



Projeto Marketing Analytics

Otimização da nova campanha de Marketing do
iFood - (iFood DAd test)

Lucas Reis

lucas.fraga@ifood.com.br



Contexto atual

Para iniciar o projeto, fez-se importante entender o contexto atual da empresa e da área em que o projeto será desenvolvido, de forma a auxiliar na proposta de uma solução que melhor se adeque às necessidades e à maturidade analítica da área.



O time de Marketing é desafiado a utilizar seu budget anual de forma mais sábia.



Para isso, um pequeno time de Cientistas de Dados é contratado para desenvolver um modelo preditivo capaz de otimizar a taxa de sucesso da próxima campanha



Assim, uma campanha piloto é criada, onde 2240 clientes foram aleatoriamente selecionados e contatados para a criação da base de dados que dará origem ao modelo



Definições do Projeto

A fim de alinhar a comunicação e garantir que todas as partes envolvidas no projeto estão cientes do caminho a ser percorrido, utilizamos a definição dos 4P's de Data, que consiste na especificação do **problema** que gerou a dor, do **potencial** de ganho do projeto, do **produto** a ser desenvolvido e na **proposta** de ação a ser tomada.

Problema:

A campanha piloto de marketing gerou um prejuízo de 3.046MU e possui uma taxa de sucesso relativamente baixa (15%).

Potencial:

Com os dados obtidos com os clientes durante a campanha piloto, é possível criar uma estratégia de segmentação mais assertiva e aumentar a receita proveniente da nova campanha.

Produto:

Modelo preditivo que indique quais clientes deverão ser impactados pela nova campanha.

Proposta:

Conduzir a nova campanha de Marketing priorizando os públicos com maior propensão a compra de nossos produtos, de forma a aumentar a taxa de sucesso e consequentemente a receita da campanha.



Como performaram as campanhas anteriores?

Antes de tentar prever o que irá acontecer na sexta campanha, é importante entender o que aconteceu no passado, isso trará insights sobre o caminho que estamos seguindo e sobre uma eventual necessidade de mudança de rota.

"Para a seguinte análise, iremos assumir que todas as pessoas contatadas pela campanha piloto, já eram clientes quando houveram as veiculações das campanhas passadas."



Campanha 1



Campanha 2



Campanha 3



Campanha 4



Campanha 5

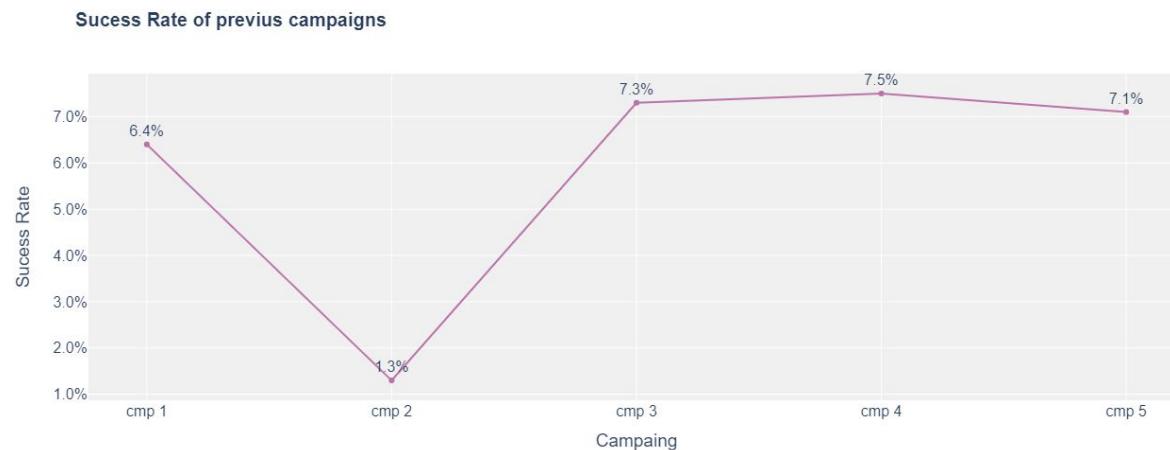


Como performaram as campanhas anteriores?

Primeiramente buscamos entender qual foi a taxa de sucesso das campanhas anteriores entre os clientes contatados.

No gráfico abaixo podemos analisar que todas as campanhas, tiveram taxas de sucesso menores do que 7,5%, o que é menor do que a taxa de sucesso da campanha piloto, que possui uma taxa de sucesso de 15% e mesmo assim não se mostrou assertiva o suficiente para gerar lucro.

Dessa forma, caso as campanhas anteriores tenham seguido os padrões de custo e receita por cliente da campanha piloto, fica claro que a falta de estratégia para a segmentação de clientes para a veiculação de campanhas é um problema a ser corrigido.





Quem são as pessoas que aceitaram a campanha piloto?

Como a campanha piloto foi criada com o intuito de gerar o modelo de segmentação da sexta campanha, através de uma análise descritiva, inicialmente iremos focar nossos esforços em entender o perfil dos clientes que aceitaram a oferta da campanha piloto.



Campanha Piloto

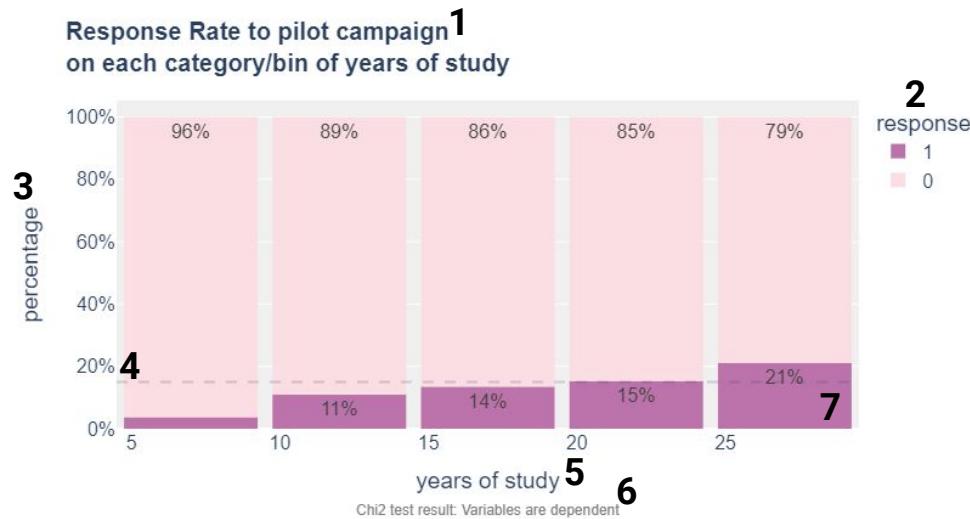


Entendendo os gráficos



Para essa análise, foi escolhido o gráfico de barras totalmente empilhadas, a fim de explicitar as diferentes proporções de clientes que aceitaram a oferta da campanha piloto para cada categoria de cada variável (variáveis numéricas foram categorizadas para auxiliar a visualização), além disso, alguns clientes foram retirados da análise, por terem comportamentos considerados atípicos.

O exemplo a seguir ilustra o formato das análises:

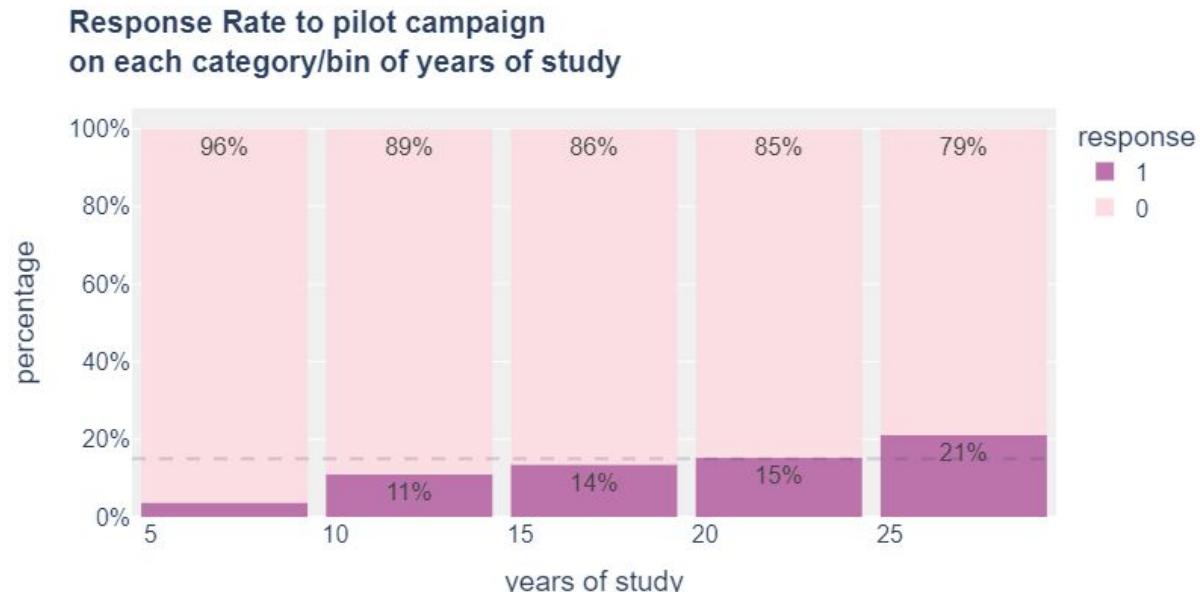


1. Título do gráfico: Indica a variável que estamos analisando junto à taxa de aceite da campanha piloto;
2. Legenda: 1(roxo) para clientes que aceitaram a oferta da campanha piloto e 0(rosa) para clientes que não;
3. Eixo Y: Indica o percentual para cada grupo analisado;
4. Linha média: Indica a taxa de sucesso da campanha piloto para toda a base, que é de 15%;
5. Eixo X: Indica as categorias ou agrupamentos da variável que estamos analisando, no exemplo temos 5 agrupamentos para anos de estudo, de 5 a 10 anos, de 10 a 15 anos, de 15 a 20 anos, de 20 a 25 anos e de 25 a 30 anos;
6. Resultado do teste Qui-Quadrado: Indica se a mudança na proporção da taxa de sucesso da campanha para cada uma das categorias da variável analisada possui significância estatística suficiente ou não, quando sim, dizemos que as variáveis são dependentes entre si, e quanto não, dizemos que são independentes;
7. Interpretação dos resultados: No exemplo, temos que 21% dos clientes com 25 a 30 anos de estudo aceitaram a oferta da campanha piloto.



Ter estudado por mais tempo impacta a propensão do cliente de aceitar a oferta da campanha?

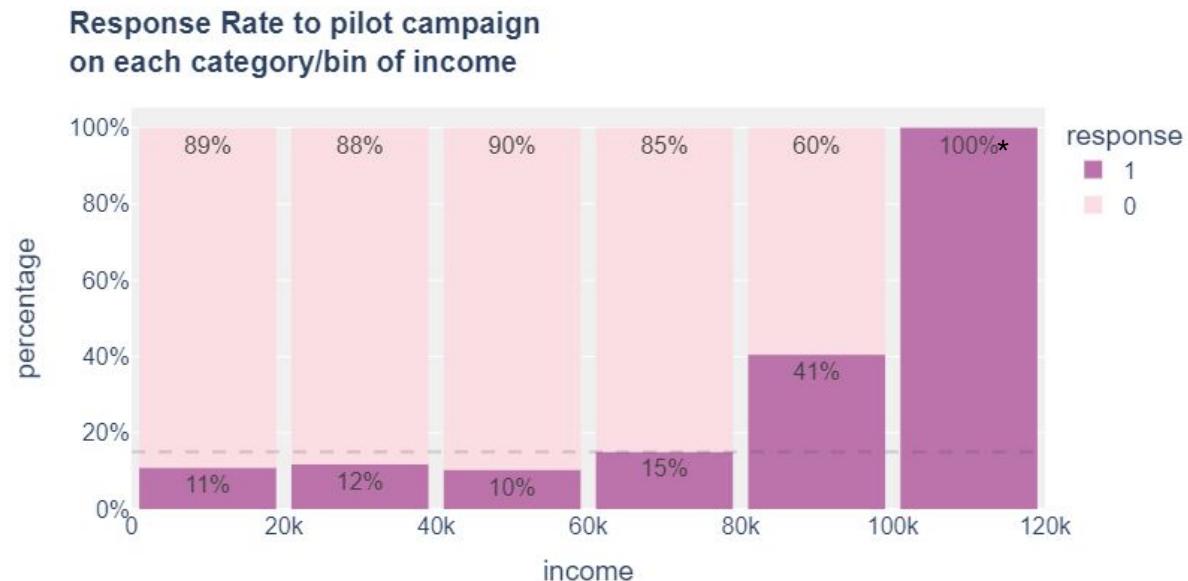
Sim! é possível analisar que conforme os clientes possuem mais tempo de estudo, as taxas de aceite da campanha vão subindo gradativamente, até chegar a um patamar de 21% para pessoas com mais de 25 anos estudo (equivalente a doutores).





Cientes de alta renda são mais propensos a aceitar a campanha?

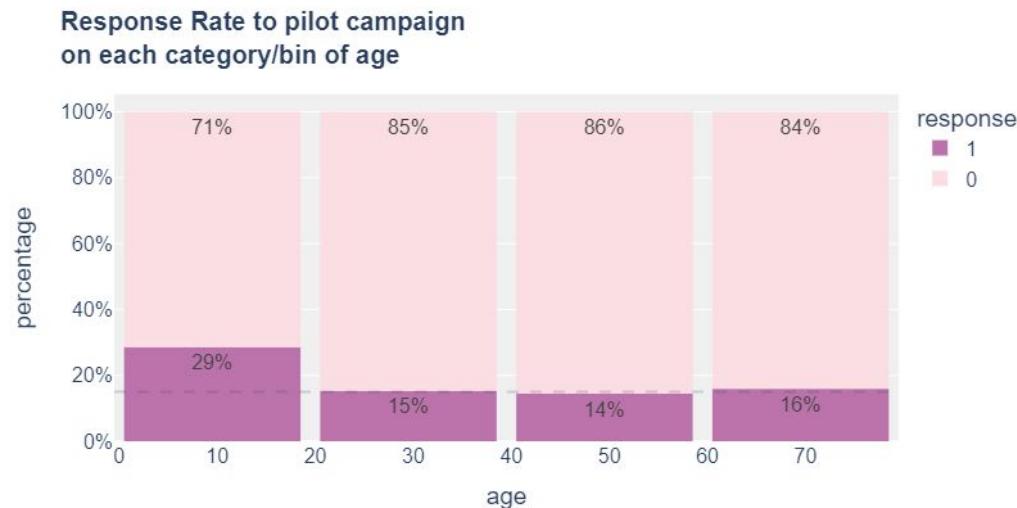
Sim, clientes que possuem rendas mais altas (acima de \$80k/ano) são clientes que proporcionalmente aceitaram mais a oferta da campanha.





Alguma faixa etária foi mais impactada do que as outras pela campanha piloto?

Apesar do gráfico indicar que para a faixa etária de até 20 anos, a taxa de sucesso da campanha foi bem superior às outras, o teste qui-quadrado indicou que as variáveis são independentes, o que acontece é que apesar deste grupo mais jovens possuir uma alta taxa de aceite da campanha, se trata de um grupo que possui poucas pessoas, enquanto para os outros grupos, que possuem uma parte mais significativa dos clientes, a taxa de sucesso se manteve estável e próximo da taxa de sucesso geral da campanha.

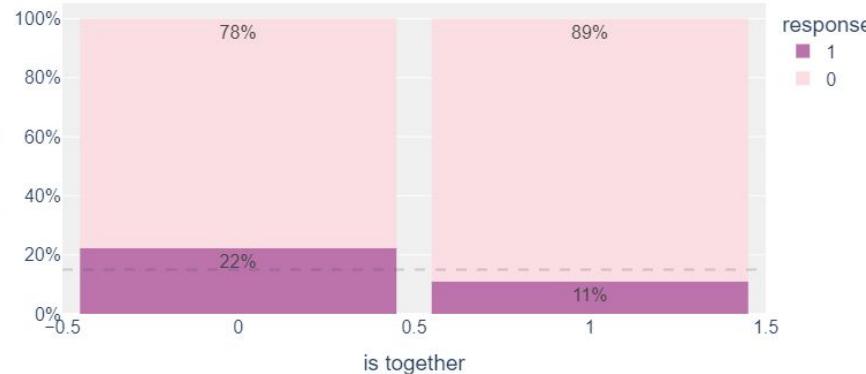




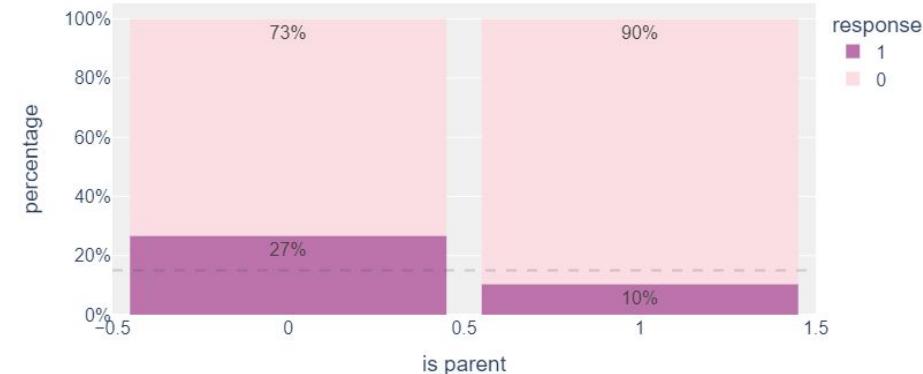
O estado civil ou o fato de ter filho(s) se mostrou relevante para o aceite da campanha piloto?

Segundo os gráficos abaixo, podemos afirmar que sim! Clientes que estão sozinhos ou não são pais tiveram maiores taxas de aceite da campanha.

Response Rate to pilot campaign on each category/bin of is together



Response Rate to pilot campaign on each category/bin of is parent

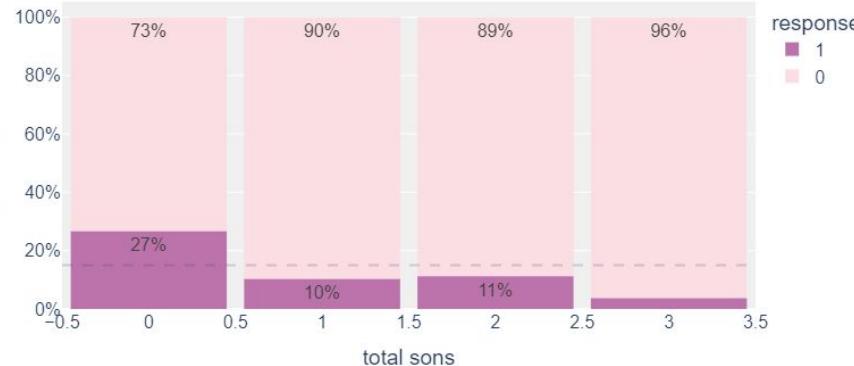




E a quantidade de filhos e o tamanho da família, também foram relevantes?

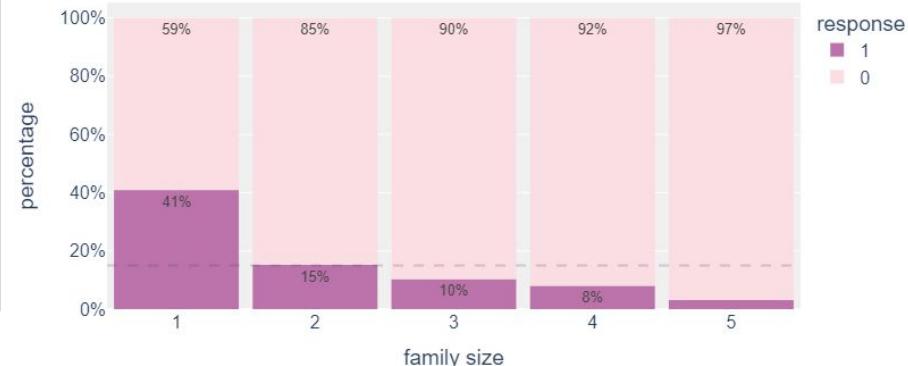
Aqui também podemos afirmar que sim, são relevantes! Clientes com mais filhos e por consequência maiores famílias, tendem a ter uma taxa de aceite da campanha mais baixa, o que está em linha com o slide anterior.

Response Rate to pilot campaign
on each category/bin of total sons



Chi2 test result: Variables are dependent

Response Rate to pilot campaign
on each category/bin of family size

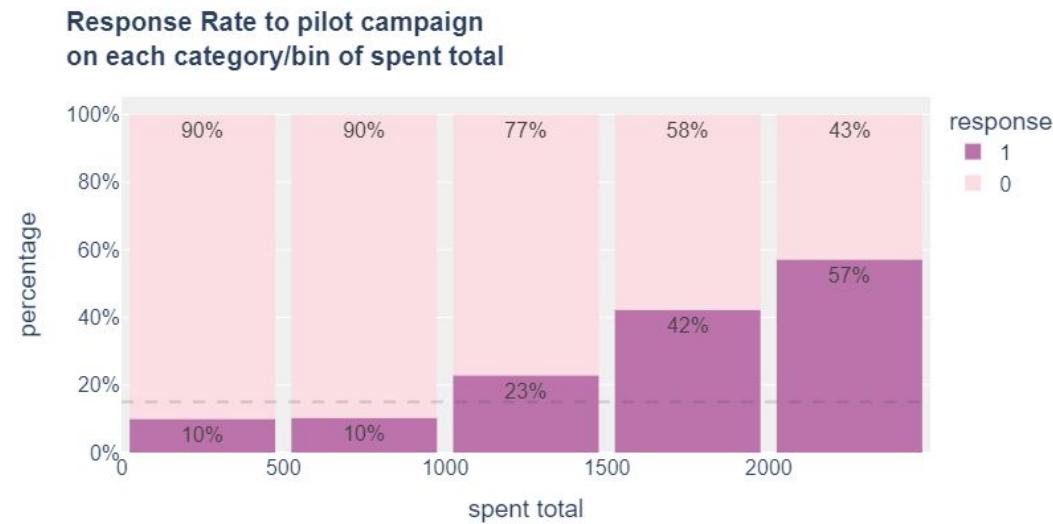


Chi2 test result: Variables are dependent



Cientes que gastaram mais conosco no passado, aceitaram mais a oferta da campanha piloto?

Como era de se esperar, para grupos de clientes que gastaram mais conosco no passado a campanha teve uma alta taxa de sucesso, chegando a 57% para clientes que gastaram acima de \$2000, enquanto para clientes que gastaram até \$1000 a taxa de sucesso atingiu somente 10%.

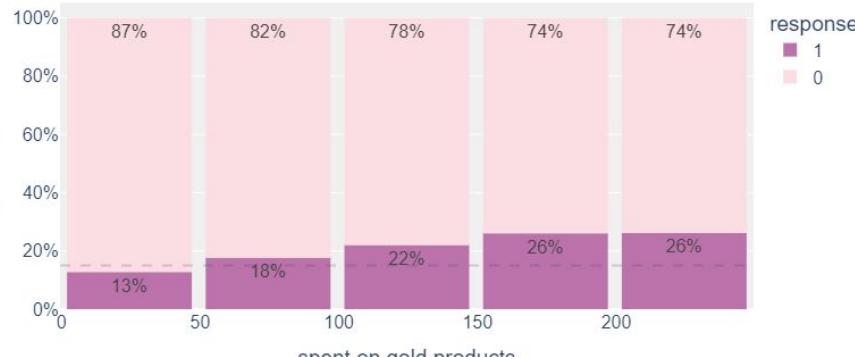




E para os gastos em produtos das categorias “Gold” e “Regular” separadamente?

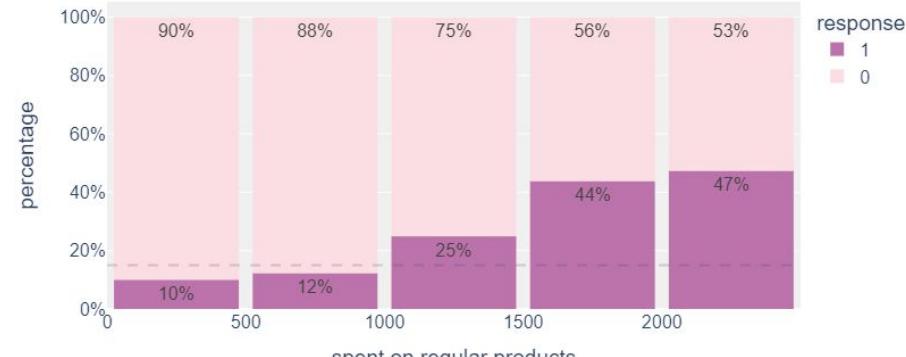
Nossos produtos são divididos em duas categorias “Gold” e “Regular”, para ambas as categorias, clientes que tiveram maior gasto no passado, também tiveram maior propensão a aceitar a campanha, porém para clientes da categoria “Regular” a campanha se mostrou mais efetiva para clientes que gastaram mais, chegando a uma taxa de sucesso superior a 40% para os grupos de clientes que gastaram mais de \$1500 na categoria.

Response Rate to pilot campaign
on each category/bin of spent on gold products



Chi2 test result: Variables are dependent

Response Rate to pilot campaign
on each category/bin of spent on regular products



Chi2 test result: Variables are dependent

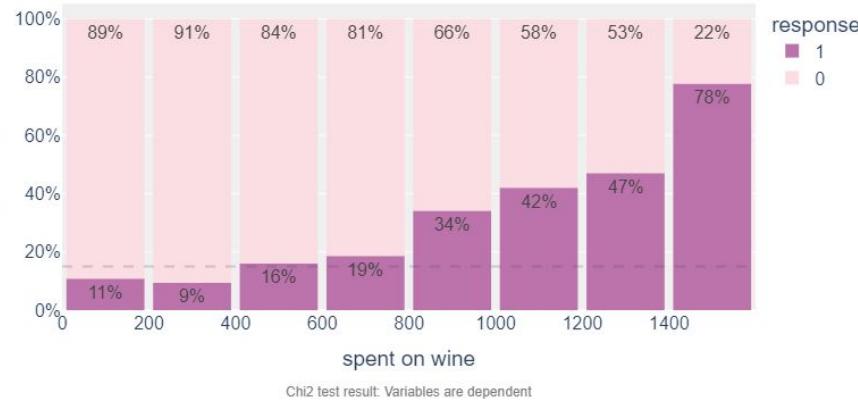


E se olharmos para o valor gasto em cada tipo de produto?

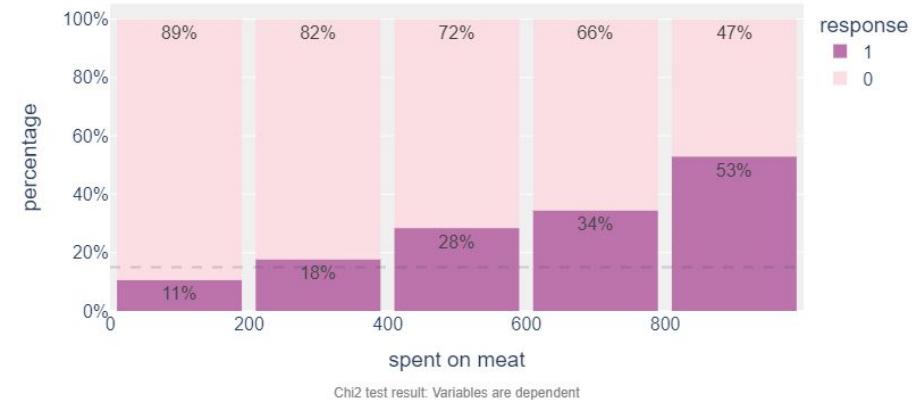
Além das categorias, os produtos também são divididos em tipos, que são vinhos, carnes, frutas, peixes e doces.

Ao olharmos para os valores gastos em vinhos e carnes, encontramos uma alta taxa de aceitação da campanha piloto entre clientes com maiores gasto, chegando até a 78% para clientes que gastaram mais de 1400\$ em vinhos e 53% para clientes que gastaram acima de \$800 em carnes.

Response Rate to pilot campaign
on each category/bin of spent on wine



Response Rate to pilot campaign
on each category/bin of spent on meat

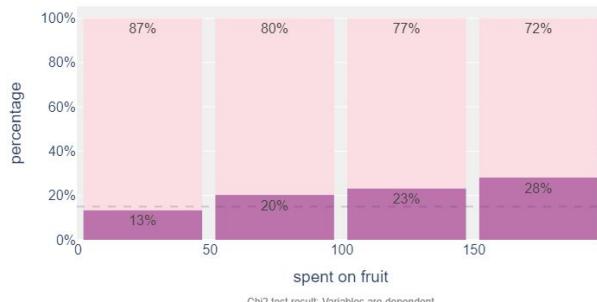




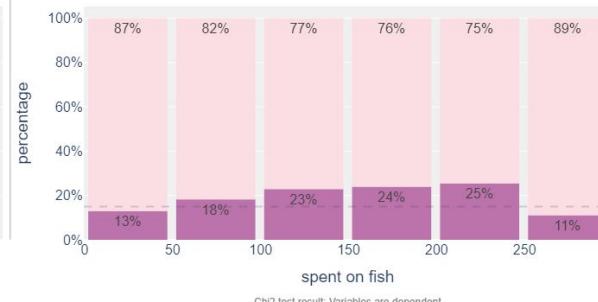
E se olharmos para o valor gasto em cada tipo de produto?

Quando analisamos os gastos em frutas, peixes e doces, também é possível identificar que para algumas faixas de valores gastos a taxa de aceitação da campanha é mais elevada do que para outras, apesar que de forma menos agressiva do que os dois exemplos anteriores (vinhos e carnes), podemos sim afirmar que o valor gastos nestes produtos impacta a aceitação da oferta feita pela campanha piloto.

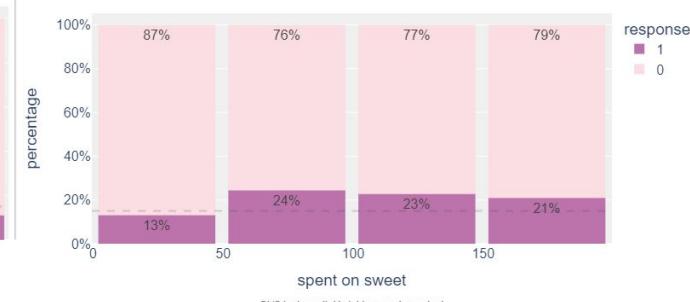
Response Rate to pilot campaign
on each category/bin of spent on fruit



Response Rate to pilot campaign
on each category/bin of spent on fish



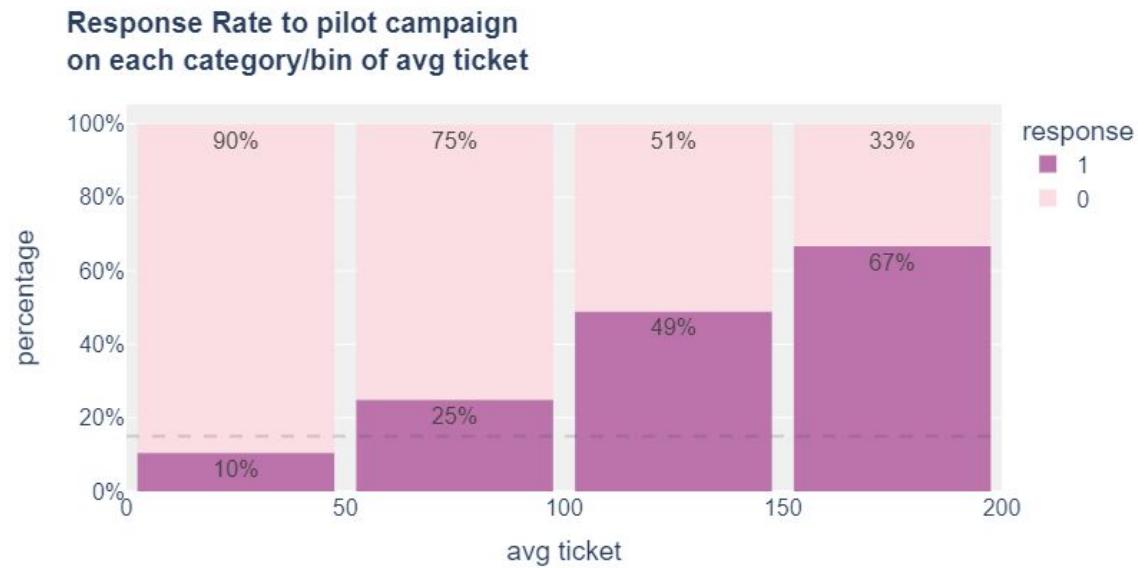
Response Rate to pilot campaign
on each category/bin of spent on sweet





O valor do ticket médio, se mostrou impactante?

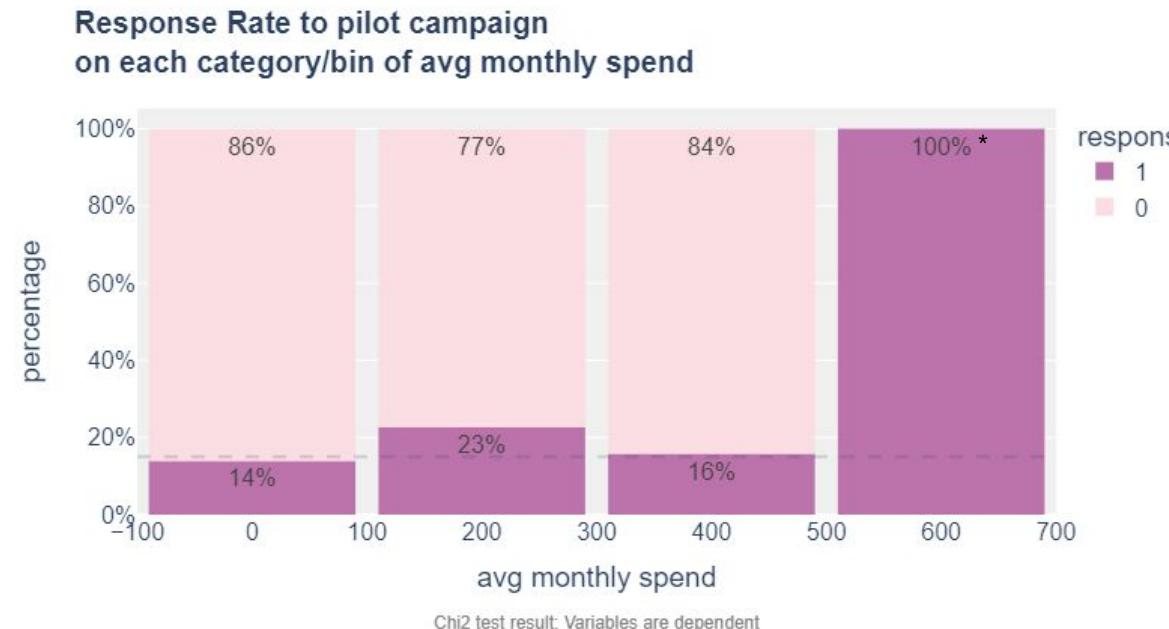
Sim! clientes que possuem tickets médio mais altos aceitaram mais a oferta da campanha piloto, onde para o grupo de clientes cujo o ticket médio é menor de \$50 a taxa de aceite da campanha piloto foi de 10%, percentual menor que a média geral da campanha.





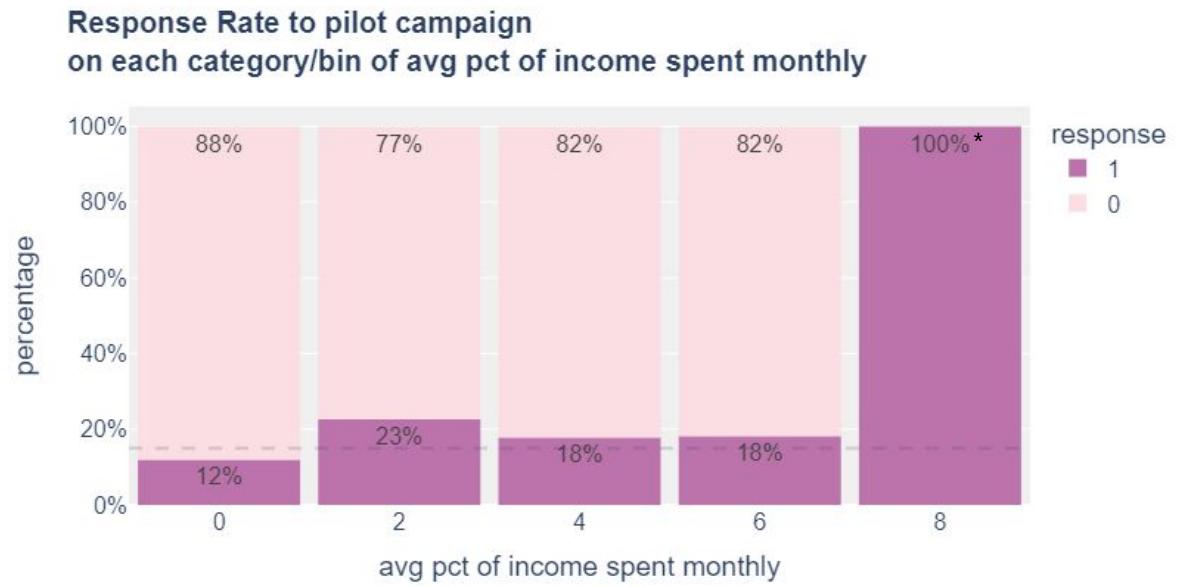
E o valor médio gasto no mês, se mostrou impactante?

Apenas de haver uma grande variação de taxa de aceite da campanha para os grupos, o teste qui-quadrado indicou que as variáveis são dependentes entre si, logo, apesar de visualmente os grupos não apresentarem certa linearidade entre as variáveis, sim, podemos afirmar que o valor médio gasto no mês foi impactante para a taxa de sucesso.



O percentual da renda mensal gasta conosco impactou a taxa de sucesso da campanha?

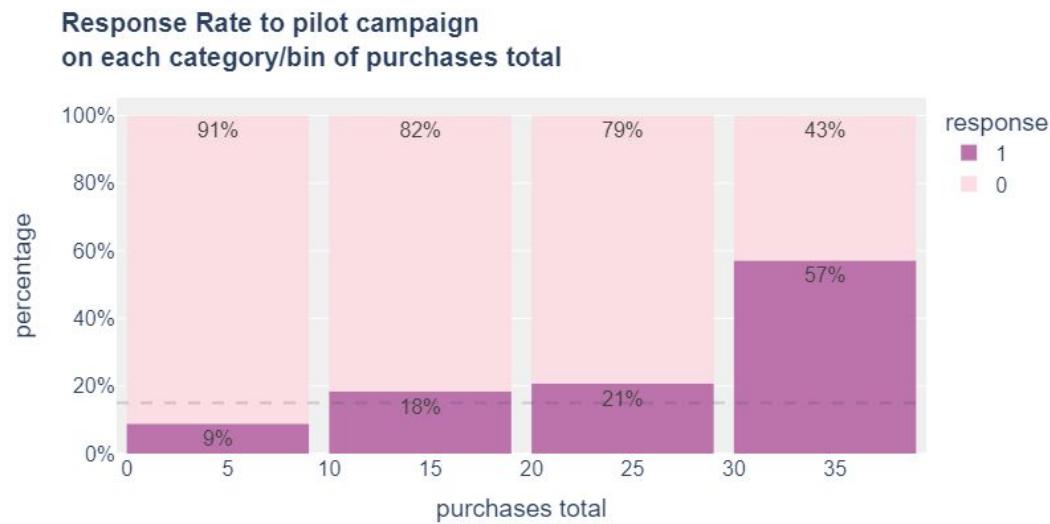
Sim! para grupos com percentual da renda mensal gasta conosco abaixo de 2%, a taxa de sucesso da campanha foi de apenas 12%, a menor entre todos os grupos e também menor do que a taxa geral de sucesso.





Já quando deixamos de falar de valor gasto e sim quantidade de compras feitas, algo muda?

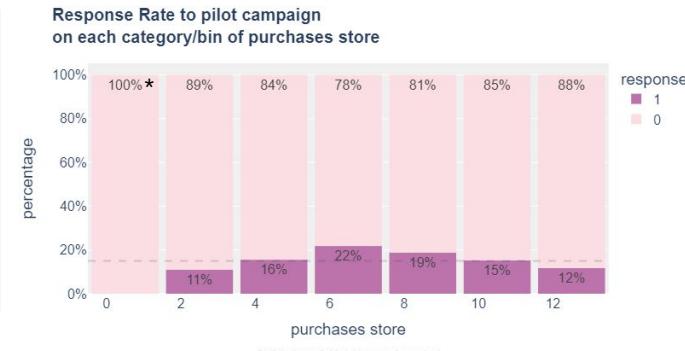
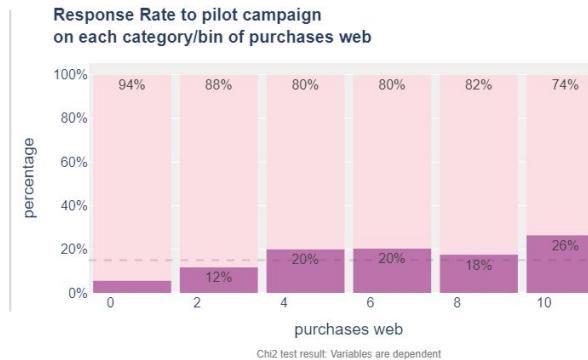
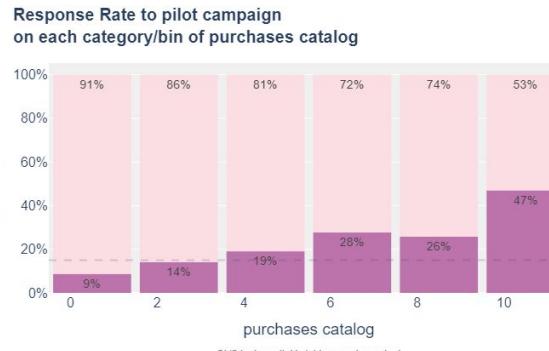
Assim como o esperado, o comportamento se mantém semelhante, assim como grupos de clientes que gastaram mais, tiveram maior taxa de aceitação da campanha, o mesmo se repetiu para grupos de clientes que compraram mais vezes, chegando a uma taxa de aceitação da campanha piloto de 57% para o grupo de clientes que comprou de 30 a 39 vezes.





E quando olhamos para os canais onde essas compras foram feitas, qual o impacto?

O impacto para compras feitas no catálogo e na web é parecido com o comportamento geral, onde grupos de pessoas que mais compraram, tiveram uma maior taxa de aceite da campanha, enquanto que o comportamento de quem comprou diretamente nas lojas varia um pouco de acordo com a quantidade de compras feitas, onde o pico da taxa de aceitação não é encontrado no grupo que mais comprou e sim em um grupo intermediários, mas de qualquer forma, para os 3 canais de compras analisados, todos tiveram grupos que se mostraram impactantes na aceitação da oferta da campanha piloto.



*apenas 3 pessoas



Para os clientes que compraram esse produtos em oferta, existe algum impacto nesse tipo de compra?

Ao comparar o comportamento de compra de grupos de clientes que compraram em ofertas ou não, encontramos que para ambos os grupos, a quantidade de compras feitas em cada tipo de compra se mostrou relevante para a taxa de aceitação da campanha piloto, porém o grupo de clientes que compraram mais fora de oferta, se mostrou mais propensos a aceitar a oferta da campanha piloto.

Response Rate to pilot campaign
on each category/bin of purchases deals



*apenas 1 cliente

Response Rate to pilot campaign
on each category/bin of purchases no deals

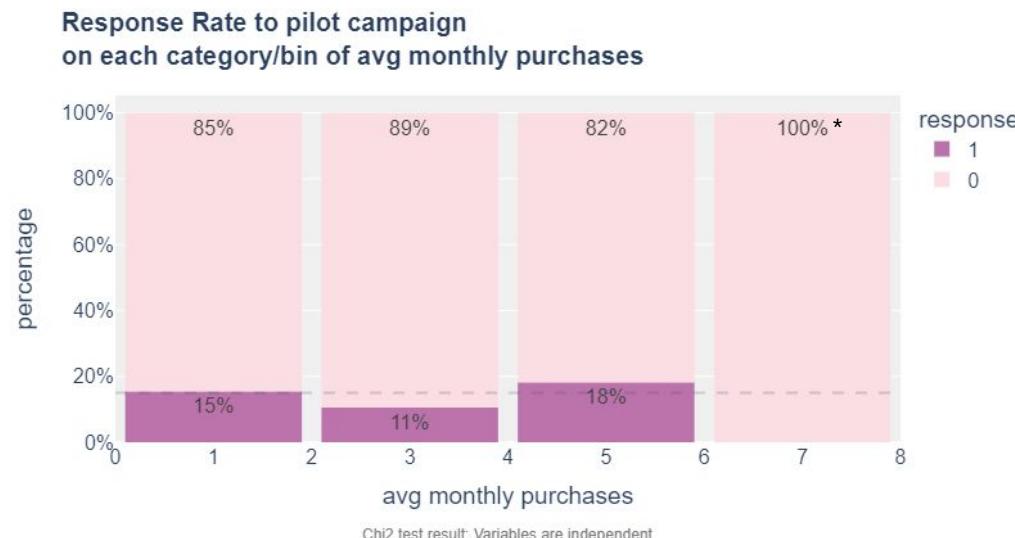


Chi2 test result: Variables are dependent



Quantidade média de compras feitas no mês

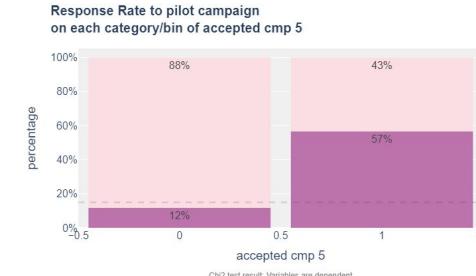
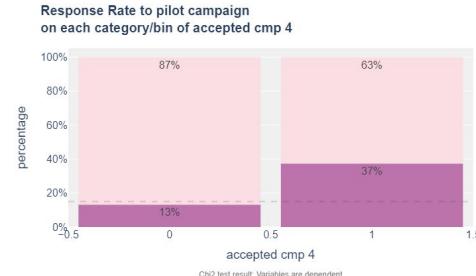
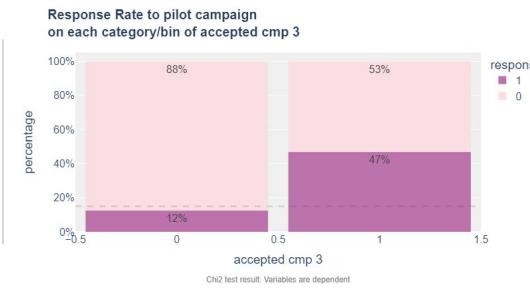
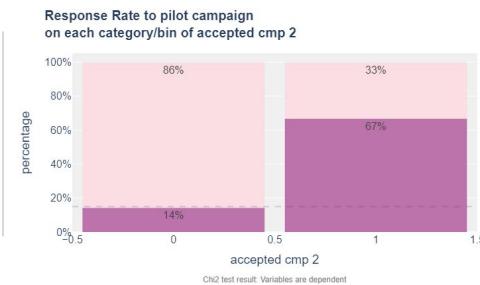
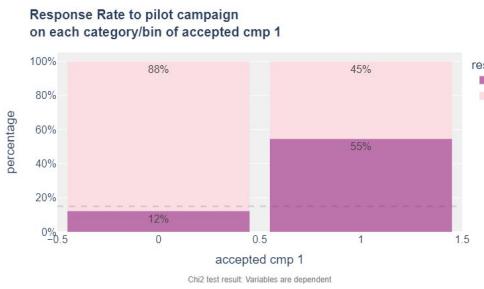
Já ao compararmos diferentes grupos pela quantidade de compras feitas no mês, a variável se mostrou independente com a taxa de sucesso da campanha.





Ter aceitado uma campanha anterior, aumenta a propensão a aceitar a campanha atual?

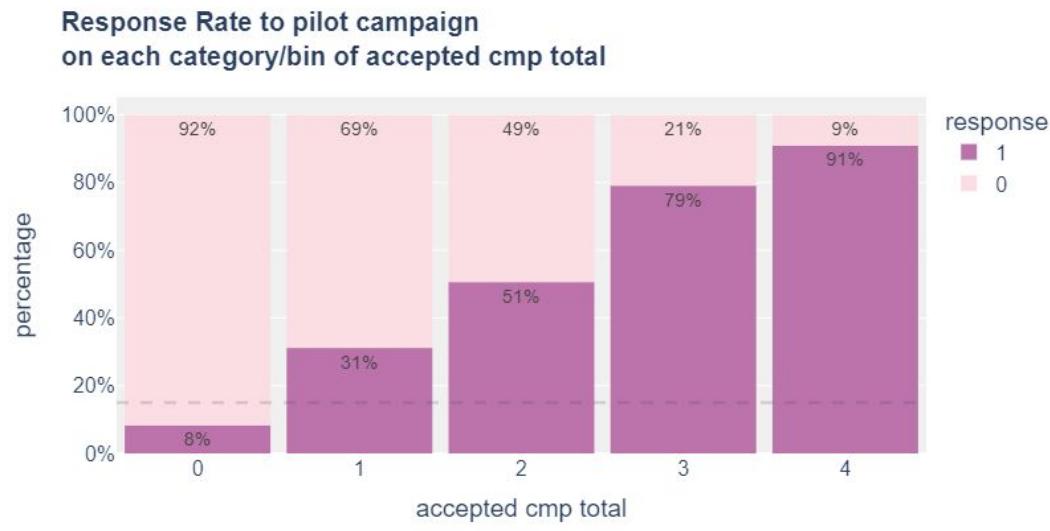
Clientes que aceitaram campanhas anteriores se mostraram mais propensos a aceitarem a campanha piloto, com destaque para os clientes que aceitaram a campanha 1, 2 e 5, onde esses clientes tiveram taxas de aceitação da campanha piloto acima de 55%.





E ter aceitado mais campanhas passadas, tem algum impacto para aceitar a campanha piloto?

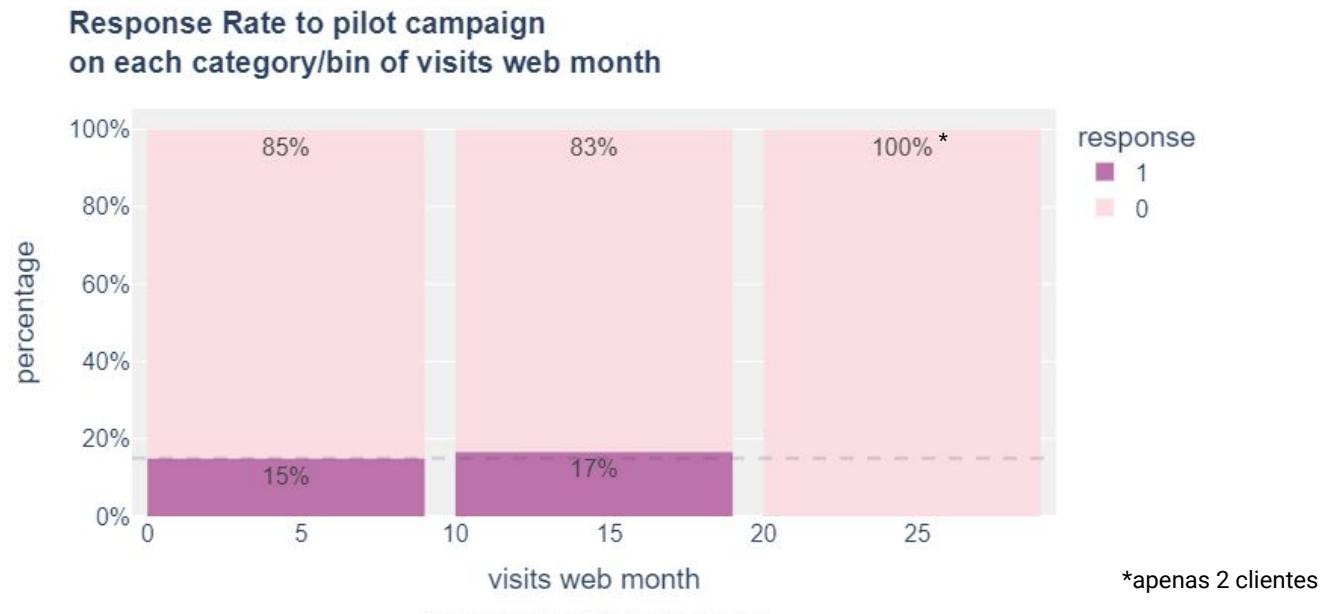
A resposta é sim! como podemos ver no gráfico abaixo, quanto mais campanhas passadas o cliente aceitou, maior a taxa de aceitação da campanha piloto, partindo de 8% para clientes que não aceitaram nenhuma campanha anterior, até chegar em incríveis 91% para clientes que aceitaram 4 campanhas!





A quantidade de visitas ao site da empresa impacta de alguma forma?

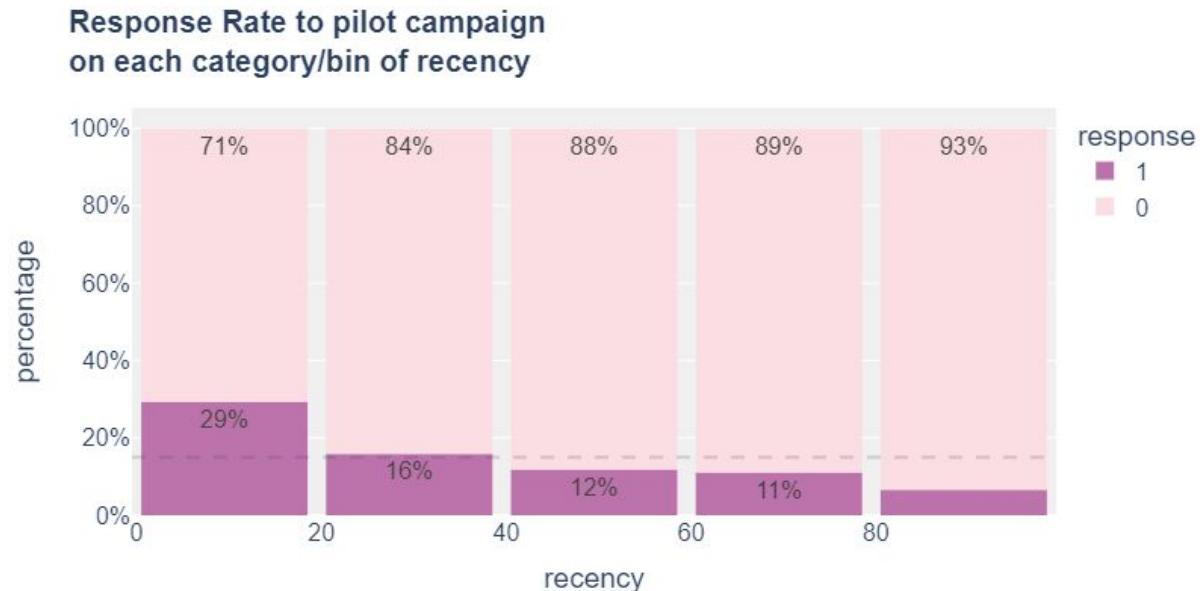
A resposta é não! O número de vezes que o cliente acessou a página da empresa pouco interfere na sua propensão a aceitar a campanha piloto, as variáveis são independentes entre si.





Qual o impacto da quantidade de dias desde que o cliente realizou sua última compra conosco (recência)?

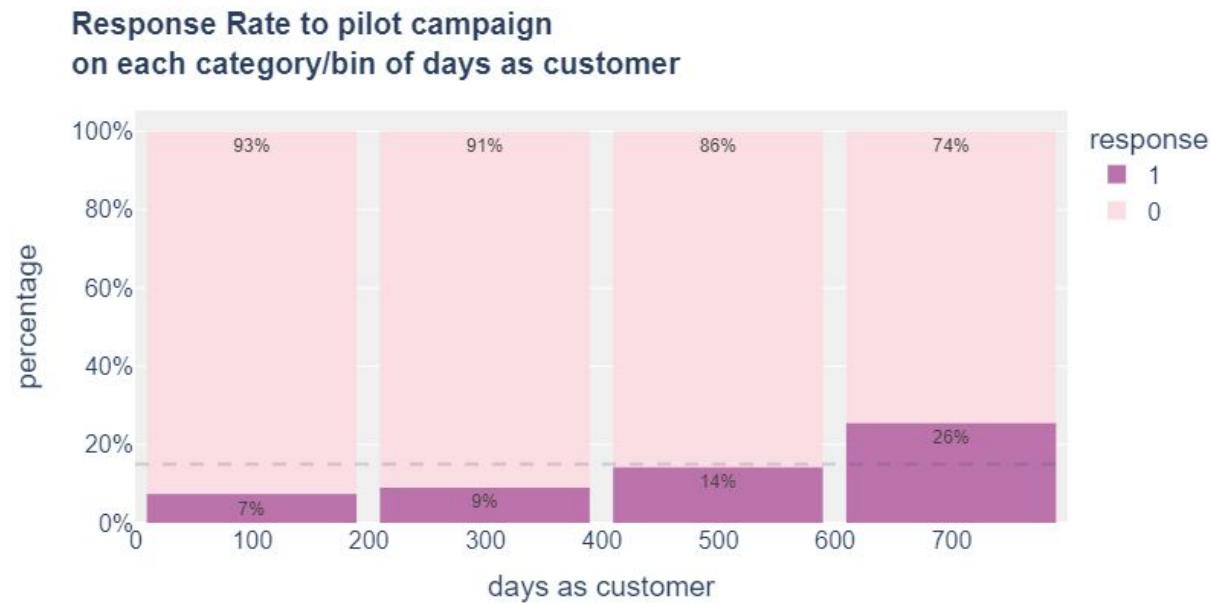
Como o esperado, clientes com valores de recência mais baixos, possuem maior taxa de aceitação da campanha piloto, onde que para clientes que fizeram sua última compra em até 20 dias, essa taxa chega a 29%.





Qual foi o comportamento dos clientes mais antigos?

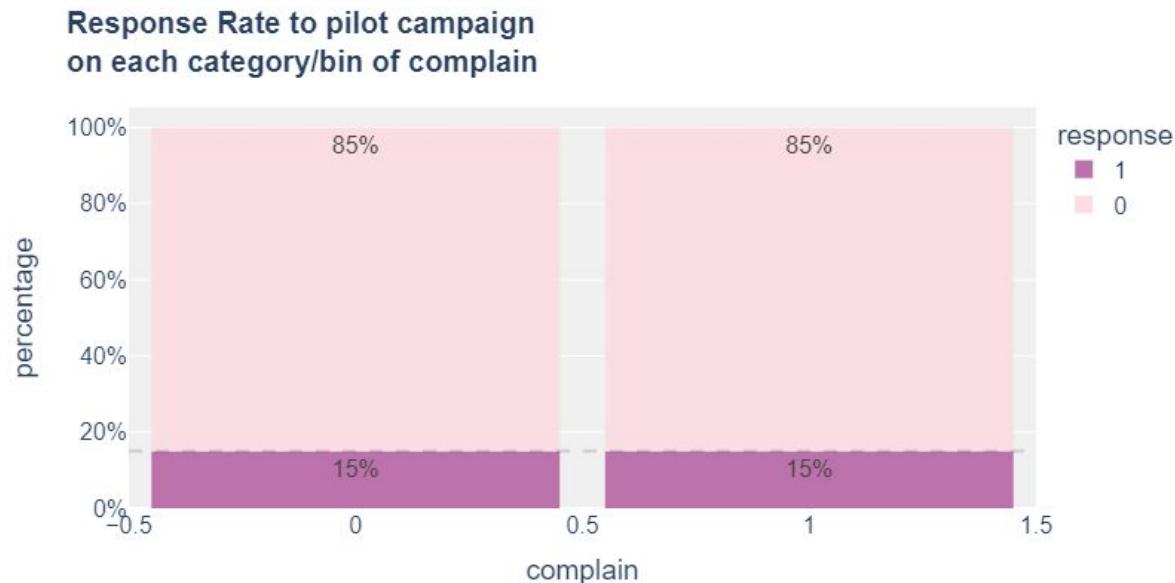
Clientes mais antigos também se mostraram mais propensos a aceitarem a oferta da campanha piloto do que clientes mais novos, foi possível observar que conforme mais tempo como cliente, maior sua propensão a aceitar a campanha, onde partimos de uma taxa de sucesso de 7% para clientes com até 200 dias, indo até 26% para clientes com mais de 600 dias.





E os clientes que tiveram alguma reclamação?

Para os clientes que tiveram alguma reclamação, a taxa de sucesso da campanha se mostrou a mesma se comparada ao grupo de clientes que não tiveram reclamações, o que indica que o fato do cliente ter reclamado não impacta a taxa de sucesso da campanha piloto.





Definição dos públicos que temos em nossa base

Com as análises anteriores conseguimos entender quais variáveis separadamente, foram mais impactantes na decisão do cliente em aceitar a oferta da campanha piloto e tornou-se possível montar uma estratégia de campanha focada em clientes que possuem maior renda, não tem filhos e aceitaram campanhas passadas por exemplo.

Porém, quando olhamos para o nosso público por completo, considerando todas as suas características, não conseguimos ter clareza se existem grupos de clientes com características em comum e quais seriam essas características.

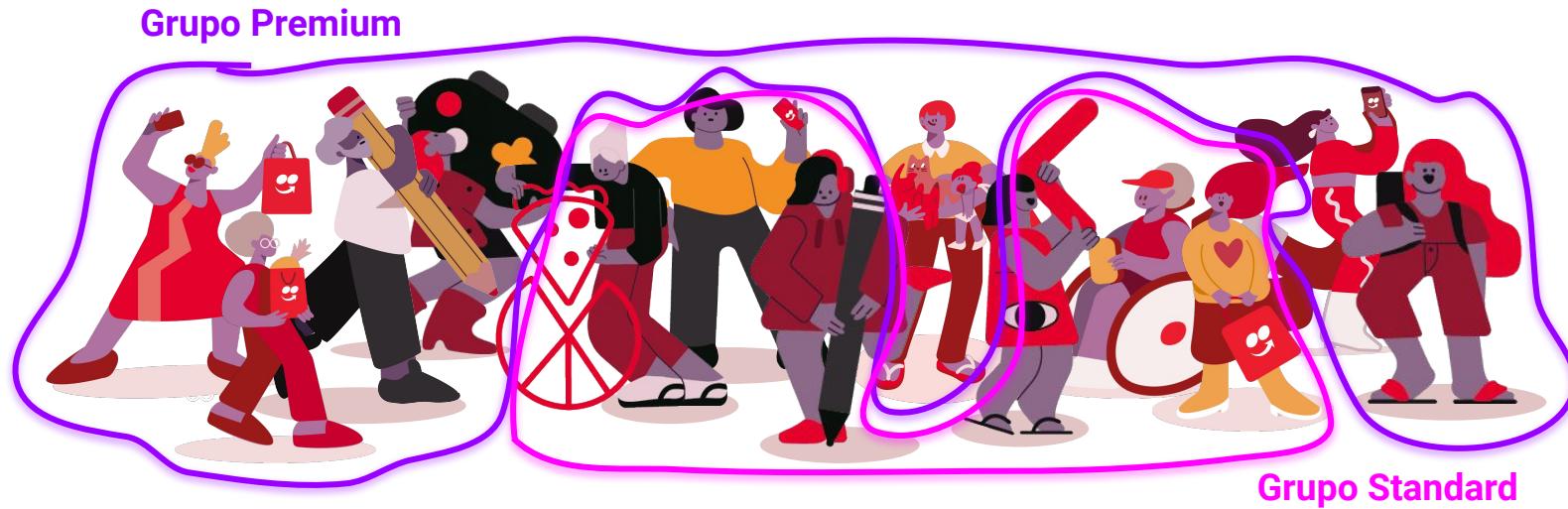




Definição dos públicos que temos em nossa base

Para resolver esse problema, utilizamos uma técnica chamada Kmeans, onde por meio da semelhança entre as características de cada indivíduo, eles são agrupados para formar grupos mais homogêneos possíveis.

Para o nosso caso, a base de clientes foi dividida em 2 grupos, e iremos chamá-los de grupos Premium e Standard, onde o grupo Premium é o grupo com maior gasto médio em nossos produtos.





Definição dos públicos que temos em nossa base

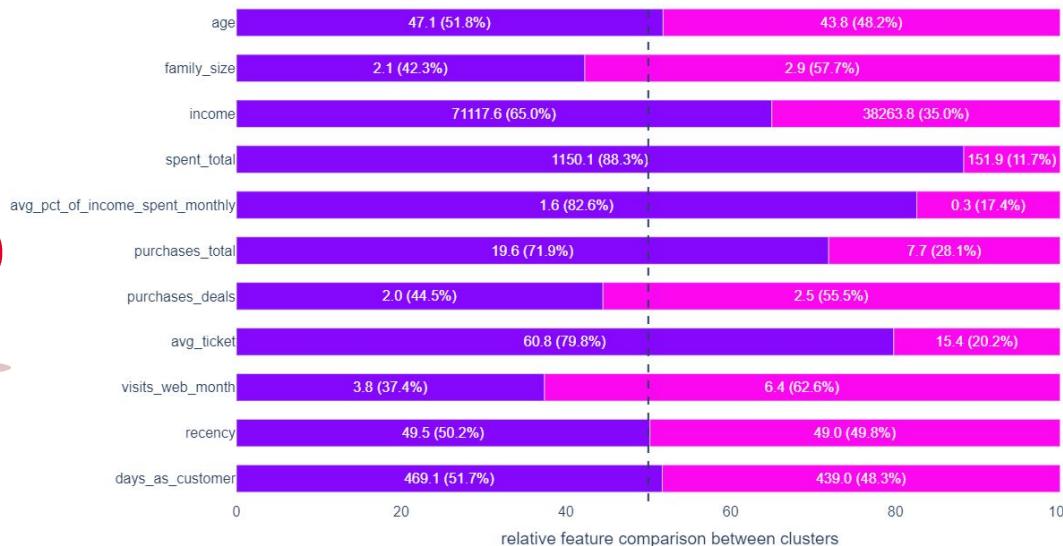
As características médias para cada grupo foram comparadas e expostas no gráfico abaixo.

No gráfico abaixo vemos a média de algumas variáveis para cada grupo e a comparação relativa entre elas, como podemos analisar, algumas características como idade, recência e dias como cliente pouco influenciaram na definição dos grupos, onde os dois grupos possuem médias muito próximas. Enquanto outras características como tamanho da família, renda, gasto total, percentual da renda gasta conosco, compras totais, compras em promoção, ticket médio e visitas web, aparecem como maiores diferenciadores entre os dois grupos.

Premium



Standard





Definição dos públicos que temos em nossa base

Dessa forma podemos definir algumas características chaves que diferenciam os dois grupos entre si.

Por termos somente dois grupos, para as características que se mostraram relevantes na divisão dos grupos, elas tenderão a ser opostas, sobre o tamanho dos grupos, o grupo Premium possui 905 pessoas, o que representa 41% da base, enquanto o grupo Standard possui 1309 pessoas, equivalente a 59% da base de clientes analisados.

Premium

41% dos clientes

- . Famílias menores (~ 2 pessoas);
- . Renda elevada (~ \$71k);
- . Gastos totais elevados (~ \$1150);
- . Percentual da renda gasta conosco elevada (~1.6%);
- . Muitas compras conosco (~ 20 produtos);
- . Menos compras em promoções (~ 2 produtos);
- . Ticket médio elevado (~ \$61);
- . Poucas visitas web (~4 visitas /mês)



Standard

59% dos clientes

- . Famílias maiores (~ 3 pessoas);
- . Renda menor (~ \$38k);
- . Gastos totais baixos (~ \$152);
- . Percentual da renda gasta conosco baixos (~0.3%);
- . Poucas compras conosco (~ 8 produtos);
- . Mais compras em promoções (~ 2.5 produtos);
- . Ticket médio baixo (~ \$15);
- . Mais visitas web (~6 visitas /mês)





Criação do modelo de propensão a aceite da oferta da Campanha 6

Agora que já entendemos o comportamento dos nossos clientes em relação a aceitação da campanha piloto e entendemos os diferentes grupos de clientes que temos em nossa base, é hora de criar um modelo que preveja qual a propensão a compra de cada um de nosso clientes baseado em suas características.

Por se tratar do primeiro uso de aprendizado de máquina da área, optamos em priorizar a explicabilidade, nesse sentido foi escolhido o modelo linear chamado **regressão logística**, onde dada as características de um indivíduo, é previsto se um evento acontecerá ou não.



Maria:

- . 28 anos;
- . Casada;
- . Não tem filhos;
- . Cliente há 180 dias;
- . Cluster Premium;
- ...



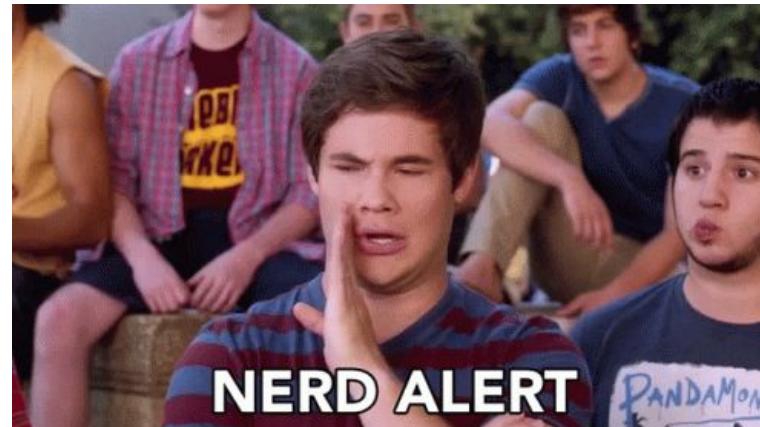
Maria vai aceitar a oferta da campanha 6?



Criação do modelo de propensão a aceite da oferta da Campanha 6

Nerd Alert! Para a criação do modelo, alguns passos precisaram ser seguidos, tais processos foram o tratamento de valores nulos, tratamento de outliers, escolha de variáveis, balanceamento da base, divisão da base, modelagem e validação cruzada.

No próximo slide iremos trazer de forma resumida as principais decisões tomadas e técnicas utilizadas para a criação do modelo, mas todos os passos da modelagem, assim como as análises anteriores estão explícitos no notebook do modelo e podem ser acessados por [aqui](#).





Criação do modelo de propensão a aceite da oferta da Campanha 6

Para os processos de **preenchimento de nulos** utilizamos o preenchimento baseado nas características de outra coluna na qual a coluna com dados faltantes tenha alta correlação. No nosso caso a coluna com dados faltantes era “Renda”, e utilizamos a coluna “Gasto total” para auxiliar no seu preenchimento.

Para o **tratamento de outliers** utilizamos uma técnica chamada IQR (Amplitude interquartil) para descartar ou valores considerados atípicos, foram encontrados valores atípicos nas colunas “Renda”, “Idade”, “Gasto total”, “Gasto com carne”, “Gasto com doces”.

Para a **seleção de variáveis**, utilizamos o cálculo de correlação entre variáveis, onde excluímos do modelo variáveis altamente (> 0.5) correlacionadas entre si, mantendo a variável mais correlacionada com a variável resposta e que faça sentido pelo contexto do problema, as variáveis selecionadas foram “Anos de estudo”, “Idade”, “Visitas web”, “Tamanho da família”, “Recência”, “Dias como cliente”, “Total de campanhas aceitas”, “Reclamação feita” e a variável resposta.

Para o **balanceamento de bases**, por temos uma das classes da variável resposta com muito menos registros do que a outra (15:85), utilizamos a técnica de superpopulação randômica, onde através aleatoriamente através dos registros da classe minoritária, criamos registros artificiais para balancear a base.

Para a **divisão da base**, utilizamos 80% como base de treino e 20% como base de teste, essa divisão permite estimar o quanto genérica a performance do modelo é.

Para a **modelagem** utilizamos regressão logística, que é um modelo estatístico utilizado para determinar a probabilidade, no nosso caso, de um cliente aceitar ou não a oferta da campanha, essa técnica se destaca sobretudo pela maior facilidade de interpretar os resultados obtidos, para o nosso caso, fizemos a modelagem para os dois clusters separadamente.

A **validação cruzada**(Kfold) é uma técnica que permite avaliar o resultado do nosso modelo por meio da quebra de diferentes subconjuntos de testes na nossa base de dados, de forma a evitar que ele seja treinado somente na mesma base e gere resultados pouco generalizados



Entendendo o Resultados do modelo



Para ilustrar o resultado do modelo, trouxemos a visualização chamada “Matriz Confusão”, ela exibe a distribuição em números absolutos dos clientes para os resultados previstos e reais de aceite de campanha, esses números absolutos fazem referência a base de teste, ou seja, 20% da base total analisada. Além disso, também são trazidas as métricas de sucesso do modelo, que são acurácia, precisão e recall.

Matriz Confusão com os dados de teste

		Previsão	Previsão
		0	1
Realidade	0	304 Verdadeiro Negativo	72 Falso Positivo
	1	16 Falso Negativo	51 Verdadeiro Positivo

Acurácia: 80%, Precisão: 81%, Recall: 76%

. **Verdadeiro Negativo (304 clientes)**: Prevemos que o cliente não iria aceitar a campanha e ele realmente não aceitou), ou seja economizamos dinheiro da campanha ao não veicular a campanha para ele.

. **Falso Positivo (72 clientes)**: Prevemos que o cliente aceitaria a campanha, mas ele não aceitou, ou seja, perdemos o dinheiro do custo de veiculação da campanha para esse cliente.

. **Falso Negativo (16 clientes)**: Prevemos que o cliente não iria aceitar a campanha, mas ele aceitou, ou seja perdemos a oportunidade de captar esse cliente.

. **Verdadeiro Positivo (51 clientes)**: Prevemos que o cliente aceitaria a campanha e ele realmente aceitou, ou seja, fomos assertivos na sua segmentação.

. **Acurácia**: Indica a performance geral do modelo;

. **Precisão**: Dentre todas as classificações da classe Positivo que o modelo fez, quantas estão corretas;

. **Recall**: dentre todas as situações de classe Positivo como valor esperado, quantas estão corretas;



Resultados do modelo

A modelagem foi feita de forma separada para cada um dos clusters, Premium e Standard, para o cluster premium, tivemos melhores resultados, chegando a uma acurácia de 81% e uma menor taxa de falsos positivos. Dessa forma, a fim de otimizar os resultados da campanha, decidimos seguir a análise somente para o cluster Premium.

Matriz Confusão (dados de teste)

		Previsão	Previsão
		0	1
Realidade	0	110 Verdadeiro Negativo	27 Falso Positivo
	1	7 Falso Negativo	37 Verdadeiro Positivo

Acurácia: 81%, Precisão: 80%, Recall: 84%

Matriz Confusão (dados de teste)

		Previsão	Previsão
		0	1
Realidade	0	182 Verdadeiro Negativo	56 Falso Positivo
	1	4 Falso Negativo	20 Verdadeiro Positivo

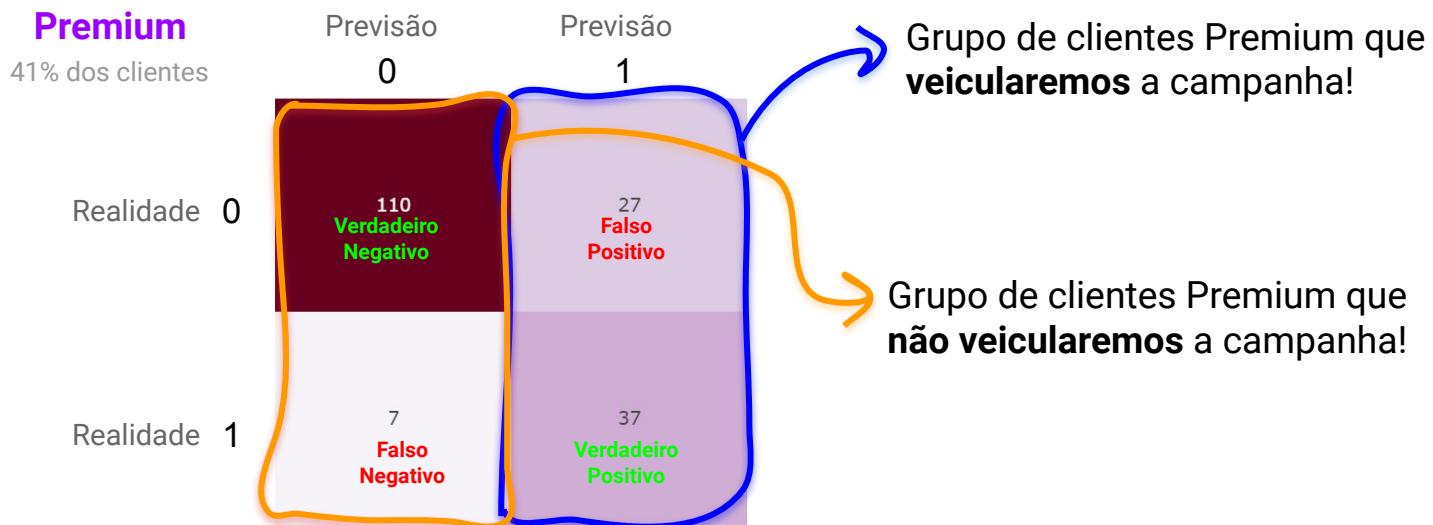
Acurácia: 77%, Precisão: 76%, Recall: 83%



Resultados do modelo

Com isso, determinamos os grupos que devem e não devem receber a campanha dentro do cluster Premium (não veicularmos para ninguém do Cluster Standard), devemos veicular a campanha para todas as pessoas que o modelo previu que aceitariam a campanha (Previsão = 1), mesmo que na realidade algumas delas não tenham aceitado, como é o caso dos falsos positivos, enquanto nós não iremos veicular a campanha para o outro grupo (Previsão = 0), mesmo que tenham pessoas que eventualmente aceitariam a campanha, os falsos negativos.

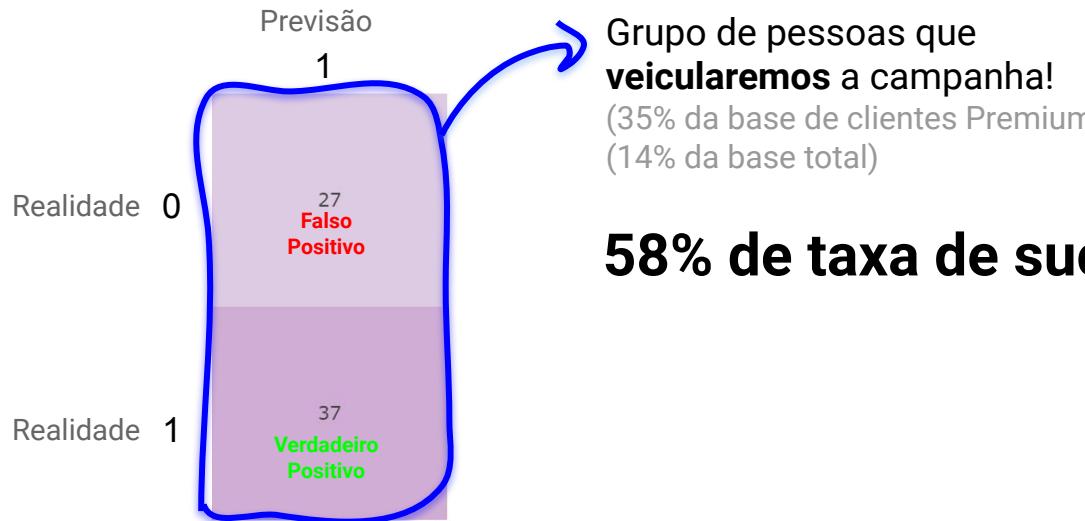
Matriz Confusão (dados de teste)





Resultados do modelo

Assim, iremos veicular a campanha somente para 35% da base de clientes Premium, o que representa 14% da base total de clientes, e olhando somente para esse grupo que receberá a campanha, temos uma taxa de sucesso da campanha estimada em 58%, valor muito acima dos 15% que tínhamos anteriormente ao não segmentar os envios.





Medindo o impacto das variáveis para o resultado

Também se fez importante entender quais variáveis foram mais impactantes para o resultado do modelo, o gráfico abaixo está ordenado da variável mais para a menos impactante em termos absolutos. Como podemos notar a variáveis como tamanho da família e recência se mostraram muito impactantes negativamente, ou seja, quanto menor os seus valores, maior o impacto para a previsão de um cliente aceitar a oferta da campanha, enquanto variáveis como campanhas aceitas e tempo como cliente, se mostraram importantes positivamente. Enquanto variáveis mais à esquerda, como valor gasto em frutas e reclamação, tiveram pouco impacto. Dessa forma, analisamos que esses valores estão em linha com o que foi visto anteriormente na análise descritiva dos clientes.





Show me the money!!!

Agora que conhecemos a taxa de sucesso estimada da campanha, o tamanho da base de clientes que a receberá e suas características, quanto de dinheiro essa campanha deve gerar? Será suficiente para termos lucro ou mesmo depois da segmentação a campanha continuará dando prejuízo?



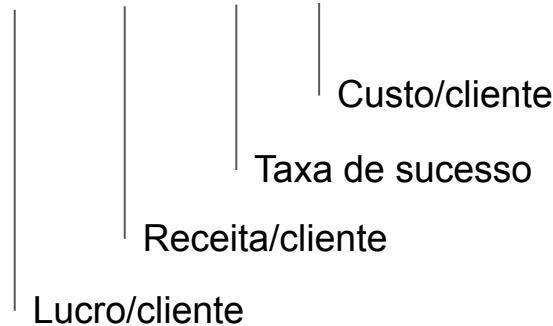


Show me the money!!!

Definimos a fórmula de lucro por cliente considerando a receita gerada caso a pessoa aceite a oferta da campanha, que é de \$11, o custo para veiculação dessa campanha por pessoa, que é de \$3 e também a taxa de sucesso da campanha, ou seja, qual o percentual das pessoas que impactamos com a campanha que realmente aceitaram a nossa oferta.

Fórmula do lucro por cliente

$$y = 11 * x - 3$$





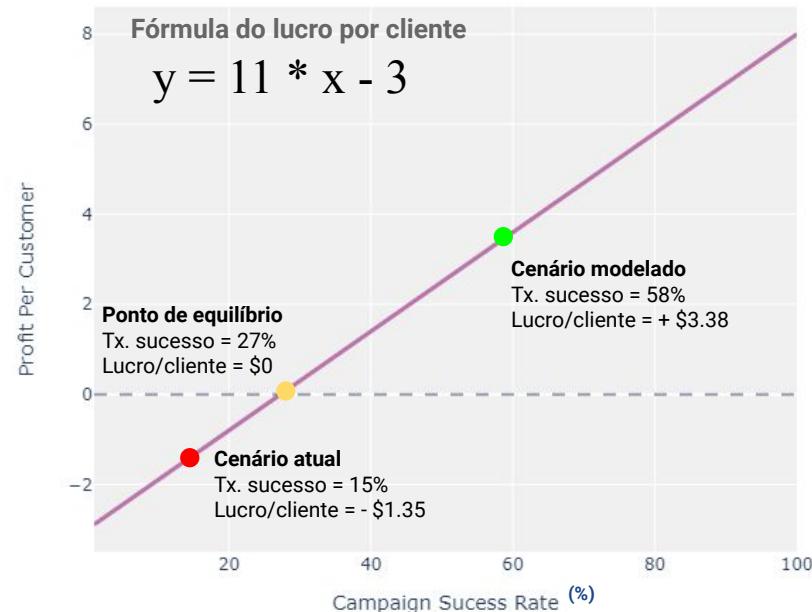
Show me the money!!!

Essa fórmula dá origem ao gráfico ao lado, que explicita diversos possíveis cenários de lucro por cliente de acordo com a taxa de sucesso da campanha.

Para o nosso **cenários atual**, onde a campanha piloto foi veiculada sem nenhum tipo de segmentação de clientes, nossa taxa de sucesso é de 15%, resultado em um prejuízo de \$1.35 por cliente.

Para que a campanha comece a dar algum lucro, a taxa de sucesso precisa ser maior que 27%.

Enquanto para o nosso **cenário modelado**, onde temos uma taxa de sucesso de 58%, conseguimos obter um lucro de \$3.38 por cliente impactado!





Show me the money!!!

Considerando uma base de possíveis clientes a serem impactados de 1 milhão de pessoas, teríamos o seguinte cenário se não realizássemos a segmentação de clientes:



\$ -1.35M 

Prejuízo total da campanha



Show me the money!!!

Já para o cenário modelado, também considerando a base de possíveis clientes a serem impactados de 1 milhão de pessoas, temos o seguinte cenário:



\$473k Lucro total da campanha



Conclusão

Após a finalização dos estudos, podemos concluir utilizando técnicas de ciência de dados é possível otimizar uma campanha através da segmentação de seus clientes, uma campanha que caso fosse veiculada sem segmentação, seria veiculada para uma base de 1 milhão de clientes, trazendo um prejuízo financeiro de \$1.35M, enquanto ao utilizarmos técnicas para segmentação, veiculamos a mesma campanha, somente para 140 mil cliente, obtendo um lucro estimado de \$473k.

Dessa forma, evitamos que clientes sejam impactados por campanhas indesejadas, concentrarmos nossos esforços em um número menor de clientes, além de termos uma campanha de custo mais baixo e o mais importante, que gera lucro para a empresa.





Próximos passos

Como próximos passos, algumas outras melhorias e testes podem ser aplicados, tais como:

- Rodar outra campanha piloto utilizando as regras de segmentação propostas nesse estudo e validar a estimativa de taxa de sucesso da campanha;
- Utilizar outros algoritmos para clusterização e segmentação dos clientes mesmo que perca em explicabilidade;
- Definir um ponto ótimo de lucro considerando a taxa de sucesso da campanha e o tamanho do público a ser impactado;
- Criação de mais variáveis utilizando a nossa base original ou unindo outras fontes de dados;



Obrigado





Apêndice

Como já citado anteriormente, o código deste projeto está disponível através deste [link](#).

E além disso, também compartilho um [documento](#) com as principais dúvidas conceituais que tive ao executar esse projeto, no qual foi muito importante para que eu revisitasse alguns conceitos e aprendesse outros novos.

Dúvidas e sugestões podem ser enviadas via slack para @lucas.fraga ou email lucas.fraga@gmail.com

Valeu! (:

