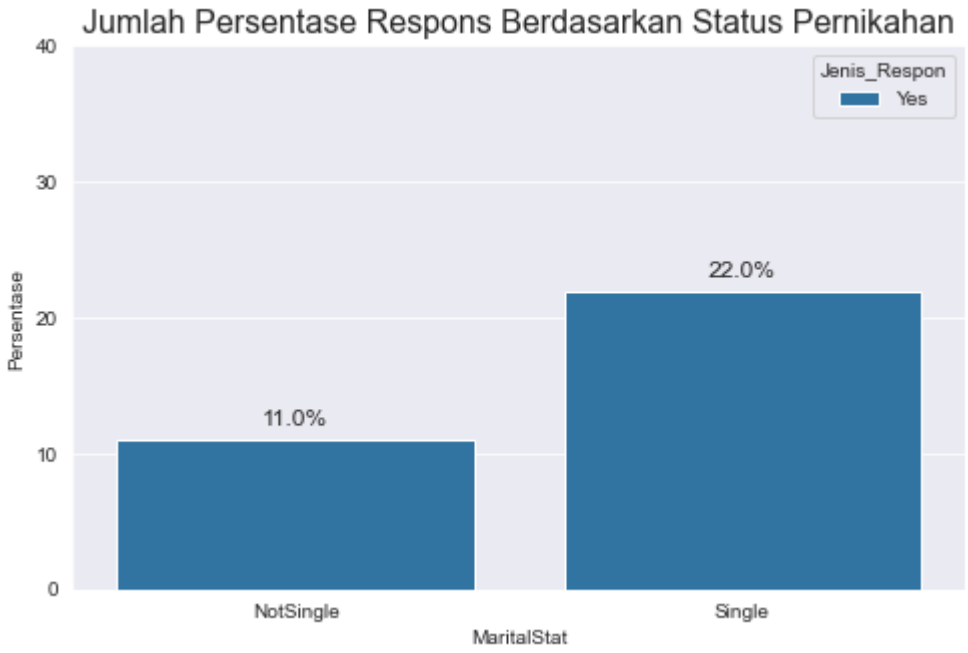
3. Marital Status

Jenis status pernikahan dari customer terbagi menjadi 2, yaitu *Single* dan *Not Single. Single* yaitu belum menikah, dan janda/duda yang sudah bercerai/ditinggal mati pasangannya. *Not Single* yaitu *customer* yang sudah menikah atau tinggal bersama. Dari kedua jenis status pernikahan ini, jumlah yang paling banyak adalah status *Single* yaitu sebanyak 1247 orang, dibandingkan yang *Not Single* yaitu 788 orang.



Figur 7 : Jumlah Status Pernikahan

Dari visualisasi ini dapat diketahui bahwa *customer* dengan status pernikahan *Single* tergolong sering melakukan transaksi di Guido's Market, hal ini dapat dikaitkan pula bahwa status pernikahan terbanyak yang dimiliki oleh penduduk di daerah tersebut adalah belum menikah ataupun janda/duda. Selain itu berarti kebutuhan produk dari golongan ini lebih banyak dibanding yang sudah berkeluarga.

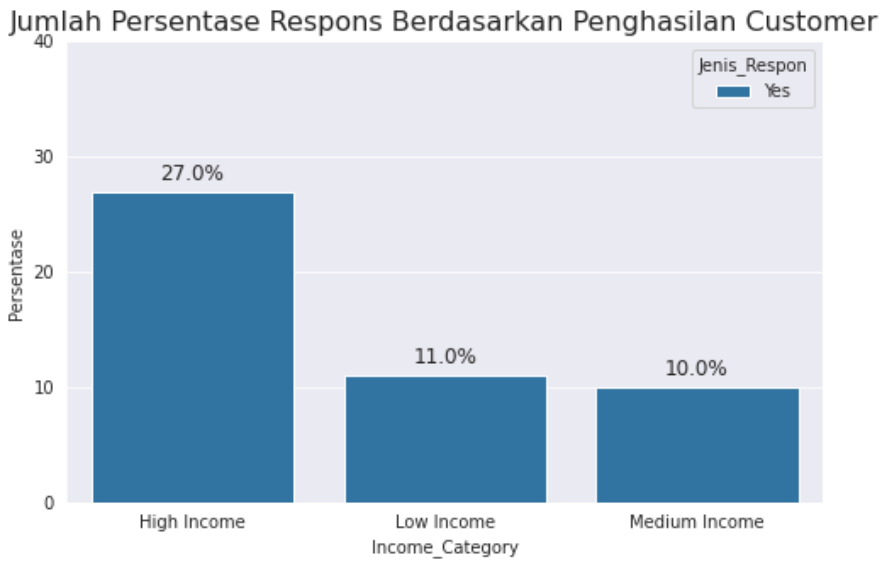


Figur 8 : Jumlah Persentase Respon *Yes* Berdasarkan Status Pernikahan *Customer*

Adapun bila dikaitkan dengan fitur target yaitu respon, terlihat bahwa *response rate* tertinggi berada pada status pernikahan *Not Single* yaitu sebesar 22%. *Response rate* disini berarti peluang terbesar yang akan memberikan respon pada *campaign* adalah dari status pernikahan yang belum berkeluarga. Dari visualisasi ini dapat diambil kesimpulan bahwa peluang *customer* yang aktif merespon adalah dari status pernikahan *Single*, yang berarti jenis *customer* ini diprediksi akan selalu menerima tawaran yang diberikan.

4. Income

Jenis penghasilan dari *customer* terbagi menjadi 3, yaitu *High Income, Medium Income,* dan *Low Income*. *High Income* yaitu yang memiliki penghasilan lebih dari $68522 (yang diambil dari nilai kuartil 3), *Medium Income* yaitu yang memiliki penghasilan kurang dari $68522 dan lebih dari $51381 (yang diambil dari nilai kuartil 2/median), dan *Low Income* yang memiliki penghasilan kurang dari $51381 (nilai kuartil 2/median). Dari ketiga jenis golongan penghasilan ini, jumlah *customer* yang paling banyak adalah yang merupakan golongan *Low Income* sebanyak 1003 orang, dan yang paling sedikit adalah golongan *High Income* sebanyak 503 orang. Dapat diambil kesimpulan bahwa *customer* yang sering bertransaksi di Guido's Store atau golongan penduduk di daerah tersebut memiliki latar belakang penghasilan yang cenderung rendah.



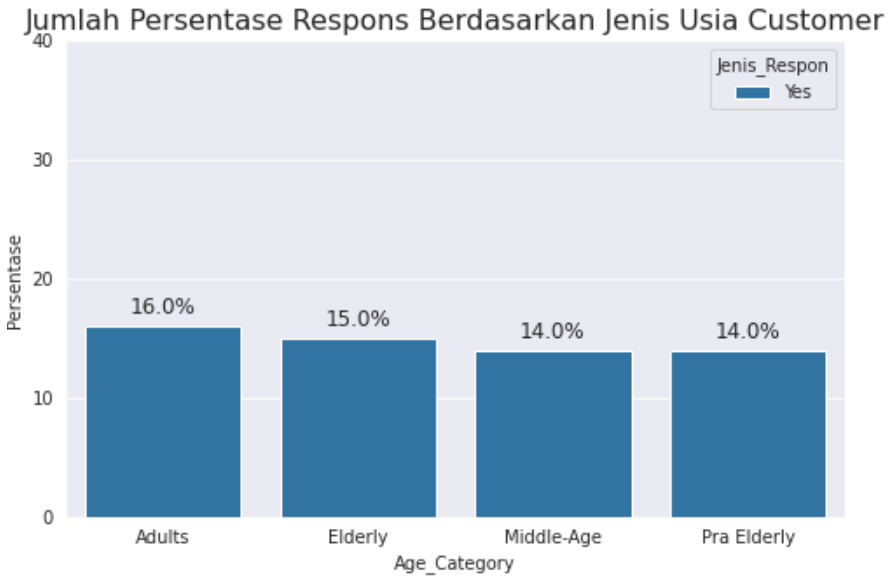
Figur 9 : Jumlah Persentase Respon *Yes* Berdasarkan Jenis Penghasilan *Customer*

Adapun bila dikaitkan dengan fitur target yaitu respon, terlihat bahwa *response rate* tertinggi berada pada latar belakang penghasilan *High Income* yaitu sebesar 27%, dan yang terendah adalah *Medium Income* sebesar 10%. *Response rate* disini berarti peluang terbesar yang akan memberikan respon pada *campaign* adalah dari latar belakang *customer* berpenghasilan tinggi.

Dari visualisasi diatas dapat diambil kesimpulan bahwa peluang *customer* yang aktif merespon adalah dari golongan penghasilan *High Income*, yang berarti jenis *customer* ini diprediksi akan selalu menerima tawaran yang diberikan. Walaupun jumlah keseluruhan *customer* nya yang bertransaksi tidak terlalu banyak, namun rata-rata orangnya selalu mau memberikan respon tiap diberikan *campaign*. Kemungkinan juga karena penghasilan mereka yang memang tinggi sehingga menjadi aktif menerima.

5. Kategori Umur

Kategori umur dari *customer* terbagi menjadi 4, yaitu *Adults, Middle Age, Pre Elderly*, dan *Elderly*. *Adults* atau Dewasa yaitu yang memiliki tahun kelahiran lebih besar dari tahun 1977 (yang diambil dari nilai kuartil 3), *Middle Age* atau Paruh Baya yaitu yang memiliki tahun kelahiran diantara tahun 1977 dan 1970 (yang diambil dari nilai kuartil 2/median), *Pre Elderly* atau Pralansia yang memiliki tahun kelahiran diantara tahun 1970 dan 1959 (nilai kuartil 1), dan *Elderly* atau Lansia yang memiliki tahun kelahiran lebih kecil dari tahun 1959.



Figur 10 : Jumlah Persentase Respon *Yes* Berdasarkan Kategori Umur *Customer*

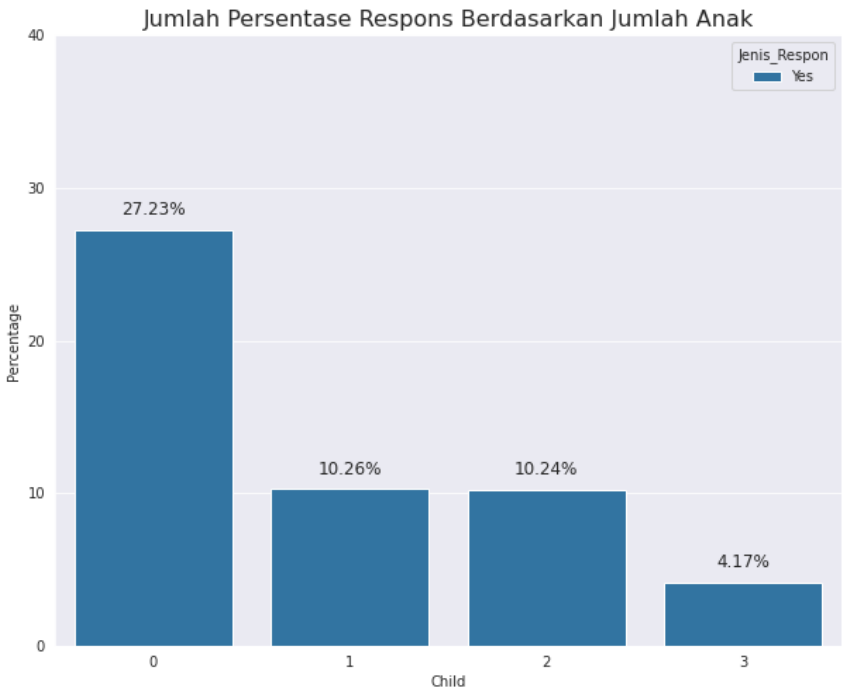
Adapun bila dikaitkan dengan fitur target yaitu respon, terlihat bahwa persentase *response rate* cenderung hampir sama antara kategori umur satu dengan yang lainnya. Namun angka tertinggi berada pada kategori umur *Adults* yaitu sebesar 16%, dan yang terendah adalah *Middle Age* dan *Pra Elderly* sebesar 14%. *Response rate* disini berarti peluang terbesar yang akan memberikan respon pada *campaign* adalah dari kategori usia dewasa, walaupun perbedaan persentasenya sangatlah tipis dengan yang berusia lanjut.

Dari visualisasi diatas dapat diambil kesimpulan bahwa peluang *customer* yang aktif merespon adalah dari golongan usia *Adults*, yang berarti jenis *customer* ini diprediksi akan selalu menerima tawaran yang diberikan. Rata-rata orang dari jenis usia ini selalu mau memberikan respon, kemungkinan karena merupakan usia yang produktif untuk berbelanja dan menerima promo.

6. Child

Jumlah anak yang dimiliki dari *customer* terbagi menjadi 4, yaitu tidak memiliki anak sama sekali (0), memiliki 1 anak, 2 anak, dan 3 anak. Adapun jumlah anak ini diambil dari kolom *Kidhome* dan *Teenhome*, yang berarti jumlah anak kecil dan remaja yang ditanggung oleh pelanggan.

Dari visualisasi dibawah ini, bila dikaitkan dengan fitur target yaitu respon, terlihat bahwa persentase *response rate* paling tinggi yaitu pada golongan yang belum memiliki anak, yaitu sebesar 27,23% dan yang paling rendah adalah yang memiliki 3 anak, sebesar 4,17%. Hal ini berarti peluang *customer* yang akan menerima respon adalah dari golongan yang memang belum memiliki anak.

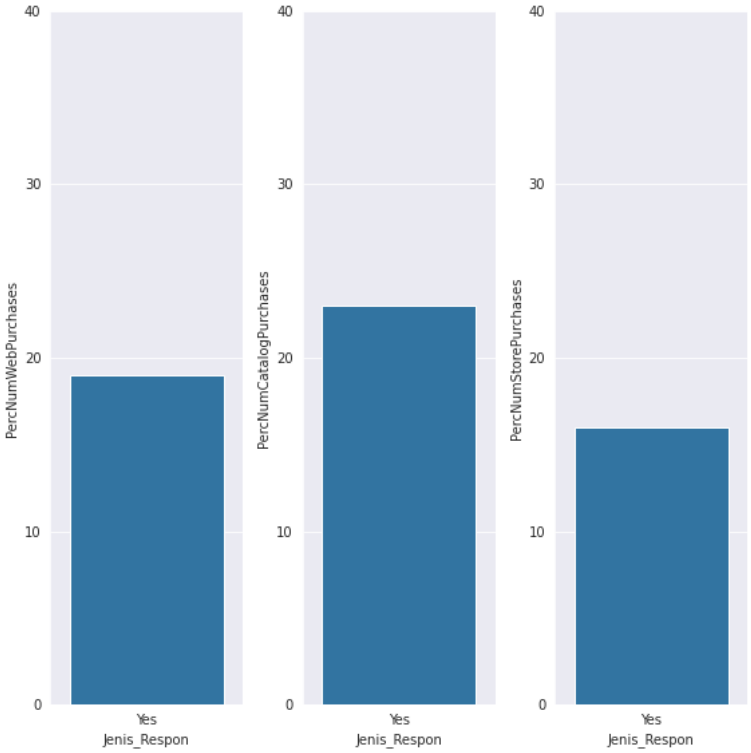


Figur 11 : Persentase Respon *Yes* Berdasarkan Jumlah Anak *Customer*

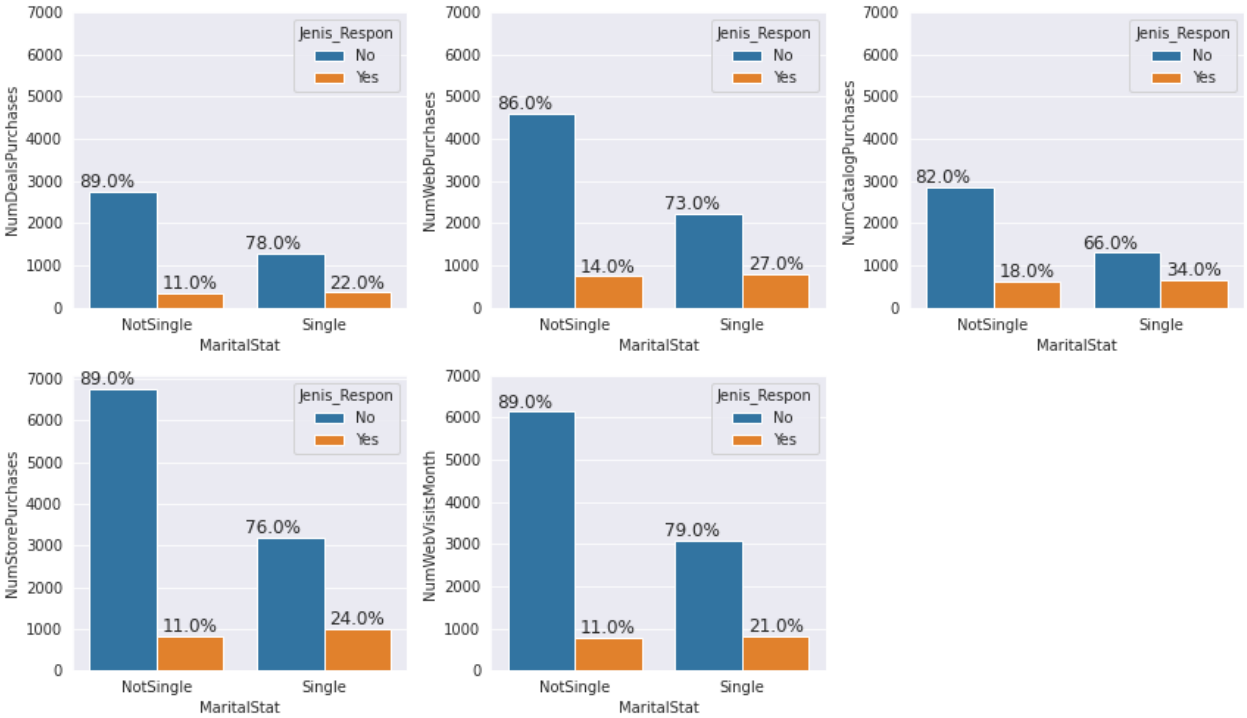
7. Purchases

Jenis *platform* pembelian terbagi menjadi 3, yaitu *Website, Catalog*, dan *Store*. *Website* yaitu pembelanjaan dari produk yang dipasang di situs *web*. *Catalog* yaitu pembelanjaan dari katalog yang disebar ke *customer*, dan *Store* merupakan jenis pembelanjaan langsung dari toko Guido's Market. Dari ketiga jenis status pernikahan ini, jumlah yang paling banyak adalah *platform* pembelian dari *Store* yaitu sebanyak 1337 transaksi, dan yang paling sedikit adalah *platform* pembelian dari *Catalog* sebanyak 144 transaksi. Dapat diambil kesimpulan bahwa banyak pelanggan yang memang menyukai metode pembelian secara langsung di toko.

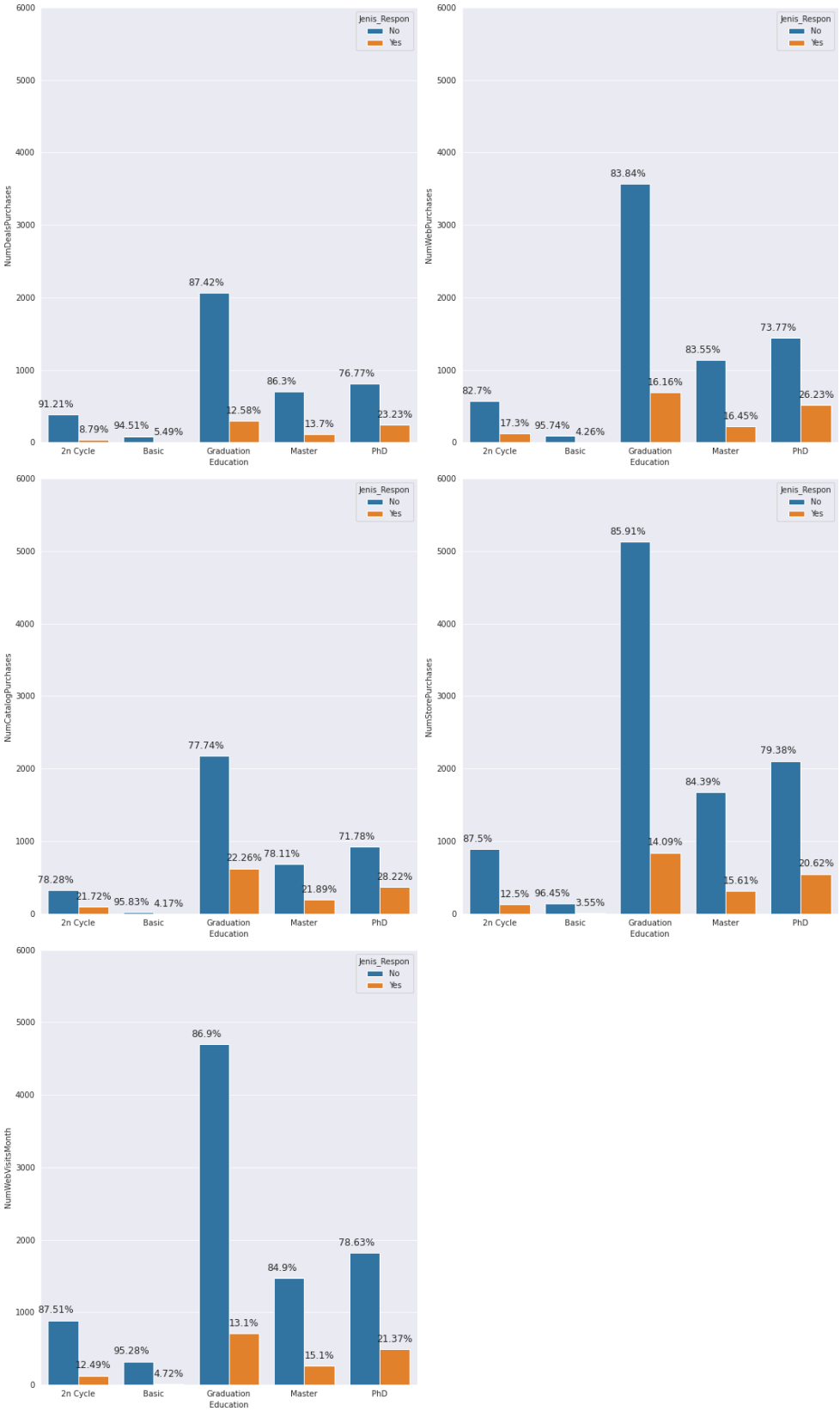
Sedangkan untuk *response rate* nya, angka paling tinggi berada pada *platform* pembelian menggunakan *Catalog* yaitu sebesar 23% dan yang paling rendah yaitu pada *Store* yang hanya sebesar 16%. Dari visualisasi dibawah ini dapat diambil kesimpulan bahwa peluang *customer* yang aktif merespon adalah yang mendapatkan sebaran katalog, dan yang paling jarang adalah dari *Store* atau toko. Hal ini berarti jenis *customer* dari katalog diprediksi lebih menyukai promo dan selalu menerima tawaran yang diberikan. Walaupun jumlah keseluruhan *customer* katalog tergolong paling rendah, namun rata-rata orangnya kemungkinan mau memberikan respon karena memang aktif dalam belanja dan pencarian promosi. Sedangkan pembelian dari toko tergolong sedikit dalam merespon, mungkin saja karena mereka lebih menyukai harga yang tertera di toko dan tidak terbiasa dalam menjumpai *marketing campaign*.



Figur 12 : Jumlah Persentase Respon *Yes* Berdasarkan Jenis *Platform* Pembelian



Figur 13 : Persentase Respon Berdasarkan *Platform* Pembelian dan Status Pernikahan *Customer*

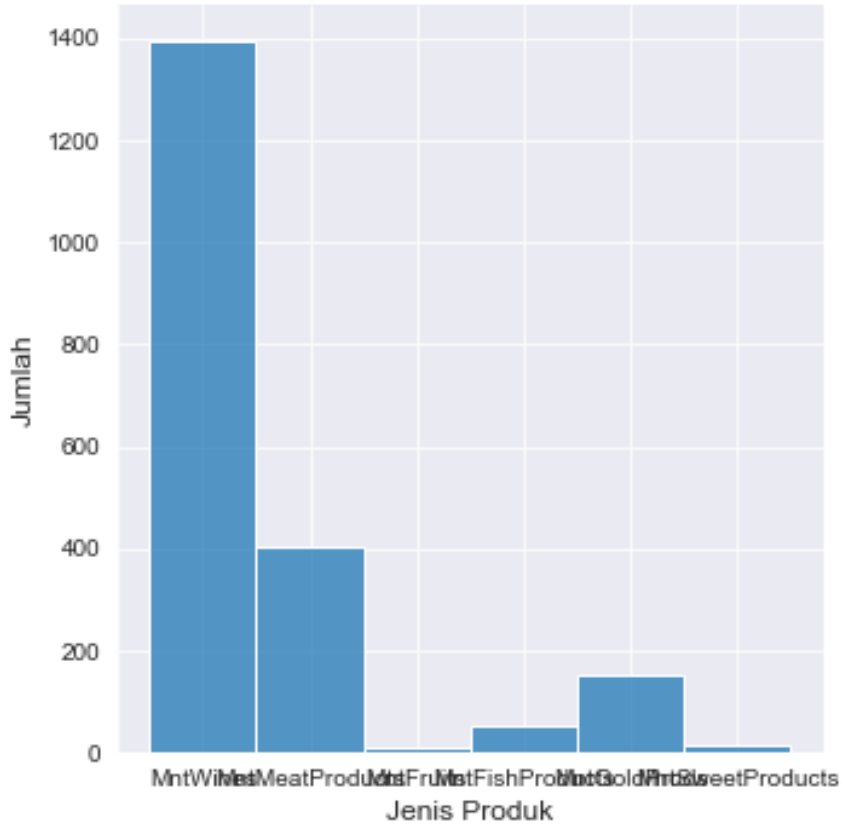


Figur 14 : Persentase Respon Berdasarkan *Platform* Pembelian dan Jenis Pendidikan *Customer*

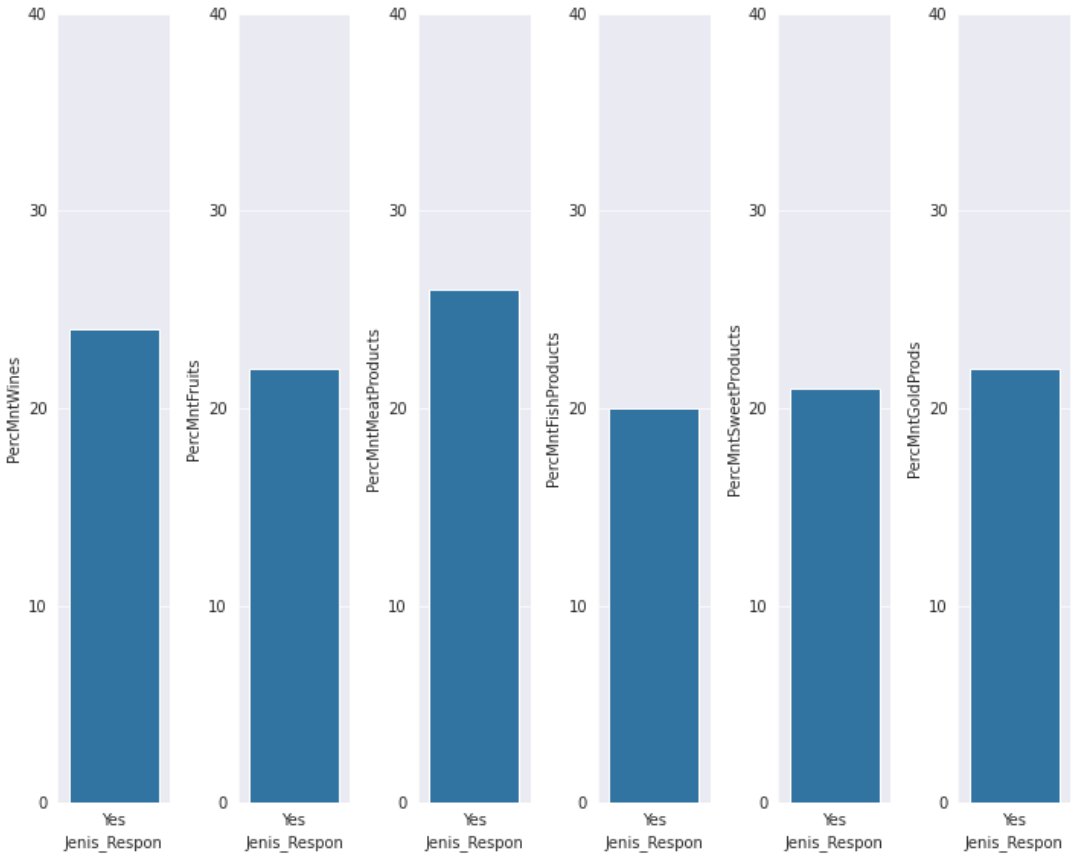
Beberapa visualisasi dari Figur 13 dan Figur 14 menunjukkan bahwa peluang jenis pendidikan yang paling banyak merespon dengan *platform* pembelian *Catalog* adalah dari *PhD* sebesar 28,22%, dan status pernikahan terbanyak adalah *Single*, yang juga menggunakan *platform* katalog sebesar 34%. Hal ini membuktikan keterkaitan *response rate* antara *platform* pembelian *website* dengan jenis pendidikan dan status pernikahan yang juga tergolong tinggi pada dua kategori tersebut.

8. Products

Jenis produk pembelian terbagi menjadi 6, yaitu *Wine* atau minuman anggur, *Fruit* atau buah-buahan, *Meat* atau daging, *Fish* atau ikan, *Sweet* atau makanan manis, *Gold* atau emas. Dari kelima jenis produk ini, jumlah yang paling banyak adalah pembelian *Wine* yaitu sebanyak 1397 transaksi, dan yang paling sedikit adalah pembelian buah-buahan hanya sebanyak 12 transaksi. Dapat diambil kesimpulan bahwa banyak pelanggan yang memang menyukai produk minuman anggur.



Figur 15 : Jumlah Produk Pembelian

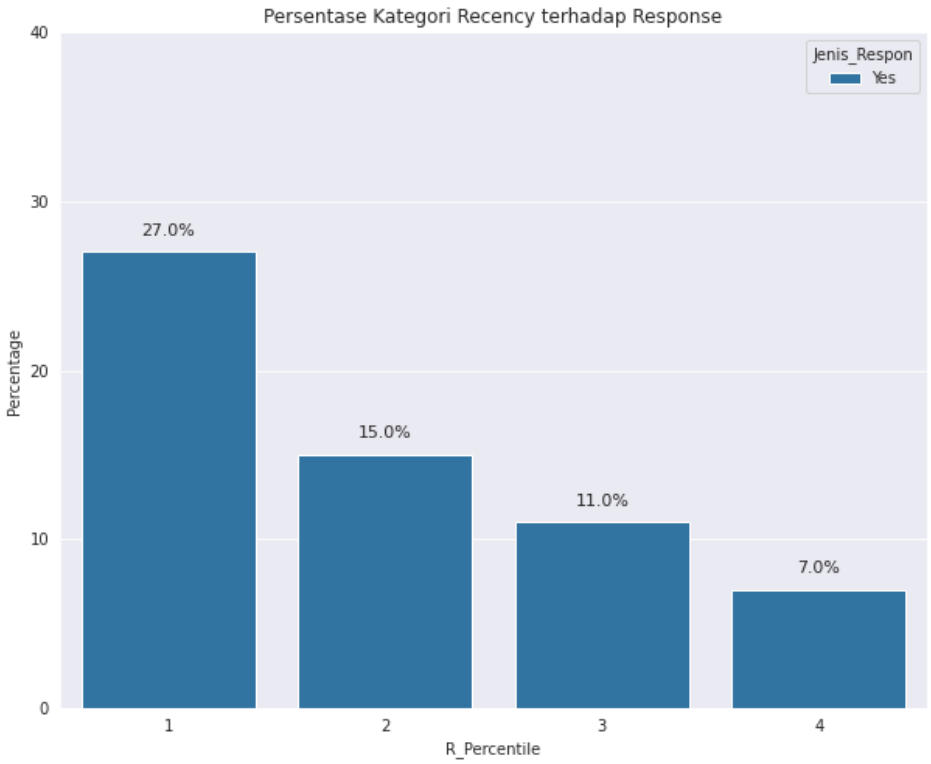


Figur 16 : Jumlah persentase Respon *Yes* Berdasarkan Produk Pembelian

Sedangkan untuk *response rate* nya, angka paling tinggi berada pada pembelian produk *Meat* yaitu sebesar 26% dan yang paling rendah yaitu produk *Fish* sebesar 20%. Dari visualisasi diatas dapat diambil kesimpulan bahwa peluang *customer* yang aktif merespon adalah yang sering membeli produk daging, dan yang paling jarang adalah yang membeli produk ikan. Hal ini berarti jenis *customer* yang membeli daging diprediksi lebih menyukai promo dan selalu menerima tawaran yang diberikan. Walaupun jumlah *customer* yang membeli daging kurang begitu banyak, namun rata-rata orangnya mau memberikan respon pada *campaign*.

9. Recency

Recency merupakan rentang waktu transaksi terakhir pelanggan hingga saat ini. Nilai *recency* ini dikategorikan menjadi 4 kelompok, yaitu Kategori 1 merupakan *customer* yang melakukan transaksi terakhir selama 0 - 23 hari kebelakang, Kategori 2 merupakan *customer* yang melakukan transaksi terakhir selama 24 - 48 hari kebelakang, Kategori 3 merupakan *customer* yang melakukan transaksi terakhir selama 49 - 73 hari kebelakang, dan Kategori 4 yang merupakan *customer* yang melakukan transaksi terakhir selama 74 - 99 hari kebelakang.

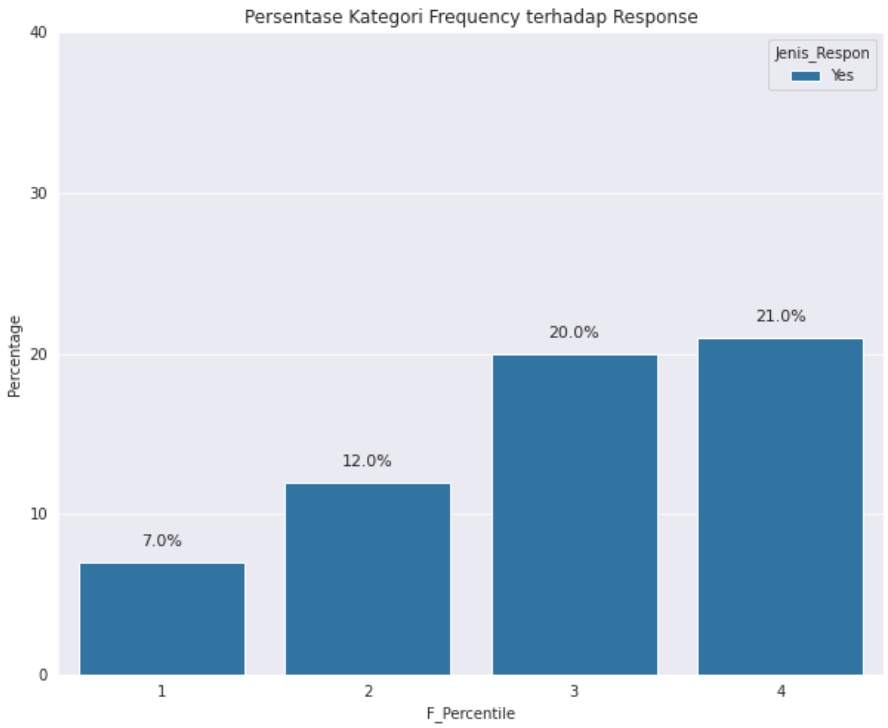


Figur 17 : Jumlah Persentase Respon *Yes* Berdasarkan *Recency*

Dari keempat jenis kategori *Recency* tersebut, yang memiliki peluang paling banyak memberikan respon adalah kategori 1, yaitu *customer* yang melakukan transaksi terakhir dibawah 24 hari, dengan *response rate* sebesar 27%. Hal ini berarti bahwa prediksi golongan *customer* yang aktif menerima *campaign* adalah yang memang melakukan transaksi paling terbaru.

10. Frequency

Frequency merupakan seberapa banyak transaksi yang dilakukan per *customer* dilihat *platform* pembelian (*NumWebPurchases, NumCatalogPurchases, NumStorePurchases*). Nilai *frequency* ini dikategorikan menjadi 4 kelompok, yaitu Kategori 1 merupakan *customer* yang melakukan transaksi sebanyak 0 - 5 kali, Kategori 2 merupakan *customer* yang melakukan transaksi sebanyak 6 - 11 kali, Kategori 3 merupakan *customer* yang melakukan transaksi sebanyak 12 - 18 kali, dan Kategori 4 yang merupakan *customer* yang melakukan transaksi terakhir sebanyak 18 - 32 kali.

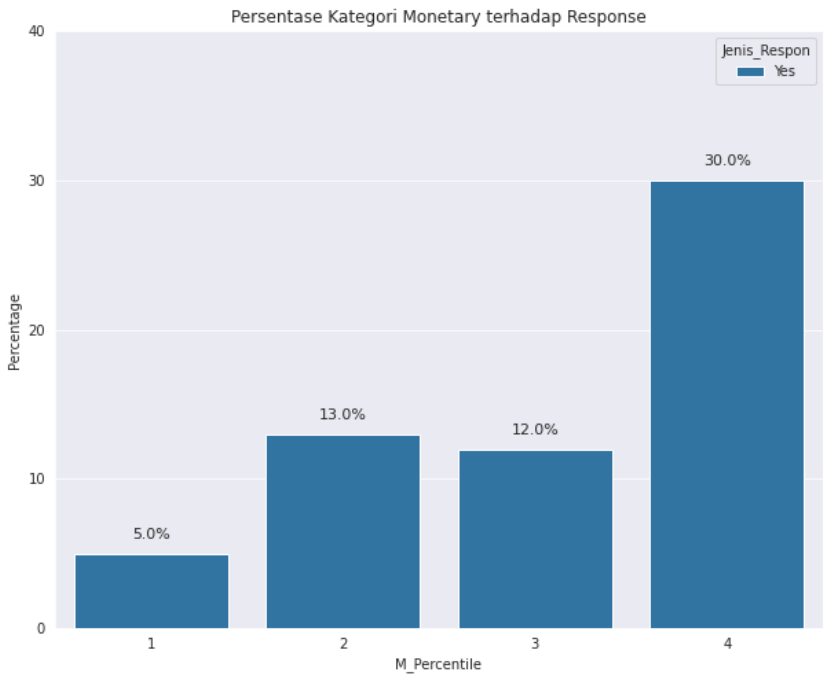


Figur 18 : Jumlah Persentase Respon *Yes* Berdasarkan *Frequency*

Dari keempat jenis kategori *Frequency* tersebut, yang memiliki peluang paling banyak memberikan respon adalah kategori 4, yaitu *customer* yang melakukan transaksi terakhir sebanyak 18 - 32 kali dengan *response rate* sebesar 21%. Hal ini berarti bahwa prediksi golongan *customer* yang aktif menerima *campaign* adalah yang memang melakukan banyak transaksi.

11. Monetary

Monetary merupakan total pembelian yang dilakukan oleh *customer* selama 2 tahun terakhir. Nilai *monetary* ini dikategorikan menjadi 4 kelompok, yaitu Kategori 1 merupakan *customer* yang melakukan total pembelian sebanyak $5 - $68, Kategori 2 merupakan *customer* yang melakukan total pembelian sebanyak $69 - $395, Kategori 3 merupakan *customer* yang melakukan total pembelian $396 - $1044, dan Kategori 4 merupakan *customer* yang melakukan total pembelian $1045.

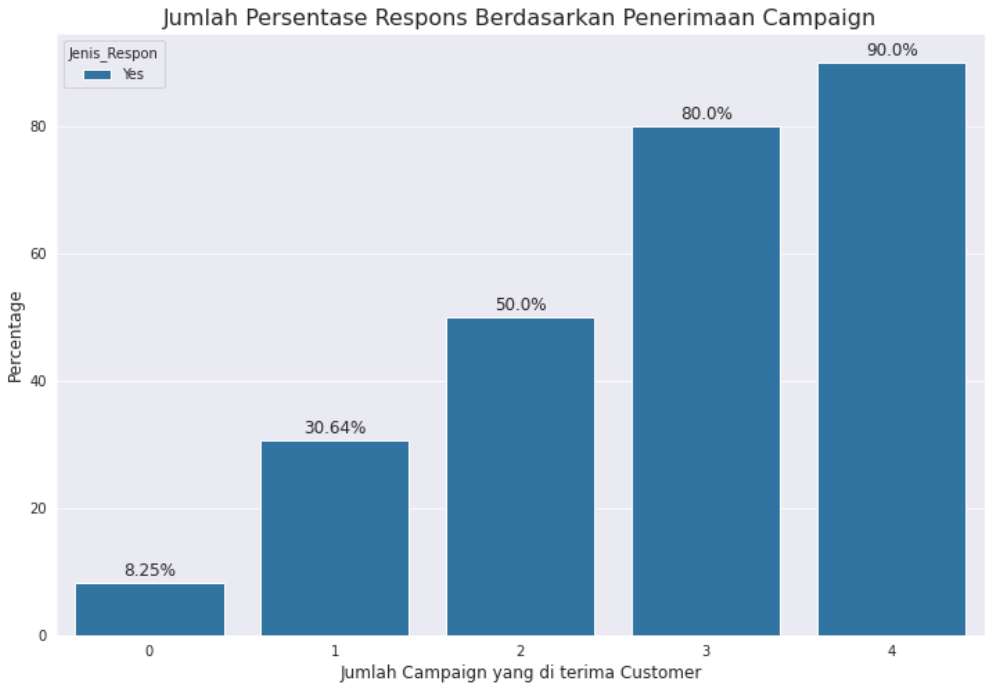


Figur 19 : Jumlah Persentase Respon *Yes* Berdasarkan *Monetary*

Berdasarkan *Monetary* dapat disimpulkan *response rate* tertinggi yaitu pada kategori 4 yaitu *customer* yang melakukan total pembelian $1045, sebesar 30%. Hal ini berarti peluang *customer* yang akan menerima *campaign* adalah dari kategori *monetary* 4 yaitu yang banyak melakukan pembelian.

12. Accepted Campaign

Berdasarkan tipe *campaign* yang diterima, semakin sering *customer* menerima campaign maka kemungkinan untuk menerima *campaign* selanjutnya lebih besar. Hal ini dibuktikan dengan angka *response rate* paling tinggi adalah pada *customer* yang menerima *campaign* sebanyak 4 kali, yaitu sebesar 90%, dibandingkan dengan yang tidak pernah menerima *campaign* sama sekali. Yang berarti peluang penerimaan responnya akan semakin besar bila semakin banyak *customer* tersebut menerima *campaign*.



Figur 20 : Jumlah Persentase Respon *Yes* Berdasarkan Penerimaan *Campaign*

Insights

Dari hasil eksplorasi data awal ini dapat diambil *insight* yaitu *customer* yang memiliki peluang paling banyak dalam memberikan respon pada *campaign* adalah *customer* yang memiliki latar belakang pendidikan *PhD*, dengan status pernikahan *Single*, tidak memiliki anak, kategori usia dewasa, memiliki penghasilan tinggi, dan sering melakukan pembelian melalui *platform Catalog*. Sedangkan untuk produk pembelian yang paling banyak adalah produk daging. Untuk *recency*, yang memiliki peluang paling banyak memberikan respon adalah kategori *customer* yang melakukan transaksi terakhir di bawah 24 hari, dengan *frequency* atau yang melakukan transaksi terakhir sebanyak 18 - 32 kali, dan *monetary* yaitu *customer* yang melakukan total pembelian $1045. Dan semakin sering *customer* menerima *campaign* maka peluang untuk menerima *campaign* selanjutnya lebih besar.

Hasil eksplorasi tersebut dapat dijadikan sebagai rekomendasi bisnis yaitu menyesuaikan target *customer* yang menerima *campaign* dengan kriteria-kriteria tersebut. Hal ini dapat meningkatkan kemungkinan *customer* memberikan respon terhadap *campaign* selanjutnya.

Modeling Experiment

Dari data yang dimiliki dan *feature engineering* yang telah dibuat, tim *data science* selanjutnya membuat model untuk menentukan fitur-fitur yang paling berpengaruh terhadap *Response* dengan menggunakan *Feature Importance*. Langkah-langkah *Machine Learning Modelling* yang dilakukan sebagai berikut:

1. Data yang sudah melalui tahap *pre-processing* di split menjadi *test* dan *train set* dengan rasio 70:30 dan *random state* 50.
2. Setelah di *split*, data di *balance* menggunakan metode *SMOTE, Oversampling*, dan *Undersampling*.
3. Membuat model dengan metode *Decision Tree, Random Forest, XGBoost*, dan *AdaBoost* untuk data dari masing-masing metode *balancing.*
4. Membuat model setelah dilakukan *Hyperparameter Tuning* untuk setiap *Machine Learning Modeling* yang dicoba.
5. Hasil dari setiap *modelling* berisi nilai *Accuracy (Test Set)*, *Precision (Test Set), Recall (Test set)*, *F-1 Score (Test Set)*, *AUC, Train Score*, dan *Test Score*.
6. Menunjukkan *Feature Importance score* dan *plot.*

Hasil dari *modelling* sebagai berikut:

	Decision Tree	***Random Forest***	***XGBoost***	***AdaBoost***
Accuracy	0.83	0.90	0.90	0.90
Precision	0.36	0.69	0.68	0.61
Recall	0.42	0.38	0.39	0.54
F-1	0.39	0.49	0.50	0.57
AUC	0.66	0.68	0.68	**0.75**
Train Score	1.00	1.00	0.80	0.72
Test Score	0.66	0.68	0.68	0.75

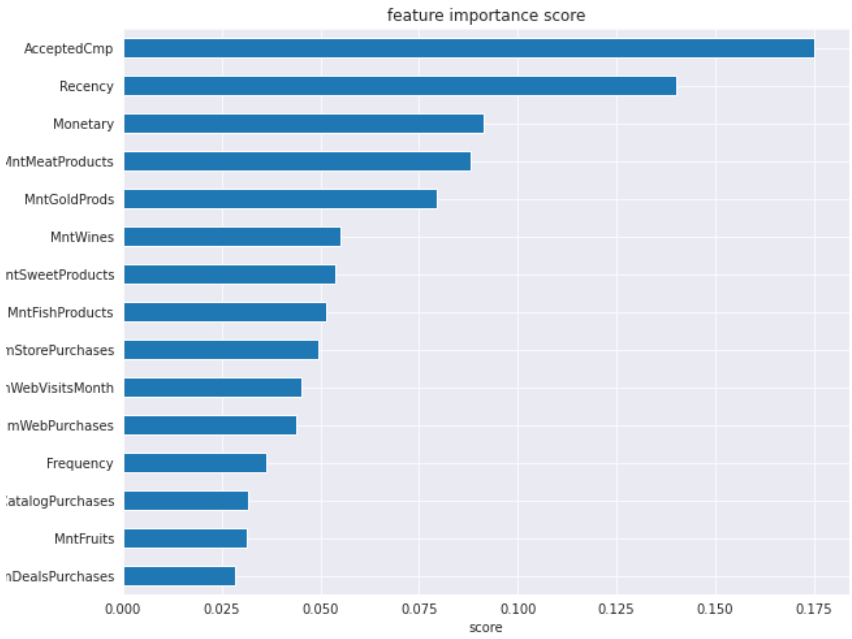
Tabel 1: Hasil *modelling* sebelum *Hyperparameter Tuning*

Hasil setelah dilakukan *Hyperparameter Tuning*:

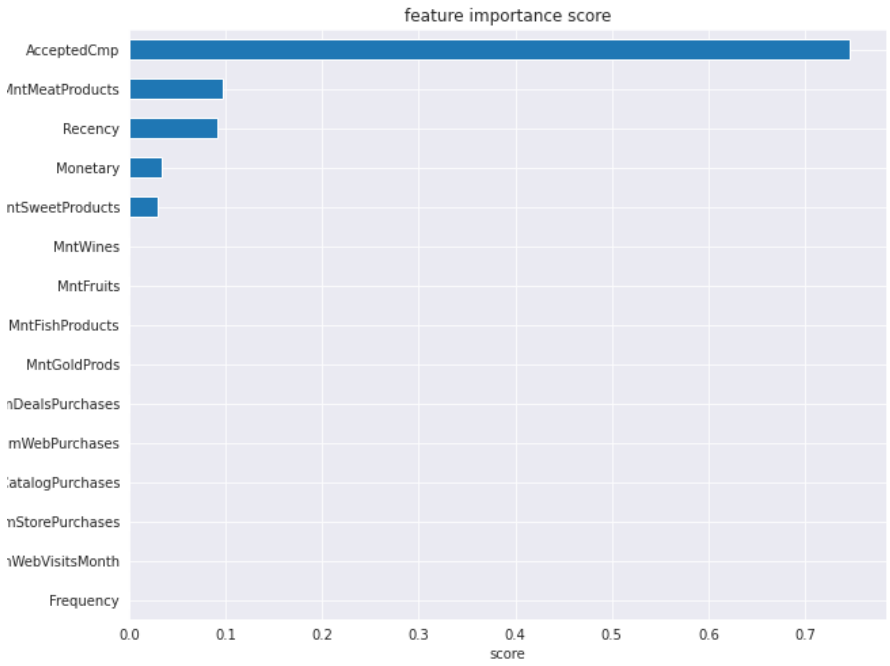
	Decision Tree	***Random Forest***	***XGBoost***	***AdaBoost***
Accuracy	0.88	0.90	0.90	0.88
Precision	0.53	0.78	0.67	0.71
Recall	0.32	0.28	0.38	0.07
F-1	0.40	0.41	0.49	0.12
AUC	0.64	0.63	0.68	0.53
Train Score	0.62	0.79	0.83	0.55
Test Score	0.64	0.63	0.68	0.53

Tabel 2: Hasil *modelling* setelah *Hyperparameter Tuning*

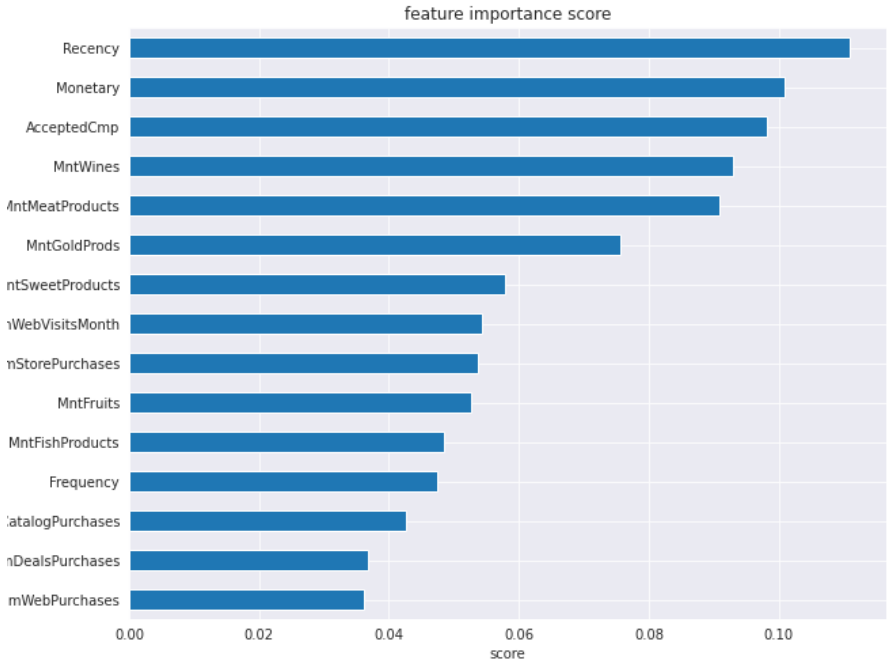
Hasil *Feature Importance*:



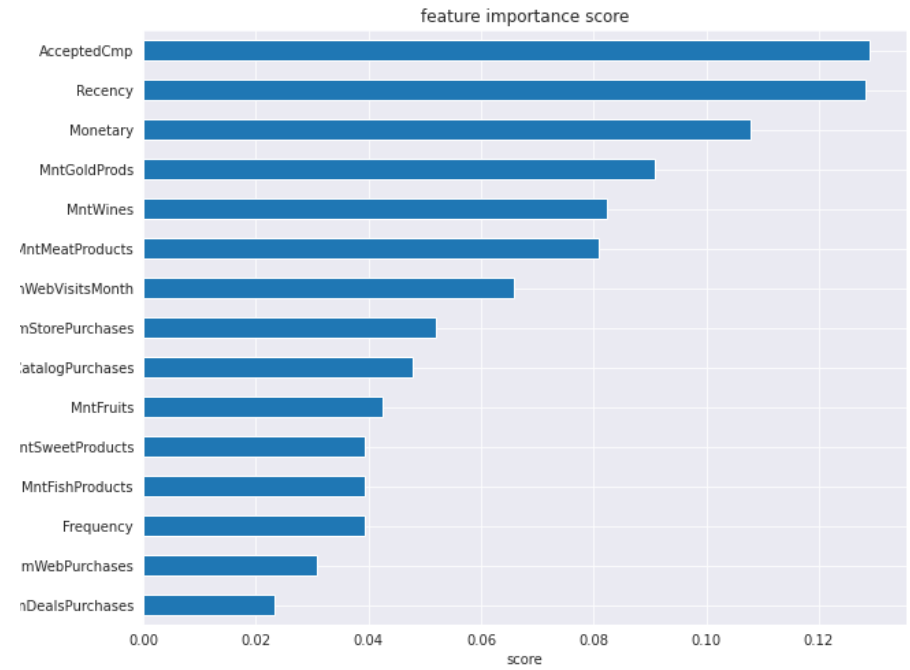
Figur 21: *Feature Importance Decision Tree* sebelum *Hyperparameter Tuning*



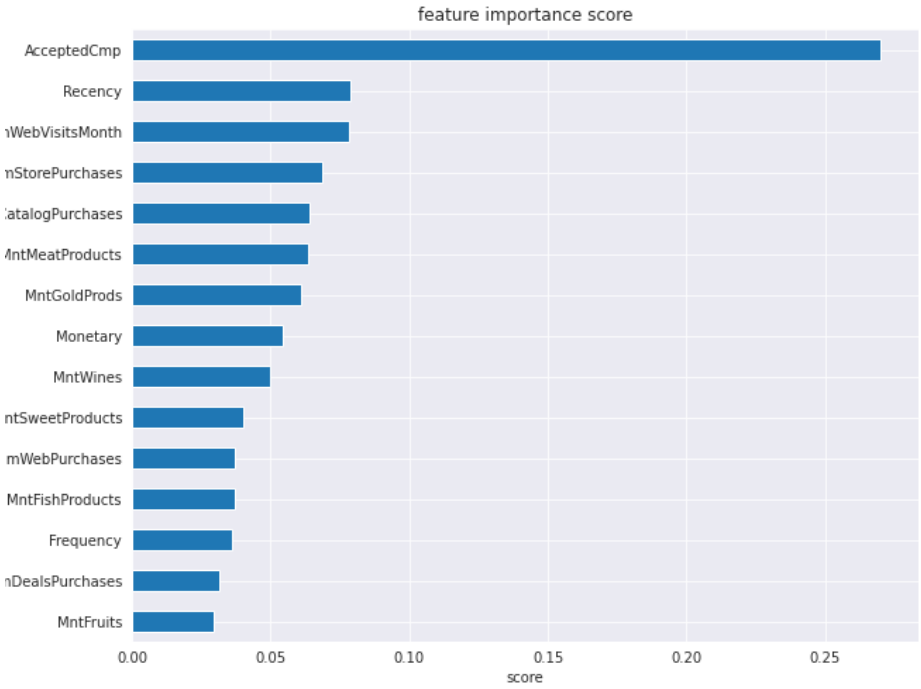
Figur 22: *Feature Importance Decision Tree* setelah *Hyperparameter Tuning*



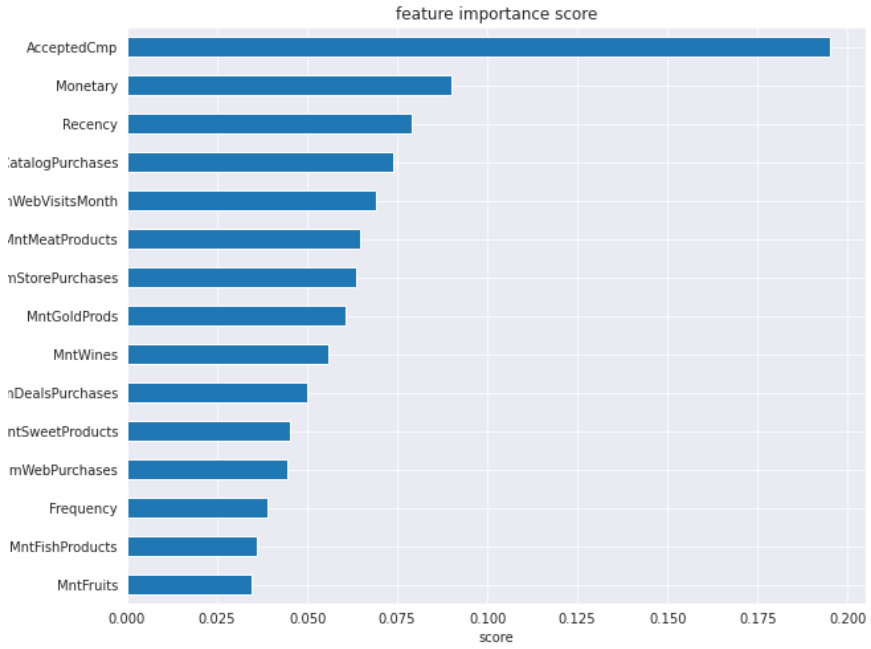
Figur 23: *Feature Importance Random Forest* sebelum *Hyperparameter Tuning*



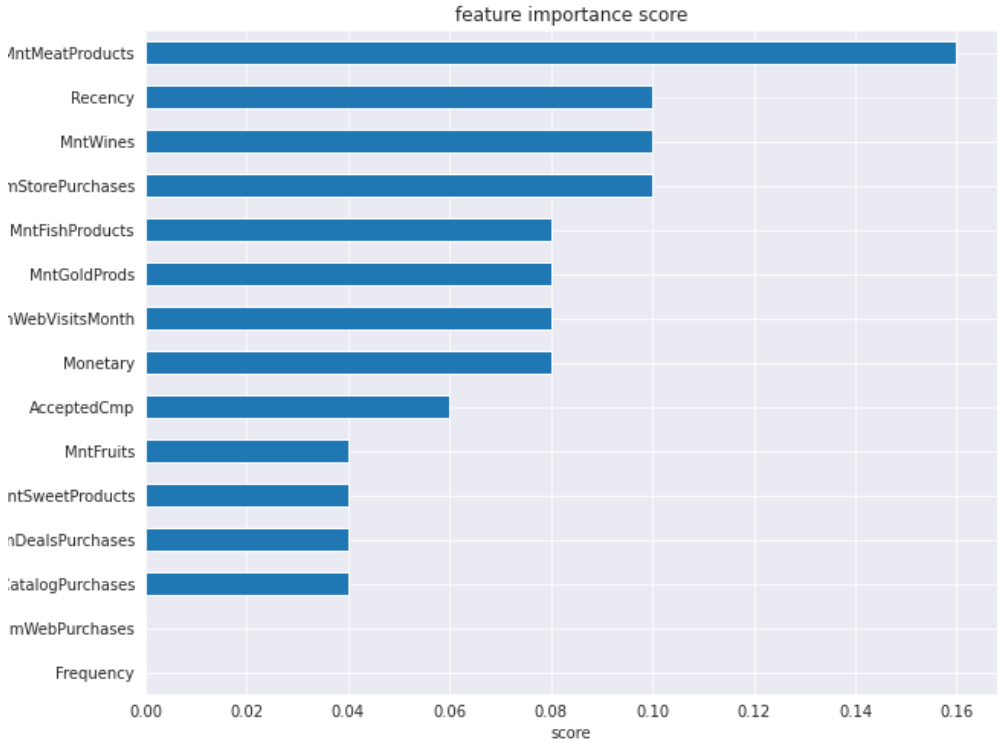
Figur 24: *Feature Importance Random Forest* setelah *Hyperparameter Tuning*



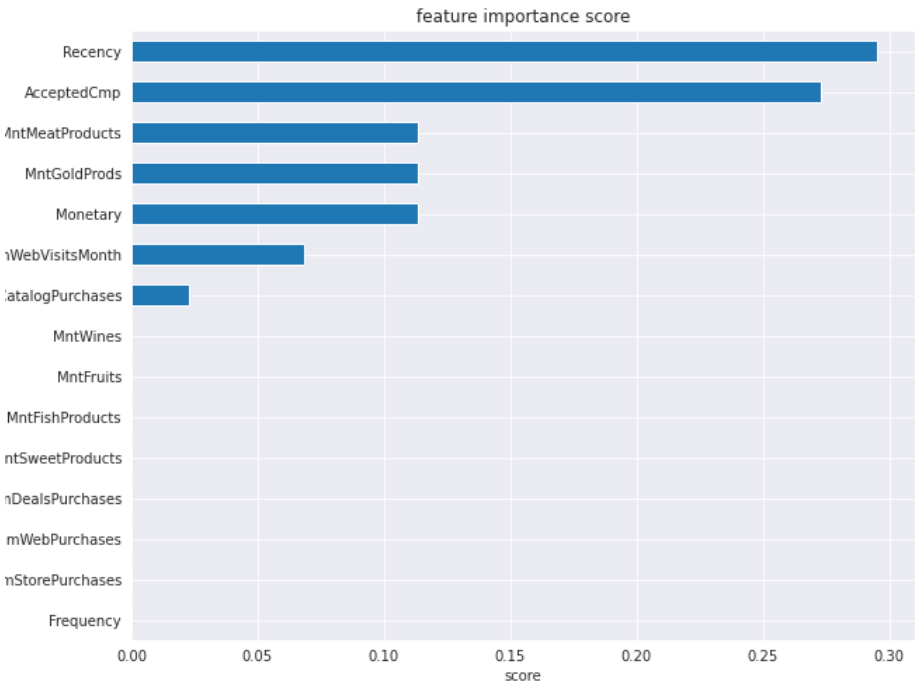
Figur 25: *Feature Importance XGBoost* sebelum *Hyperparameter Tuning*



Figur 26: *Feature Importance XGBoost* setelah *Hyperparameter Tuning*

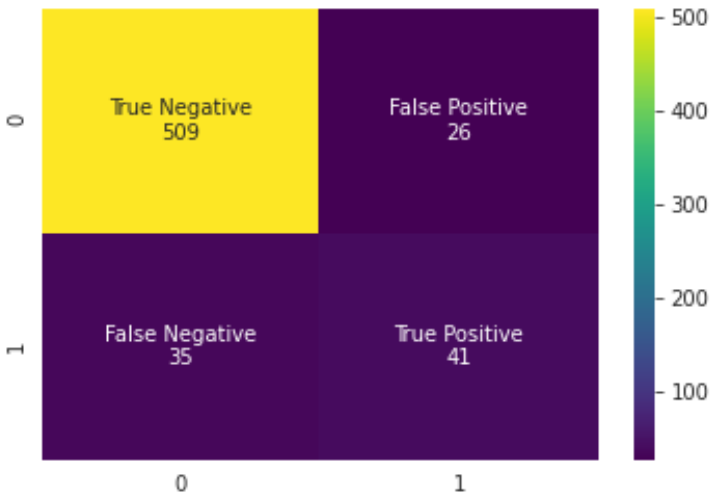


Figur 27: *Feature Importance AdaBoost* sebelum *Hyperparameter Tuning*



Figur 28: *Feature Importance AdaBoost* setelah *Hyperparameter Tuning*

Untuk menentukan model terbaik, *metrics* yang digunakan adalah *AUC* karena data yang digunakan *Imbalance*. *Modelling* yang memiliki *AUC* terbaik merupakan *AdaBoost* sebelum melakukan *Hyperparameter Tuning* dengan nilai *AUC* 0.75. Menurut hasil dari *AUC*, 75% penerima *campaign* dari hasil prediksi, benar. Melihat *Feature Importance* nya, empat *Feature Importance* tertinggi dari model *AdaBoost* merupakan *MntMeatProducts* (16%), *Recency* (10%), *MntWines* (10%), dan *NumStorePurchases* (10%). Hasil model dan *feature importance* tersebut cukup bagus dan dapat membantu untuk memberikan rekomendasi bisnis.



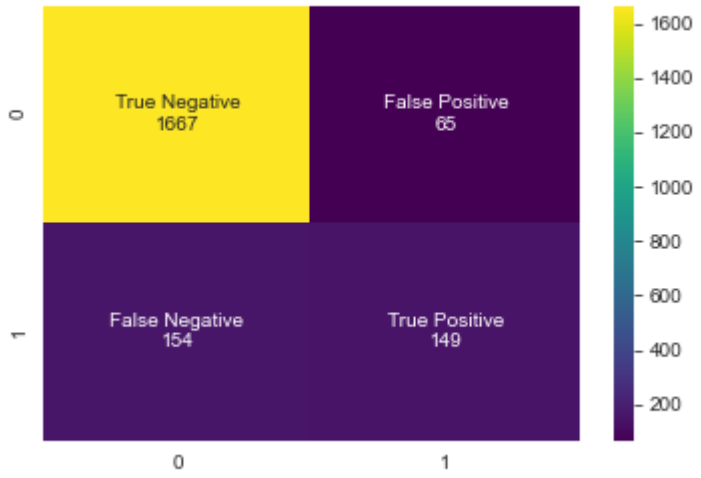
Figur 29: *Confusion Matrix data train*

Berdasarkan hasil *Confusion Matrix* yang didapatkan dari hasil *modeling* dengan data *train*, dapat dilihat bahwa *True Negative* berjumlah 509, *True Positive* berjumlah 41, *False Negative* berjumlah 35, dan *False Positive* berjumlah 26. Diprediksikan 544 customer akan merespon tidak terhadap campaign, akan tetapi 509 yang prediksinya benar dalam kategori ini. Dari 67 customer yang diprediksikan akan menerima campaign, 41 benar-benar merespon terhadap *campaign,* sementara 26 lainnya tidak benar.

Conclusion & Recommendations

Setelah melalui proses *pre-processing*, EDA, dan *modeling*, dapat disimpulkan bahwa model terbaik untuk data *marketing campaign* adalah *AdaBoost* dengan *AUC* 0.75. Dari *feature importance*, diketahui bahwa empat *Feature Importance* tertinggi dari model *AdaBoost* merupakan *MntMeatProducts, Recency, MntWines*, dan *NumStorePurchases*. Hasil *modelling* kemudian dapat digunakan untuk memprediksi *customer-customer* yang akan merespon *campaign* selanjutnya sehingga *campaign* tersebut dapat diarahkan ke *customer-customer* yang diprediksi akan menerima respon untuk menaikan *response rate*.

Model yang telah dibuat kemudian digunakan untuk memprediksi kembali *dataset* yang telah di *cleansing* secara menyeluruh untuk digunakan sebgai acuan untuk *marketing campaign* selanjutnya. Hasil dari prediksi tersebut adalah sebagai berikut:



Figur 30: *Confusion matrix full dataset*

Strategi *marketing campaign* selanjutnya adalah dengan mengirimkan *campaign* hanya kepada *customer-customer* yang diprediksi oleh model akan menerima *campaign* tersebut. Dengan begitu biaya yang digunakan untuk marketing campaign dapat ditekan dibandingkan dengan mengirimkan marketing ke seluruh customer.

Dengan Model		Tanpa Model	
campaign dikirim	214	campaign dikirim	2035
Menerima campaign	149	Menerima campaign	303

Tanpa Model			Dengan Model			Selisih		
Biaya campaign	$	6.105,00	Biaya campaign	$	642,00	$	-5.463,00
pendapatan campaign	$	3.333,00	pendapatan campaign	$	1.639,00	$	-1.694,00
profit	$	-2.772,00	profit	$	997,00	$	3.769,00
Response rate		15%	Response rate		70%		55%

Selain menekan biaya campaign sebesar $5463, response rate yang didapat juga menjadi naik secara signifikan yaitu sebanyak 55%. Cara penghitungan tabel diatas adalah sebagai berikut:

Biaya Campaign = Campaign dikirim ×$3
Pendapatan Campaign = Menerima campaign ×$11
Profit = Biaya Campaign - Pendapatan Campaign
Response rate = Menerima campaign/Campaign dikirim x 100%

Dari hasil perhitungan diatas dan juga *feature importance* yang didapatkan, beberapa rekomendasi tambahan yang dapat diberikan untuk tim bisnis adalah untuk menyesuaikan *marketing campaign* dengan *Recency* dan memberikan promo kepada pelanggan yang berbelanja *in-store*. Untuk menyesuaikan *marketing campaign* dengan *Recency, Guido's Market* dapat memberikan promo kepada pelanggan yang sudah lama tidak melakukan transaksi di Guido's Market untuk menarik pelanggan. Promo bagi pelanggan yang berbelanja *in-store* juga direkomendasikan karena dibandingkan dengan platform pembelian lainnya pembelian lewat toko merupakan jenis platform belanja dimana pelanggan paling sedikit merespon terhadap *campaign*. Selain itu, dapat diberikan juga promo untuk kategori barang *wine* dan daging.

Untuk kedepannya, tindak lanjut yang dapat dilakukan adalah memperbaiki model agar hasilnya dapat lebih baik lagi. Untuk sekarang, hasil AUC dengan nilai 0.75 merupakan yang tertinggi, akan tetapi idealnya nilai yang dipilih harus lebih tinggi lagi. Selain itu, dapat dibuat hitungan untuk ROI (*Return on Investment)* sebagai dasar dalam mengambil keputusan dan menentukan hasil akhirnya.