Desafio Q-Bem

Ciência de Dados (mar/22)

Olá!

Me chamo Euclides Formica

Você me encontra em

(11) 97444-2729 | eformica@gmail.com
www.linkedin.com/in/eucformica

Roadmap

1. Entendendo o Desafio

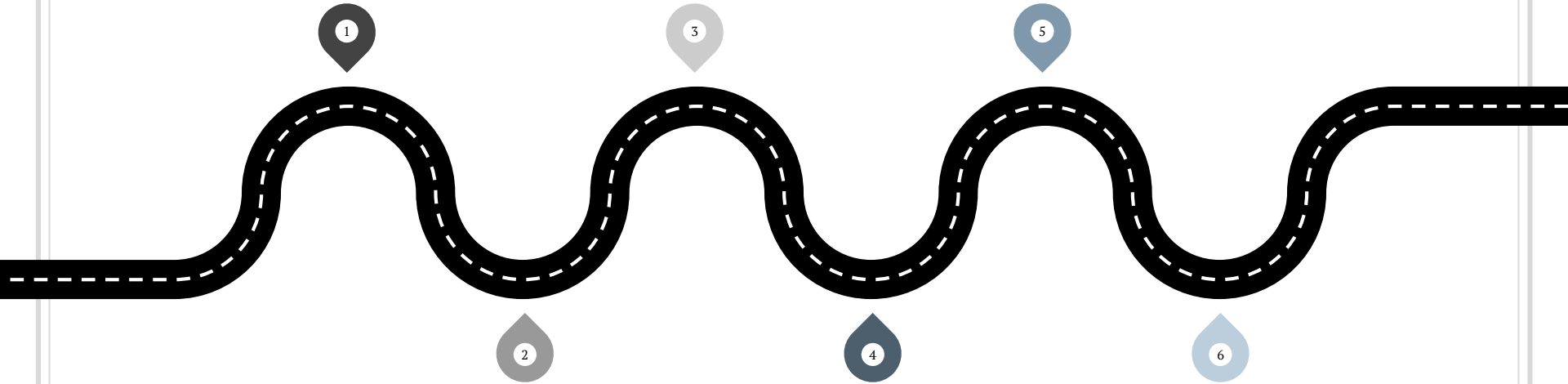
2. Quem são nossos clientes?

3. Expandindo Horizontes

4. A Campanha de Marketiing

5. Próximos Passos

6. Perguntas



1.
Entendendo o Desafio

Empresa

- Considere uma empresa bem estabelecida que opera no setor de varejo de alimentos.
- Têm cerca de centenas de milhares de clientes cadastrados e atendem a quase um milhão de consumidores por ano.
- Teve receitas sólidas nos últimos 3 anos, mas as perspectivas de crescimento do lucro não são promissoras, por isso, várias iniciativas estratégicas estão sendo consideradas para reverter a situação.

Departamento de Marketing

- Pressionado a gastar seu orçamento anual com mais sabedoria, o CMO percebe a importância de ter uma abordagem mais quantitativa na tomada de decisões, razão pela qual uma pequena equipe de cientistas de dados foi contratada para construir um modelo preditivo que apoiará as iniciativas de marketing direto.

Campanha Piloto

- O objetivo da equipe é construir um modelo preditivo que produza o maior lucro para o próxima campanha de marketing direto, prevista para o próximo mês.
- A nova campanha visa vender um novo gadget para o banco de dados de clientes.
- Para construir o modelo, foi realizada uma campanha piloto envolvendo 2.240 clientes.
- Os clientes foram selecionados aleatoriamente e contatados por telefone sobre a aquisição do gadget.
- Durante os meses seguintes, os clientes que compraram o oferta foram devidamente rotulados.
- O custo total da campanha de amostra foi de 6.720MU e a receita gerado pelos clientes que aceitaram a oferta foi de 3.674MU.
- A campanha teve um lucro de -3.046MU. A taxa de sucesso da campanha foi de 15%.

Produtos

Carnes

Frutas Exóticas

Vinhos

Preparados de Peixe

Doces

Linhas

Gold / Comum

Canais de Venda

Catálogo / Site / Loja Física

Objetivos

1. Explorar os dados

Você precisa fornecer à equipe de marketing uma melhor compreensão das características dos entrevistados; Como as variáveis se conectam com as taxas de resposta? Que outras relações entre variáveis são interessantes para o negócio? Quais ações podemos tirar da EDA?

2. Propor e descrever uma segmentação com base nos comportamentos dos clientes

Quantos e quais perfis existem no banco de dados? Como a segmentação se relaciona com o retorno financeiro da campanha?

3. Criar um modelo preditivo que permita maximizar o lucro da próxima campanha de marketing

Qual é a melhor métrica que se correlaciona com a lucratividade da campanha? Simplicidade e consciência do que está acontecendo são preferíveis a implementações de algoritmos complexos.

4. Criar uma apresentação de negócios altamente eficaz

O caso deve conter uma apresentação que, ao mesmo tempo, traga força técnica, insights e ações, mas se comunique com um público não técnico, como um CMO. Leve o público em uma viagem. Ajude-os a ver a história de sucesso e o que ela trará.

2.
Quem são nossos clientes?

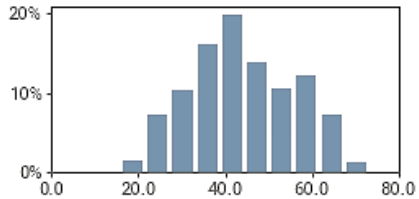
Perfil dos Entrevistados e Correlações Interessantes

Segmentação por Nível de Consumo

Segmentação por Grupos de Interesses

Perfil dos Entrevistados

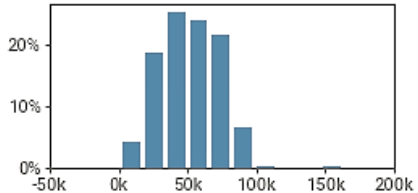
Idade



44 anos
é a idade média

60,3%
possui idade entre
34 e 56 anos

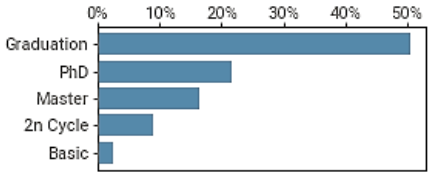
Renda



$ 51.958,81
é a renda média

73,5%
possui renda entre
$30.000 e $80.000

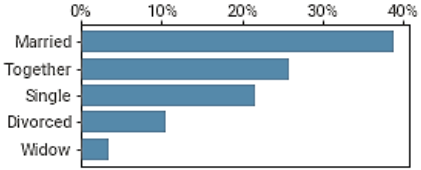
Escolaridade



50,4%
possuem Graduação

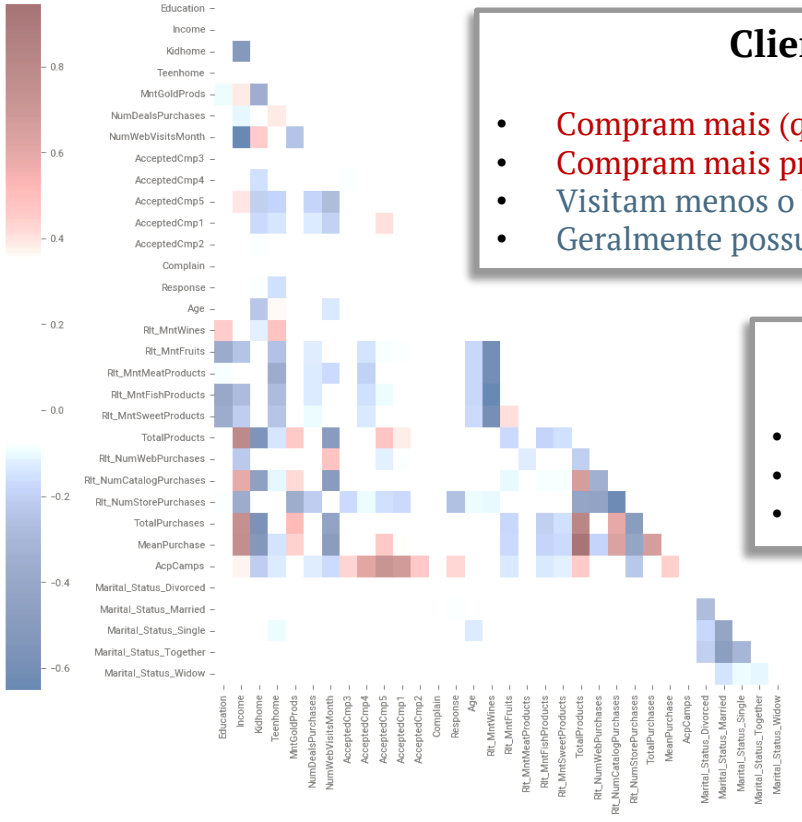
33,2%
possuem PhD ou Master

Estado Civil



64,5%
São casados ou vivem
em união estável

Correlações Interessantes



Clientes com Maior Renda

- Compram mais (quantidade, volume e produtos por compra);
- Compram mais produtos Gold e por catálogo;
- Visitam menos o Website da empresa;
- Geralmente possuem poucas crianças em casa;

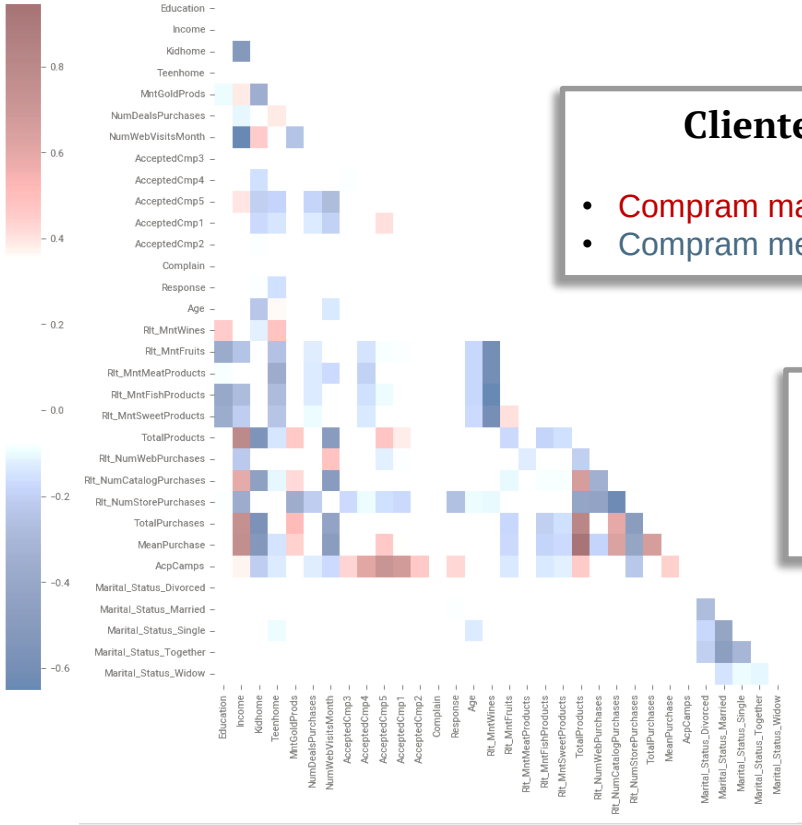
Clientes com Crianças

- Acessam mais o website;
- Geram menos vendas;
- Compram menos produtos Gold e por Catálogo;

Clientes que Compram Mais Vinhos

- Compram menos produtos de outras categorias

Correlações Interessantes



Clientes que Compram mais por Catálogo

- Compram mais e maior relação valor dos produtos por compra
- Compram menos na loja física

Clientes que Compram Mais na Loja Física

- Compram menos, e com uma relação gasto/compra

Segmentando por Nível de Consumo

Vamos identificar os clientes com maior nível de compra de produtos da Empresa

Segmentando por Nível de Consumo

Estratégia

- Algoritmo k-means;
- O método Silhouette Score foi empregado para definição do número de grupos;
- Os dados foram analisados e caracterizados com os softwares Tableau e Orange Data Mining;

Variáveis de Entrada

- TotalProducts: soma de compras de produtos de todas as categorias;
- TotalPurchases: soma do número de compras de todos os canais de venda;
- MeanPurchase: valor médio por compra;

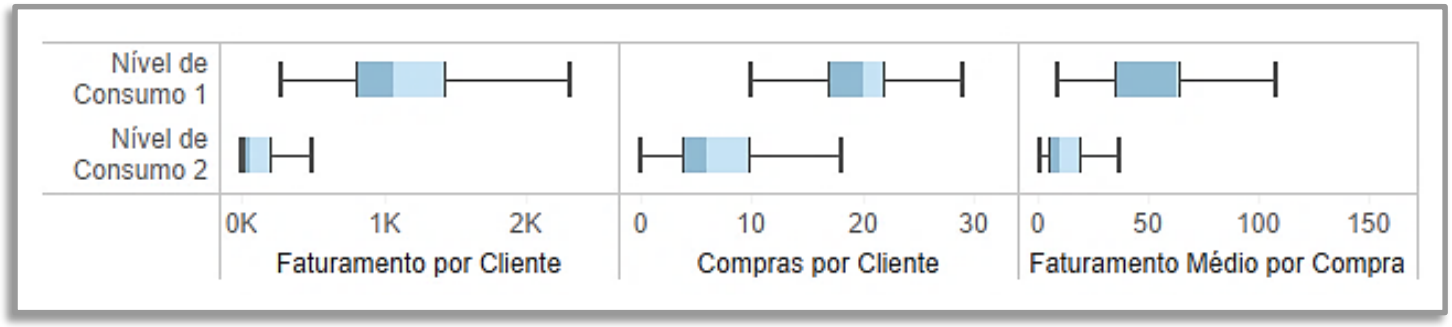
Resultados

- O número de grupos foi definido em 2, que apresentou um Silhouette Score (eficiência da separação) de **62%**;
- A distribuição observada foi:

929 clientes nível 1 (42,2%) para 1.273 clientes nível 2 (57,8%)

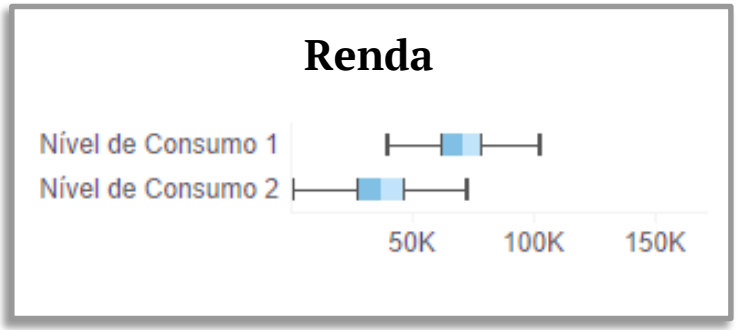
Principais Diferenças entre os Dois Níveis de Consumo Modelados

Os 929 clientes do nível 1 gastam na empresa, em média, 8,2 vezes que os 1.283 do nível 2:

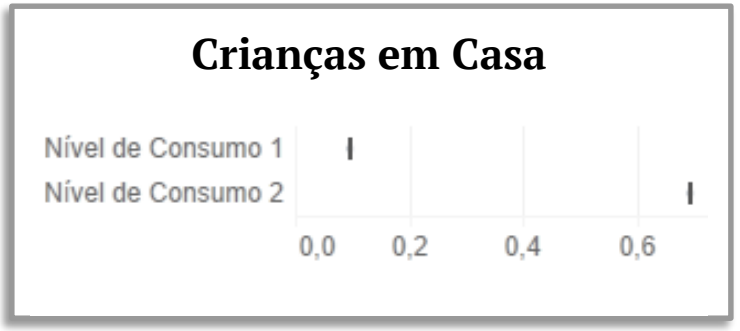


Principais Diferenças entre os Dois Níveis de Consumo Modelados

...tem renda cerca de **1,8 vezes** maior:

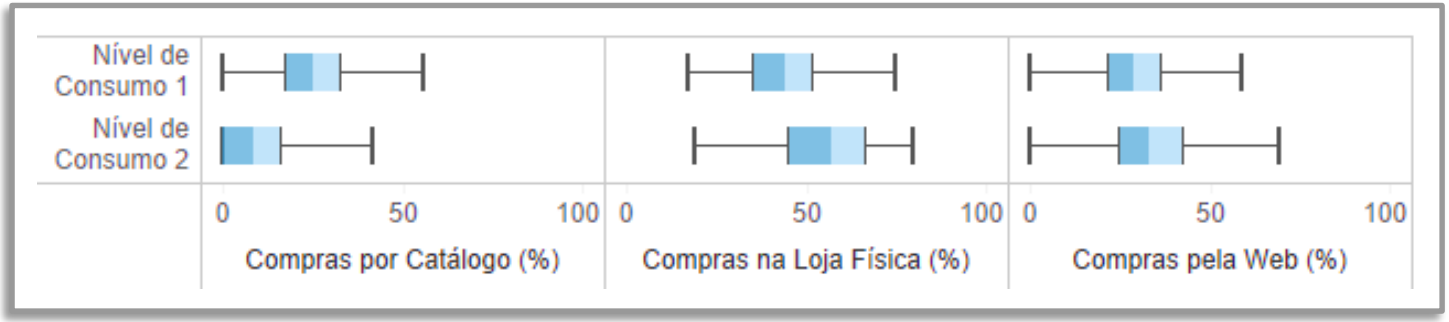


... e possuem bem **menos crianças em casa**:



Principais Diferenças entre os Dois Níveis de Consumo Modelados

Diferenças significativas também são observadas nos canais de compra de preferência catálogo e loja física:



Segmentando por Grupos de Interesses

Vamos identificar o seguimento de produtos preferidos de cada cliente

Segmentando por Grupos de Interesses

Estratégia

- Algoritmo k-means;
- O método Silhouette Score foi empregado para definição do número de grupos;
- Os dados foram analisados e caracterizados com os softwares Tableau e Orange Data Mining;

Variáveis de Entrada

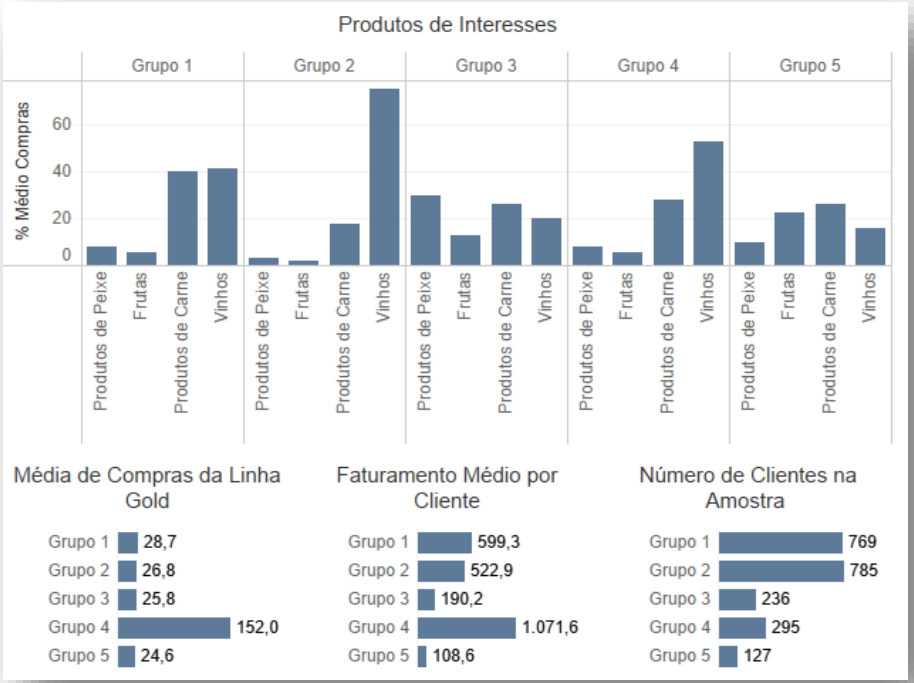
- Rlt_MntWines, Rlt_MntFruits, Rlt_MntMeatProducts, Rlt_MntFishProducts, Rlt_MntSweetProducts, MntGoldProds

Resultados

- O número de grupos foi definido em 5, que apresentou um Silhouette Score (eficiência da separação) de **31%**;

Perfis de Interesses

- **GRUPO 1**: Vinhos e produtos de carne. Pouco interesse pelas demais categorias. Grande número de clientes na amostra; Faturamento alto;
- **GRUPO 2**: Forte interesse por vinhos. Pouco interesse pelas demais categorias. Grande número de clientes na amostra; Faturamento alto;
- **GRUPO 3**: Interesses equilibrados, com destaque para produtos de peixe e carne; Pequeno número de clientes na amostra; Faturamento baixo;
- **GRUPO 4**: Interesse principal por produtos da linha Gold, principalmente por vinhos e produtos de carne; Faturamento bastante alto;
- **GRUPO 5**: Interesses equilibrados, com destaque para produtos de carne e frutas. Pequeno número de clientes na amostra; Faturamento baixo;



Os Grupos 4, 2 e 1 apresentam os maiores faturamentos por cliente!

3.
Expandindo Horizontes

Modelos Preditivos para Prospecção de Novos Clientes

Modelos Preditivos para Prospecção de Novos Clientes

Estratégia

- Vários algoritmos de aprendizado supervisionado foram comparados para prever o Nível de Consumo e Grupo de Interesses a partir de dados que independem do histórico de compras.

Variáveis de Entrada

- Education, Income, Kidhome, Teenhome, Age e Material_Status.

Validação

- Validação cruzada k-fold, com k = 5.

Métricas

- AUC, Acurácia (CA), F1, Precisão e Recall

Metaparâmetros

Modelo	Metaparameters
Gradient Boosting	Method=Gradient Boosting (scikit-learn); Number of trees=100; Learning rate=0.1; Limit depth individual trees=3; Replicable training=True; Do not split substes smaller than=2; Fraction of training=1;
Random Forest	Number of trees=50; Replicable training=True;
Naive Bayes	-
kNN	Number of neighbors=5; Metric=Euclidean; Weight=Uniform;
SVM	SVM Type=SVM; Cost=1;Regression Loss Epsilon=0.1; Kernel=RBF; Numerical tolerance=0.001; Iteration Limit=100
AdaBoost	Number of Estimators=50; Learning Rate=1; Classification algorithm: SAMME.R; Regression loss function=Linear;
Tree	Induce binary tree=True; Min. number of instances in leaves=2; Do not split subsets smaller than=5; Limit the maximal tree depth to=100; Stop when majority reaches[%]=95

Modelos Preditivos para Prospecção de Novos Clientes

Resultados

- Modelos para predição do Grupo de Consumo apresentaram melhor desempenho;
- O Gradiente Boosting apresentou o melhor métrica AUC (área sobre a curva ROC em ambos os alvos;
- As métricas obtidas são suficientes para identificar possíveis clientes em potencial em outras bases de dados;

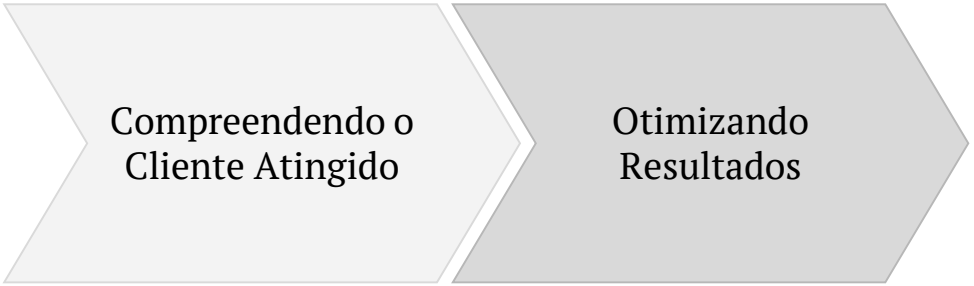
Modelos Avaliados para Predição do Grupo de Consumo

Model	AUC	CA	F1	Precision	Recall
Gradient Boosting	0,967	0,903	0,903	0,904	0,903
Random Forest	0,962	0,906	0,906	0,906	0,906
Naive Bayes	0,954	0,900	0,900	0,902	0,900
kNN	0,937	0,876	0,876	0,876	0,876
SVM	0,934	0,812	0,802	0,841	0,812
AdaBoost	0,872	0,874	0,874	0,874	0,874
Tree	0,848	0,873	0,873	0,873	0,873

Modelos Avaliados para Predição do Grupo de Interesses (Média sobre as Classes)

Model	AUC	CA	F1	Precision	Recall
Gradient Boosting	0,772	0,522	0,506	0,506	0,522
Random Forest	0,762	0,545	0,539	0,538	0,545
Naive Bayes	0,757	0,492	0,481	0,480	0,492
SVM	0,708	0,443	0,422	0,428	0,443
Tree	0,668	0,473	0,470	0,470	0,473
AdaBoost	0,640	0,473	0,477	0,481	0,473
kNN	0,633	0,420	0,400	0,402	0,420

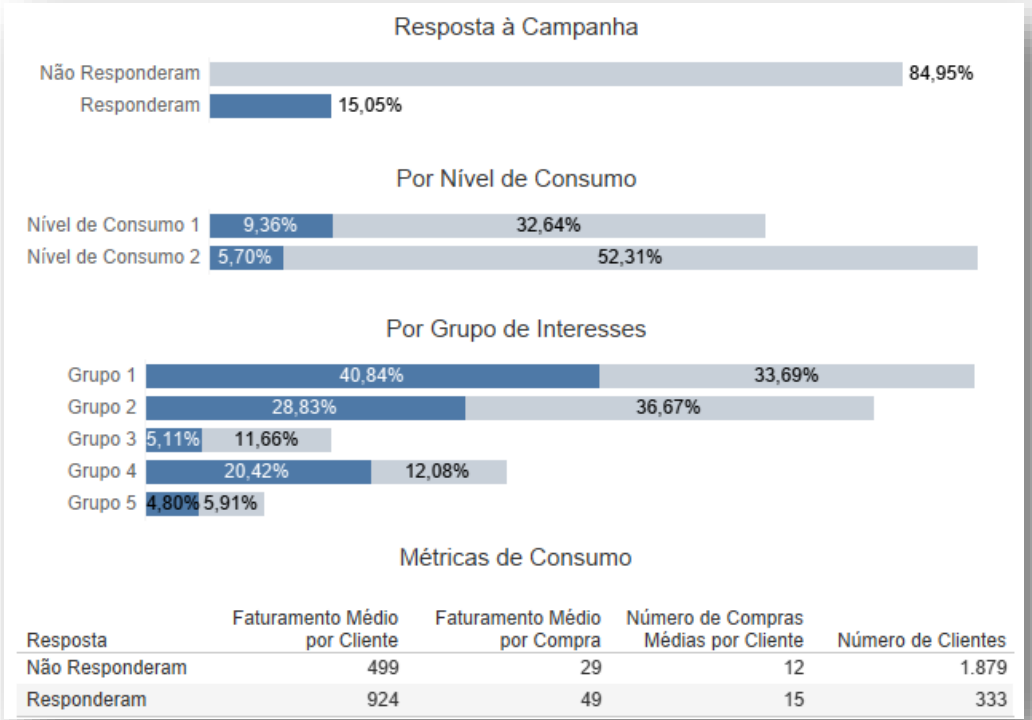
4.
A Campanha de Marketing



Perfil dos Clientes que Responderam à Campanha

A campanha está alinhada ao perfil dos clientes com maiores padrões de consumo

- A maior parte dos consumidores que responderam a campanha enquadram-se nos níveis de consumo 1 e grupos de interesses 1, 2 e 4.
- Observam-se diferenças percentuais de 85,2%, 69,0% e 25,0% para faturamento médio por cliente, faturamento médio por compra e números de compras médio por clientes, respectivamente.



Métricas de Consumo

Resposta	Faturamento Médio por Cliente	Faturamento Médio por Compra	Número de Compras Médias por Cliente	Número de Clientes
Não Responderam	499	29	12	1.879
Responderam	924	49	15	333

Otimizando Resultados

Vamos Maximizar o Direcionamento e o Retorno Financeiro da Campanha

Otimizando Resultados

Estratégia

- O algoritmo **Decision Tree (Scikit Learn)** foi escolhido em razão da fácil interpretabilidade do modelo;
- Foram testados **diferentes cutoffs** para a probabilidade do modelo: de 0,15 a 0,7, variando-se de 0,05 em 0,05;
- A amostragem e treino foram **repetidos por 100 vezes** para cada cutoff testado, a fim de aferirem-se as **médias e desvios-padrão das métricas**;
- Os dados gerados foram usados para **modelar o retorno financeiro previsto da campanha e sua provável variação** (nível de confiança de 95%), considerando diferentes cutoffs para as predições de probabilidade;

Validação

- A cada repetição, o dataset foi divido em grupos distintos de treino e teste, na **proporção de 80% para treino e 20% teste**;

Variáveis de Entrada

- InterestGroup, ConsuptionLevel e AcpCamps;
- Optou-se por poucas variáveis de entrada buscando-se criar um modelo de baixa complexidade, humanamente explicável, e de alto poder de generalização;

Metaparâmetros

- max_depth = 10

Otimizando Resultados

Métricas

- Algumas métricas foram criadas especificamente para o problema de maximização do lucro da campanha, levado em consideração os dados de custo e faturamento levantados com a campanha piloto ($3 custo de contato, $11 faturamento por venda):

Eficiência (%):	Lucro por Disparo ($):	Proporção de Disparos (%):
$\frac{\text{[lucro obtido]}}{\text{[lucro máximo possível, considerando acerto de 100\% das predições]}} * 100$	$\frac{\text{[lucro obtido]}}{\text{[número de clientes alvo da campanha]}}$	$\frac{\text{[número de clientes alvo da campanha]}}{\text{[número de clientes na base dados]}} * 100$

- Também foi avaliado: **Acurácia, Precisão e Recall;**

Modelagem do Lucro Previsto para cada 10 mil clientes na base de dados:

- Usados 2 desvios-padrão para confiança de 95%;

Previsão Pessimista:

$$[lucro\ prev] = \min\left((média[lucro\ disparo] - 2 * desvio\ padrão[lucro\ disparo]) * \frac{(média[proporçao\ disparos] \pm 2 * desvio\ padrão[proporçao\ disparos])}{100} * 10.000\right)$$

Previsão Média (mais provável):

$$[lucro\ prev] = média[lucro\ disparo] * \frac{média[proporçao\ disparos]}{100} * 10.000$$

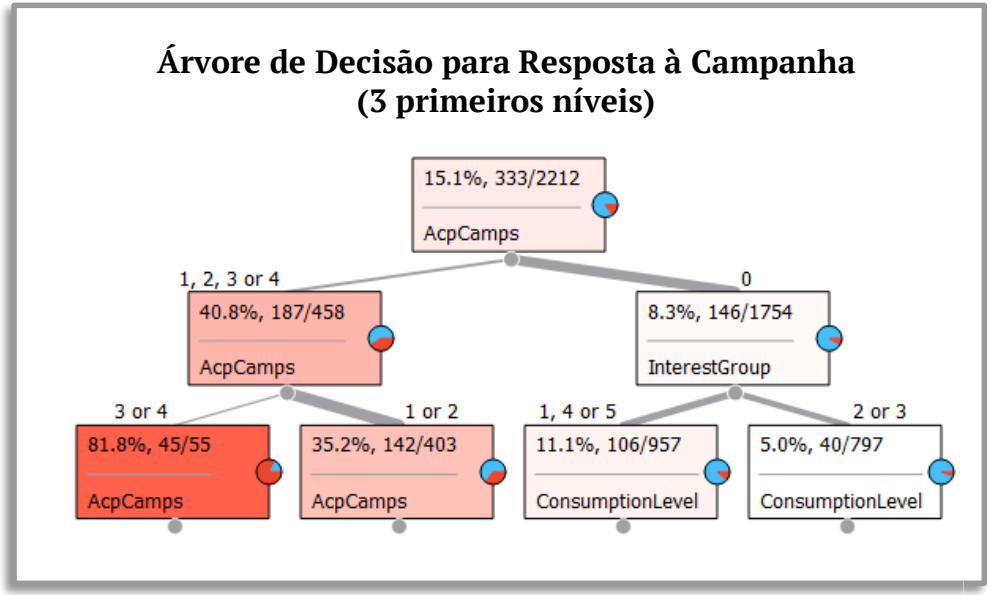
Previsão Otimista:

$$[lucro\ prev] = (média[lucro\ disparo] + 2 * desvio\ padrão[lucro\ disparo]) * \frac{(média[proporçao\ disparos] + 2 * desvio\ padrão[proporçao\ disparos])}{100} * 10.000$$


RESULTADOS

RESULTADOS

A principal variável determinante é o **AcpCamps**, o que demonstra que **clientes com histórico de resposta a campanhas anteriores se mostraram muito mais receptivos à campanha atualmente em estudo**.



As métricas obtidas foram as seguintes:

	Eficiencia (%)		Lucro por Disparo ($)		Proporção de Disparos (%)		Acuracia		Precisao		Recall	
Cutoff	Média	Std	Média	Std	Média	Std	Média	Std	Média	Std	Média	Std
0,15	16,5	5,89	1,34	0,55	20,78	2,52	0,81	0,02	0,39	0,05	0,55	0,05
0,2	17,72	4,89	1,59	0,57	19,16	2,35	0,82	0,02	0,42	0,05	0,52	0,05
0,25	17,63	4,71	2,02	0,76	15,36	3,01	0,83	0,02	0,46	0,07	0,45	0,07
0,3	18,75	4,3	2,71	0,75	11,67	1,79	0,85	0,01	0,52	0,07	0,4	0,06
0,35	16,92	4,17	2,93	0,77	9,82	2,24	0,86	0,01	0,54	0,07	0,35	0,07
0,4	16,05	3,62	3,7	0,97	7,45	1,67	0,86	0,01	0,61	0,09	0,3	0,06
0,45	14,98	3,58	3,87	0,92	6,37	1,18	0,87	0,01	0,62	0,08	0,27	0,05
0,5	13,62	3,66	4,32	1,2	5,19	1,19	0,87	0,01	0,67	0,11	0,23	0,05
0,55	13,23	3,42	4,79	1,21	4,62	1,09	0,87	0,01	0,71	0,11	0,22	0,05
0,6	12,04	3,24	5,1	0,99	4,02	1	0,87	0,01	0,74	0,09	0,19	0,04
0,65	10,11	3,19	4,78	1,33	3,6	0,98	0,86	0,01	0,71	0,12	0,17	0,04
0,7	7,86	2,55	5,14	1,44	2,66	0,88	0,86	0,02	0,74	0,13	0,13	0,04

(calma, já vai ficar mais fácil de entender...)

RESULTADOS

Simulação de Lucro Previsto da Campanha, com Modelo Árvore de Decisão, para cada Dez Mil Clientes Cadastrados na Base de Dados

- O **cutoff em 0,3** apresentou as maiores previsões de lucro médio e pessimista, ao passo que 0,25 traz uma previsão otimista 17,8% maior, porem com previsão pessimista 52,2% inferior.

- **A recomendação é pelo cutoff em 0,3**.

- Cutoffs entre 0,35 e 0,6 podem ser empregados caso deseje-se promover uma campanha menor (visando reduções de custo), porém com perspectivas de lucro significativamente menores.

Cutoff	Previsão Pessimista ($)	Previsão Média ($)	Previsão Otimista ($)
0,15	377,76	2.784,52	6.300,08
0,2	650,70	3.046,44	6.513,78
0,25	467,00	3.102,72	7.568,52
0,3	978,89	3.162,57	6.420,25
0,35	742,26	2.877,26	6.392,10
0,4	723,36	2.756,50	6.085,56
0,45	814,03	2.465,19	4.984,83
0,5	539,52	2.242,08	5.087,04
0,55	578,28	2.212,98	4.902,80
0,6	630,24	2.050,20	4.262,16
0,65	347,68	1.720,80	4.136,64
0,7	203,40	1.367,24	3.544,84

Então, considerando o cutoff em 0,3 teríamos...

$ 3.162,57

…de lucro médio previsto para cada 10.000 clientes na base de dados…

11,8±1,8 %

...de taxa de abrangencia esperada, em relação a base de dados!

52±7 %

...de taxa de resposta esperada!

$2,71±0,75

...de lucro previsto por disparo!

18,7±4,3 %

...é a eficiência do modelo, e...

85±1 %

...a acurácia das predições!

Modelos mais robustos e com mais variáveis de entrada podem desempenhar ainda melhor, porém com interpretabilidade reduzida.

5.
Próximos Passos

Próximos Passos

Campanha de Marketing Direto

- Recomenda-se a **continuidade da campanha com novo lote de clientes**, segmentando-se com base no modelo treinado, para teste prático da metodologia desenvolvida e apuração de novos resultados;

Segmentação dos Clientes

- Os modelos desenvolvidos podem ser usados para **construção de campanhas futuras mais específicas e personalizadas** ao público correspondente;
- Merecem especial atenção no marketing os clientes do grupo 2, que possuem grande interesse por vinhos, e do grupo 4, maiores consumidores dos produtos da linha Gold.

Prospecção de Novos Clientes

- A empresa deve buscar principalmente por **clientes potenciais com renda acima de $ 54.690** e que não possuem crianças (características correspondentes ao nível de consumo 1);
- Os modelos treinados na seção 5.1. podem ser empregados para **busca por clientes potenciais em outras bases de dados**.

Pontos de Força

- As linhas de destaque são **Vinhos e Produtos de Carne**, que correspondem aos principais interesses da maior parte dos clientes;
- **Vendas por catálogo** são o canal de preferência dos consumidores com padrão mais elevado de consumo.

Obrigado!

Perguntas?

Você me encontra em

(11) 97444-2729 | eformica@gmail.com
www.linkedin.com/in/eucformica