iFood CRM Data Analyst Case

Análise da campanha de marketing de uma empresa do setor de varejo de alimentos

Utilizei Python para realizar a análise e o *notebook*, com os comandos, está disponibilizado na minha página do *GitHub* e pode ser acessado por meio deste link

Autor: Eduardo Luís Bartholomay[1]

29 de Maio de 2020

[1] Mestrando em Economia Aplicada, Universidade Federal do Rio Grande do Sul.
E-mail: elbartholomay@gmail.com
Linkedin: linkedin.com/in/elbartholomay
Site pessoal: edubarth.github.io

SUMÁRIO

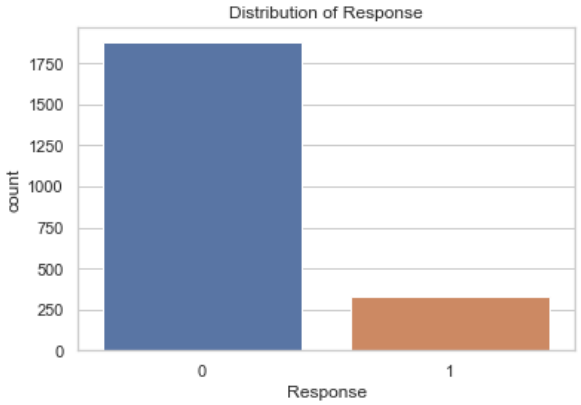
FIGURAS:

TABELAS:

1. DATA EXPLORATION

A base de dados inicial continha 2.240 clientes. Após a limpeza, remoção de nulos e *outliers*, a base final ficou com 2.208 clientes. Destes, 331 compraram o gadget oferecido e 1877 recusaram. No gráfico abaixo é possível visualizar a distribuição das respostas.

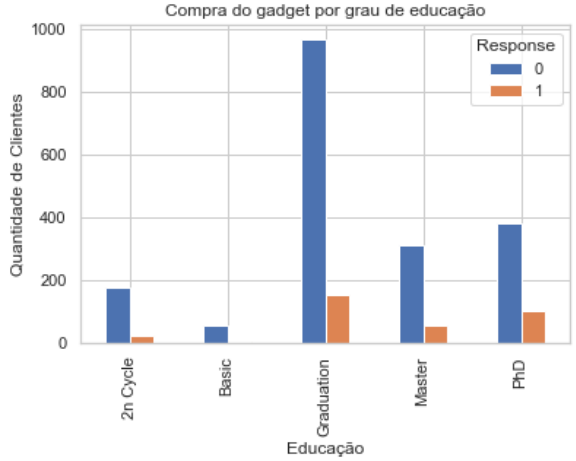
Figura 1 Distribuição das respostas.



Fonte: elaborado pelo autor.

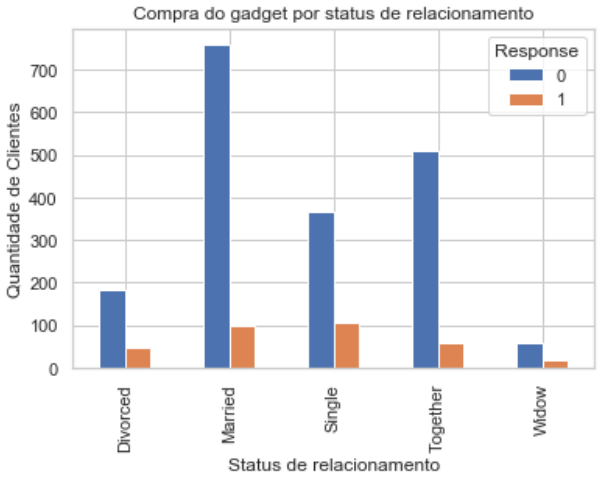
Nas figuras abaixo são apresentadas a diferença na quantidade de clientes que compraram e não compraram o *gadget*, para cada grau de escolaridade, figura 2, e tipo de relacionamento, figura 3.

Figura 2: Quantidade de clientes por educação.



Fonte: elaborado pelo autor.

Figura 3: Quantidade de clientes por status de relacionamento.

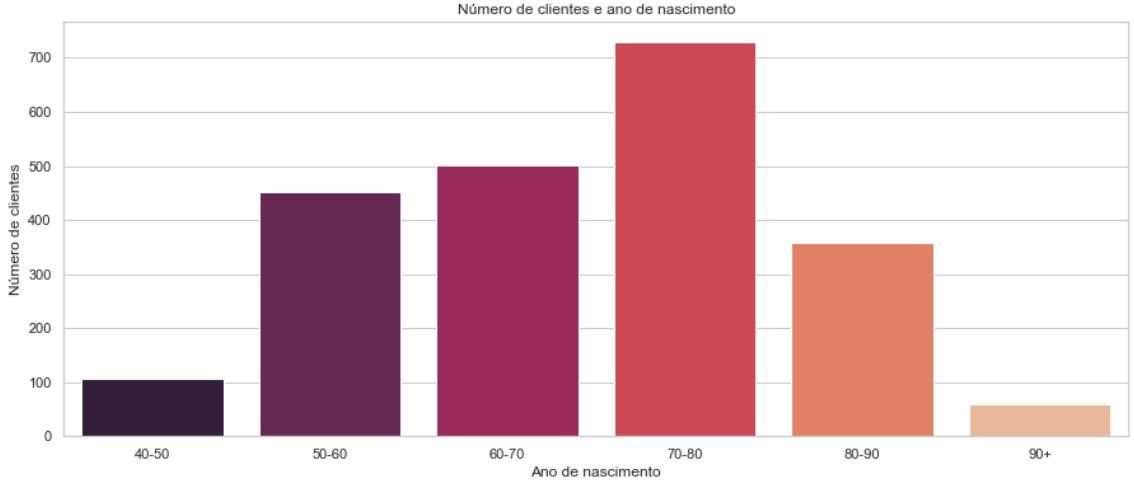


Fonte: elaborado pelo autor.

Por meio dessas distribuições, percebe-se que a os clientes que possuem graduação foram os que mais aceitaram e rejeitaram a oferta do *gadget*, entretanto ao analisar o status de relacionamento, oberava-se que os consumidores solteiros foram os que mais aceitaram a oferta, enquanto que os consumidores casados foram os que mais rejeitaram a oferta, indicando que pessoas casadas tendem a recusar mais a compra do *gadget*, mas lembrando que isto é apenas uma suposição, tendo como base uma análise gráfica, ou seja, necessitamos de mais aprofundamento para inferirmos algo.

No gráfico abaixo, podemos visualizar que a base de dados possuem uma predominância de indivíduos que nasceram entre 1970 e 1980. Esse padrão se repete quando estratificamos a mostra entre os que aceitaram e recusaram a oferta.

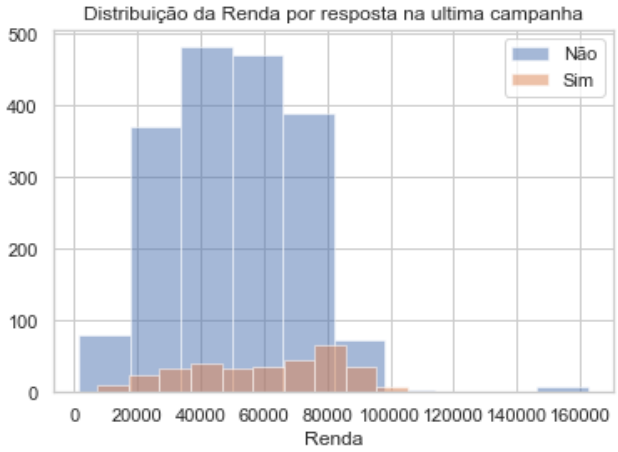
Figura 4: Distribuição do ano de nascimento



Fonte: elaborado pelo autor.

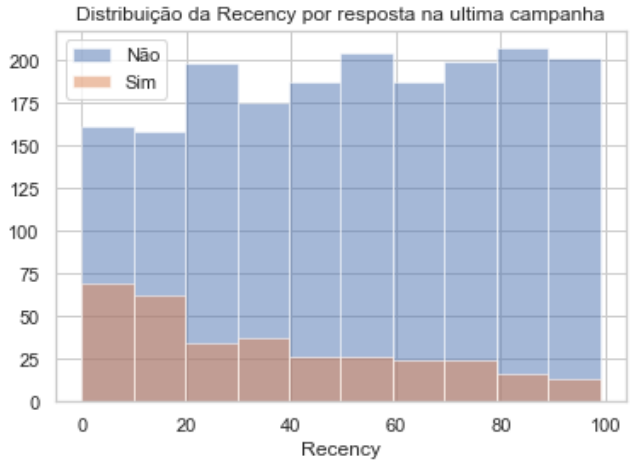
Nas figuras X e Y abaixo, verifica-se que os consumidores que mais aceitaram a proposta possuem uma renda de aproximadamente 80.000MU por ano, figura X, enquanto que os que rejeitaram possuem, em sua maioria, uma renda de aproximadamente 50.000MU anual. Já na figura Y, observa-se que os clientes que compraram algo recentemente no estabelecimento, aceitaram mais a oferta, por outro lado os que estão a muito tempo sem comprar algo, foram os que mais recusaram a oferta. Esse resultado irá corroborar com os encontrados na seção de segmentação dos consumidores.

Figura 5: Histograma da renda por resposta na última campanha.*:*



Fonte: elaborado pelo autor.

Figura 6: Histograma da *Recency* por resposta na última campanha.*:*



Fonte: elaborado pelo autor.

Para facilitar a análise para cada grupo, os que compraram o gadget (sucesso) e os que não compraram (insucesso), apresenta-se abaixo duas tabelas com as estatística

descritivas, sendo a Tabela X sobre o grupo de sucesso e a tabela Y sobre o grupo de insucesso.

Tabela 1: Estatística descritiva do grupo de sucesso.

Variável	Média	Máximo	Mínimo	Desvio Padrão
Income	60187,75	105471	7500	23231,6
Year_Birth	1969,41	1996	1943	12,3
Widow	0,05	1	0	0,23
Single	0,32	1	0	0,47
Together	0,18	1	0	0,39
Married	0,3	1	0	0,46
Divorced	0,15	1	0	0,35
Filhos	0,65	3	0	0,74
Recency	35,29	99	0	27,61
PhD	0,3	1	0	0,46
Master	0,17	1	0	0,38
Graduation	0,46	1	0	0,5
Basic	0,01	1	0	0,08
2n_Cycle	0,07	1	0	0,25
NumWebVisitsMonth	5,31	10	1	2,56
NumWebPurchases	5,07	11	0	2,57
NumStorePurchases	6,08	13	2	3,08
NumDealsPurchases	2,34	11	0	2,11
NumCatalogPurchases	4,19	11	0	3,12
MntWines	503,26	1492	1	429
MntSweetProducts	38,41	198	0	46,23
MntMeatProducts	295,01	981	1	287,52
MntGoldProds	60,76	241	0	56,75
MntFruits	37,85	193	0	45,88
MntFishProducts	51,37	250	0	61,14
Complain	0,01	1	0	0,09
AcceptedCmp5	0,27	1	0	0,45
AcceptedCmp4	0,19	1	0	0,39
AcceptedCmp3	0,23	1	0	0,42
AcceptedCmp2	0,06	1	0	0,24
AcceptedCmp1	0,24	1	0	0,43
total_spent	986,66	2525	17	720,8
total_purchases	17,68	34	4	6,88
Observações	**331**	**331**	**331**	**331**

Fonte: elaborado pelo autor.

Primeiramente, destaque-se a aprovação da quinta campanha de marketing, que foi a que obteve a melhor aceitação dos consumidores, 27 por cento, seguida pela primeira com 24 por cento. Por outro lado, a que obteve o pior desempenho foi a segunda campanha, com apenas 6 por cento de aceitação.

Tabela 2: Estatística descritiva do grupo de insucesso.

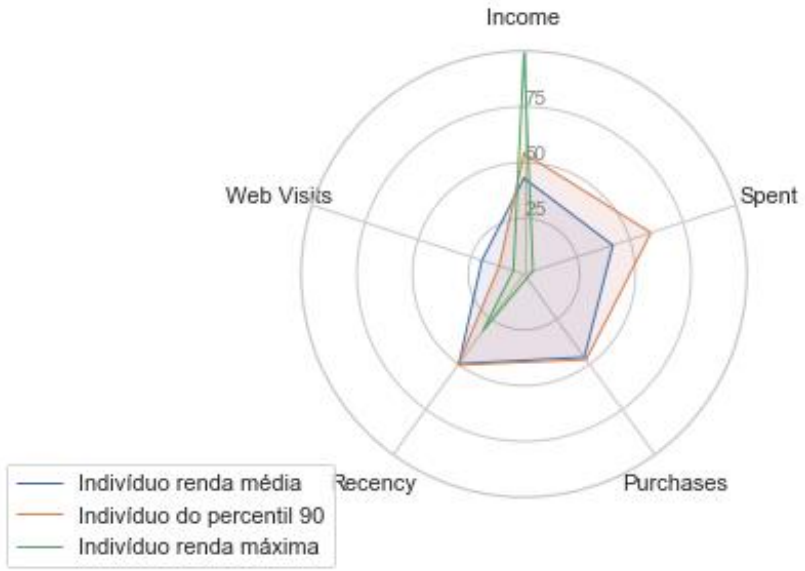
Variável	**Mean**	**Máximo**	**Minimo**	**Desvio Padrão**
Income	50489,69	162397	1730	20895,58
Year_Birth	1968,81	1996	1940	11,59
Widow	0,03	1	0	0,17
Single	0,19	1	0	0,4
Together	0,27	1	0	0,45
Married	0,4	1	0	0,49
Divorced	0,1	1	0	0,3
Filhos	1	3	0	0,74
Recency	51,48	99	0	28,49
PhD	0,2	1	0	0,4
Master	0,16	1	0	0,37
Graduation	0,51	1	0	0,5
Basic	0,03	1	0	0,16
2n_Cycle	0,09	1	0	0,29
NumWebVisitsMonth	5,32	20	0	2,4
NumWebPurchases	3,91	27	0	2,74
NumStorePurchases	5,76	13	0	3,28
NumDealsPurchases	2,32	15	0	1,89
NumCatalogPurchases	2,4	28	0	2,8
MntWines	270,31	1493	0	305,9
MntSweetProducts	25,06	262	0	39,82
MntMeatProducts	144,43	1725	0	203,11
MntGoldProds	40,79	321	0	49,96
MntFruits	24,26	199	0	38,2
MntFishProducts	35,09	259	0	52,98
Complain	0,01	1	0	0,09
total_spent	539,94	2525	5	553,28
total_purchases	14,39	44	0	7,71
Observações	**1877**	**1877**	**1877**	**1877**

Fonte: elaborado pelo autor.

Na primeira tabela, observa-se que as variáveis que se sobressaíram sobre o grupo de insucesso foram: renda, total gasto, total de compras, compras online, compras pelo

catálogo, gasto com vinho, gasto com doces, gasto com carne, gasto com fruta e com produtos “gold”, solteiro e “recency’. Estas variáveis indicam que os consumidores que aceitaram a oferta, em média, possuem uma renda maior do que os que rejeitaram, apresentam uma disposição maior a gastar e a comprar os produtos online e pelo catálogo, além de terem realizado alguma compra recente no estabelecimento.

Figura 7: *Radar-Chart* indivíduos.



Fonte: elaborado pelo autor.

Na figura acima podemos observar e comparar as características dos indivíduos que possuem a maior renda, a menor renda do percentil 90 e o indivíduo com a renda média. Notasse que o indivíduo com a maior renda não realiza tantas visitas online, não gasta e não compra muito, mas é um cliente recente, quando comparado com os outros dois indivíduos. Em relação aos indivíduos com renda média e o do percentil 90, eles possuem características semelhantes, diferenciando-se apenas no gasto e na quantidade de visitas online.

2. SEGMENTATION

2.1. RPS Score e K-Mean Clustering

Com o objetivo de encontrar os clientes de maior valor para a empresa, apliquei o RPS (Recency, Purchases e Spent) Score, o qual leva em consideração o gasto total do cliente, a quantidade de item adquiridos e quantos dias se passaram desde a sua última compra. Com base nisso, segmentando para o quantil 80, buscamos identificar o cliente que mais gasta e que fica menos dias sem realizar uma compra na loja, ou seja, o cliente mais valioso para a loja.

Nas tabelas abaixo, encontrasse, dividido em 4 grupos (vermelho = “Disengaged”, verde = “Star”, laranja = “Light”, Amarelo = “New”), a média do total gasto, o total de consumidores e a média de dias desde a última compra do cliente, para cada grupo.

Tabela 3: *RPS Score*.

Spent	**1**	**2**
2	1621	1575
1	355	362

Customers	**1**	**2**
2	87	357
1	350	1414

Recency	**1**	**2**
2	90	40
1	89	39

Fonte: elaborado pelo autor.

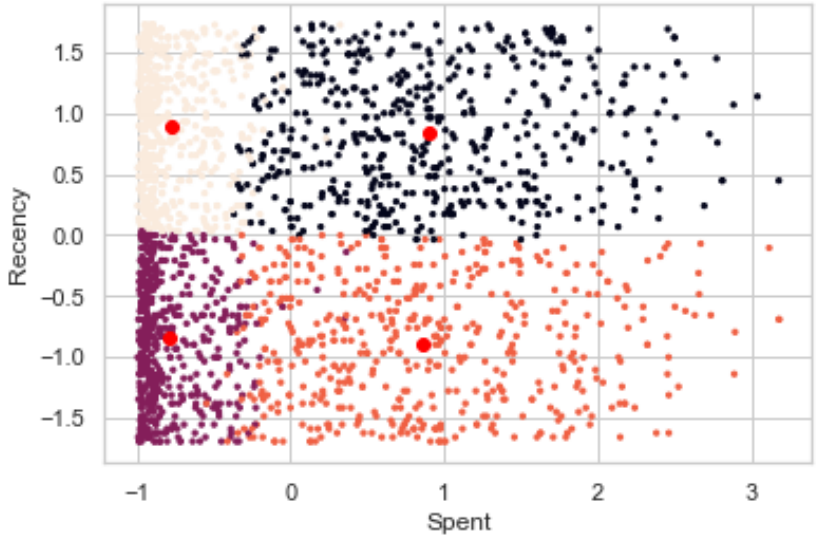
Os resultados indicam que existem 357 consumidores classificados como “Star”, ou seja, que são fiéis ao estabelecimento, e que estes consumidores gastaram em média 1.575MU na compra de produtos nos últimos 2 anos. Por outro lado, existem 87 clientes no grupo de “*Disengaged*”, que desistiram de consumir os produtos, por algum motivo não esclarecido e estes consumidores são os que possuem a maior média de gastos nos últimos dois anos. Portanto, entrar em contato com estes clientes seria uma medida para tentar traze-los de volta para o estabelecimento.

Além disso, foram identificados 350 consumidores no grupo “*Light*”, isto é, consumidores com alguma propensão a voltar a frequentar a loja, necessitando apenas algum incentivo para eles. Dessa maneira, enviar cupons para estes clientes pode ser uma ótima forma de trazer eles para o grupo “*New*” novamente.

Esse padrão podemos observar, também, quando aplicado o método de segmentação dos consumidores, como o *K-Means Clustering*. Na figura abaixo, quando

comparamos o total gasto nos últimos dois anos com os dias sem realizar uma compra, verificamos que existem 4 clusters de consumidores, um que gasta muito porem está a muitos dias sem comprar, o qual poderíamos colocá-lo no grupo de "*Disengaged*", outro que gasta muito e está a poucos dias sem comprar, que seria nossos consumidores "*Star*", ou seja, os que possuem o maior valor para o estabelecimento. Também encontramos os que estão a muitos dias sem comprar e gastaram pouco, ou seja, eram clientes novos que desistiram de voltar para a loja, portanto, novamente, ofertar cupons para estes clientes poderá torná-los em clientes novos, e entrar para o grupo "*New*", que são os que gastaram pouco e estão a poucos dias sem comprar algo.

Figura 8: *K-Means Cluestering, Recency* e *Total Spent*.



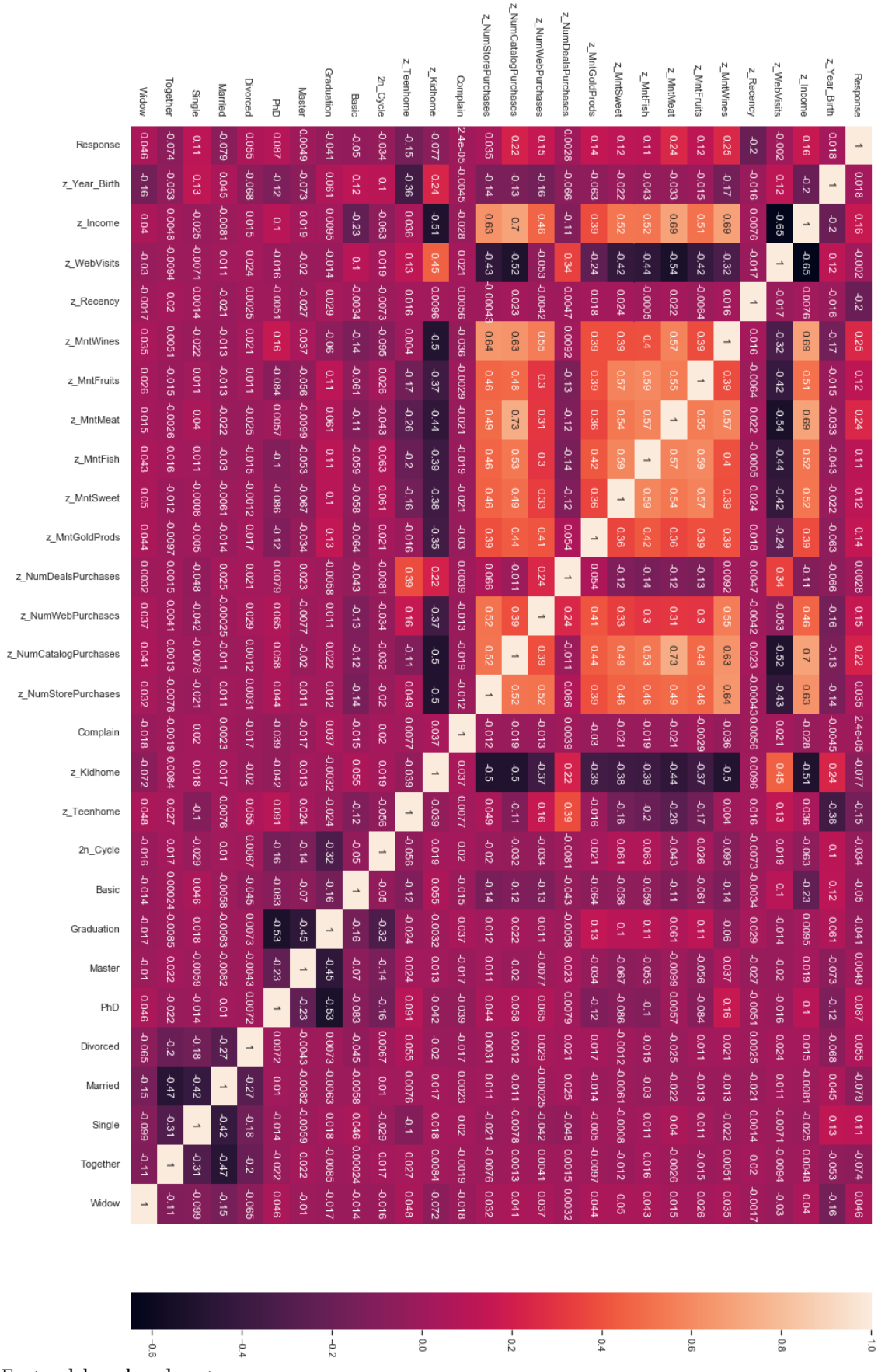
Fonte: elaborado pelo autor.

3. CLASSIFICATION MODEL

3.1. Recursive Feature Elimination (RFE)

Primeiramente, realizei a padronização das variáveis contínuas, após iniciei a análise verificando a relevância de cada variável para explicar a resposta do cliente. Portanto, na figura X abaixo, é apresentado o gráfico de correlação entre todas as variáveis. Contudo, não observamos alta correlação com a variável desejada, sendo que a correlação mais alta, em módulo, foi de 0,25 com a variável "MntWine", que indica o gasto com vinho. Verificando para todas as variáveis, as que obtiveram o maior valor, 0,73, de correlação foi entre "MntWine" e "NumCatalogPurchase", a qual indica a quantidade de compras feita pelo catálogo.

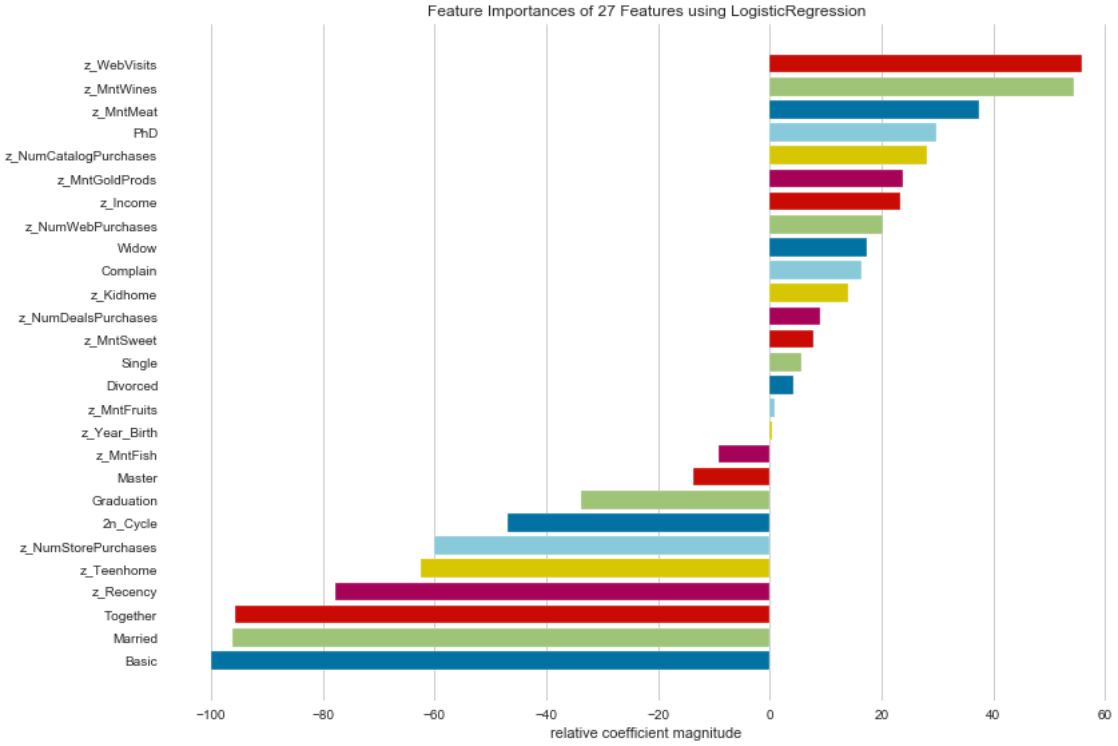
Figura 9: Gráfico de correlação.



Fonte: elaborado pelo autor.

Visando analisar a importância de cada variável em explicar a variável *target, "Response"*, apliquei também o método de Recursive Feature Elimination, e por meio da figura X, podemos ver que as variáveis "z_*Year_Birth*" e "z_*MntFruits*", possuem pouca relevância. Por conta disso, optei por retirá-la do modelo, afim de melhor a precisão.

Figura 10: *Recursive Feature Elimination*.



Fonte: elaborado pelo autor.

3.2. Aplicando modelos de Machine Learning

Tendo definido as variáveis que serão utilizadas no modelo, podemos seguir para a aplicação dos mesmos. Primeiramente, dividimos a amostra em treino, 75%, e teste, 25%, com o objetivo de avaliar o desempenho dos modelos.

Os modelos que irei aplicar e comparar serão os de Regressão Logística, de *Random Forest* e de *Decision Tree*. Por conta do fato do *Random Forest* ser uma evolução do *Decision Tree*, é esperado que o mesmo obtenha um desempenho melhor, entretanto o mesmo não podemos afirmar quando comparar com a Regressão Logística.

Na tabela X abaixo, podemos observar algumas métricas dos desempenhos dos modelos. O Score Train, mede o quão preciso o modelo foi em prever na amostra de treino, sendo que o modelo que melhor desempenhou neste quesito foi o de Decision Tree, com 99% dos resultados previstos. Já o Score Test, verifica a previsão fora da

amostra de treino, na amostra de teste, e é aqui que verificamos se quando é dado um banco de dados diferente para o modelo ele conseguirá ter o mesmo desempenho quando no treino. Observamos que a Regressão Logística foi a que obteve o melhor Score Test, com 89% dos resultados previstos. Outra métrica importante é a do *Mean Absolute Error*, o qual basicamente mede a distância do valor previsto para o valor verdadeiro, e considera essa distância como o erro do modelo, portanto quanto menor o MAE, melhor será para o modelo. Sendo assim, quem se saiu melhor no quesito MAE, novamente, foi o modelo de Regressão Logística. Por fim o critério ROC AUC é a probabilidade do modelo obter verdadeiro-positivo contra falso-positivo, e neste quesito o modelo de *Random Forest* se saiu melhor.

Tabela 4: Métricas de desempenho.

Avaliação	Logit	Random Forest	Decision Tree
Score Train	0,87	0,9	0,99
Score Test	0,89	0,87	0,82
MAE	0,11	0,13	0,18
ROC AUC	0,85	0,87	0,68

Fonte: elaborado pelo autor.

Com estes resultados já podemos descartar o modelo de *Decision Tree*. Resta agora comparar os resultados monetários entre Regressão Logística, modelo 1, e *Random Forest*, modelo 2. Para isso, iremos analisar a *Confusion Matrix* de ambos, a qual representa em uma tabela a quantidade de falsos-positivos, verdadeiros-positivos, falsos-negativos e falsos-positivos.

Figura 11: *Confusion Matrix*.

Fonte: elaborado pelo autor.

Tendo em vista de que cada ligação tem um custo de 3MU, cada pessoa que compra o gadget compra por 11MU, portanto por cada aceito existe um lucro de 8MU. Todos clientes que foram previstos com 1s serão contatados, logo, no modelo de regressão logística, para a amostra de teste, serão 37 pessoas contatadas, enquanto que no modelo de *Random Forest*, serão apenas 9.

Dessas pessoas contatadas, no modelo 1, 24 consumidores realmente aceitaram a oferta e compraram o *gadget*, tornando em uma taxa de sucesso de 65 por cento e um lucro de 153MU. Já no modelo 2, 6 consumidores aceitaram, resultando uma taxa de sucesso de 67 por cento e um lucro de 39MU.

Olhando pela ótica das pessoas que rejeitaram a compra do *gadget*, o primeiro modelo previu que 515 clientes recusariam, enquanto que o segundo modelo previu 543 recusariam. O resultado para o modelo 1 foi de 465 clientes realmente recusariam, o que resulta em uma taxa de acerto de 90 por cento e evitando uma perda de 1.395MU, porém 50 clientes teriam aceitado a oferta, o que resultaria em 400MU de receita, logo a perda liquida seria de 995MU, caso todos estes clientes fossem contatados.

Analisando para o modelo 2, vemos que o resultado mostra que 475 clientes realmente recusariam, ocasionando em uma taxa de acerto de 87 por cento e evitando uma perda de 1.425MU, entretanto 68 consumidores teriam aceitado a oferta, o que geraria uma receita de 544MU, com isso a perda liquida seria de 881MU, caso todos estes clientes fossem contatados.

4. CONCLUSÃO

Por fim, podemos concluir que o modelo de Regressão Logística se mostrou melhor para prever as pessoas que recusariam a oferta, e que o modelo de *Random Forest* obteve uma taxa de sucesso maior, se tratando em acertar os clientes que comprariam o *gadget*. Dessa maneira, a escolha do modelo dependeria do quanto a empresa é avessa ao risco, caso sua aversão ao risco seja alta, o modelo de *Random Forest* é o mais recomendado pois é capaz de identificar as pessoas que devem ser contatadas pela equipe de Marketing com mais eficiência. Porém, se a empresa for mais propensa ao risco, o modelo de Regressão Logística seria o recomendado, pois ele gera um lucro maior em troca de gerar mais falsos-positivos.