Desafio Q-Bem – Cientista de Dados

Descrição do Desafio

Os analistas de dados que trabalham na equipe de dados são constantemente desafiados a fornecer insights e valor à empresa por meio de projetos de escopo aberto. Este caso pretende simular isso. Nesse caso, você encontrará um conjunto de dados de amostra, é uma meta-informação simulada sobre as interações da campanha do cliente com esse cliente.

É seu desafio entender os dados, encontrar oportunidades e insights de negócios e propor qualquer ação orientada por dados para otimizar os resultados da campanha e gerar valor para a empresa. Este caso visa avaliar suas habilidades e conhecimento de dados para funções de análise avançada de dados: realize análises exploratórias robustas, usando ferramentas avançadas de análise e métodos estatísticos para gerar produtos de dados para otimizar resultados de negócios (modelos preditivos e de clusterização, por exemplo).

Os principais objetivos e entregas são:

- Explore os dados – seja criativo e preste atenção aos detalhes. 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?

- Propor e descrever uma segmentação de clientes 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?

- Crie um modelo preditivo que permita à empresa 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 que você não domina.

- Faça uma apresentação de negócios altamente eficaz: Lembre-se de que 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.

- Anexe os artefatos jpnb à sua resposta, comente os pontos principais e o motivo da escolha das bibliotecas e/ou frameworks. Mantenha este caso confidencial, não o publique no github e envie-o no prazo máximo de duas semanas.

A empresa

Considere uma empresa bem estabelecida que opera no setor de varejo de alimentos. Atualmente eles têm cerca de centenas de milhares de clientes cadastrados e atendem a quase um milhão de consumidores por ano.

Vendem produtos de 5 grandes categorias: vinhos, carnes raras, frutas exóticas, especialmente peixe preparado e produtos doces. Estes podem ainda ser divididos em ouro e produtos regulares. O os clientes podem encomendar e adquirir produtos através de 3 canais de venda: lojas físicas, catálogos e o site da empresa. Globalmente, a empresa teve receitas sólidas e uma linha de fundo saudável no últimos 3 anos, mas as perspectivas de crescimento do lucro para os próximos 3 anos não são promissoras...

Por isso, várias iniciativas estratégicas estão sendo consideradas para reverter essa situação. Uma é melhorar a execução das atividades de marketing, com especial enfoque nas campanhas de marketing.

O Departamento de Marketing

O departamento de marketing foi 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 com um objetivo claro em mente: construir um modelo preditivo que apoiará as iniciativas de marketing direto. Desejavelmente, o sucesso dessas atividades provará a valor da abordagem e convencer os mais céticos dentro da empresa.

Objetivo

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, sexta, visa vender um novo gadget para o banco de dados de clientes. Para construir o modelo, uma campanha piloto envolvendo Foram realizados 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 rotuladas. 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. Globalmente, a campanha teve um lucro de -3.046MU. A taxa de sucesso da campanha foi de 15%. O objetivo da equipe é desenvolver um modelo que preveja o comportamento do cliente e aplicá-lo ao restante da base de clientes.

Felizmente, o modelo permitirá que a empresa escolha os clientes com maior probabilidade de compre a oferta deixando de fora os não respondentes, tornando a próxima campanha altamente rentável. Além disso, além de maximizar o lucro da campanha, o CMO está interessado em entendimento das características dos clientes que estão dispostos a comprar o gadget.

Desafio Q-Bem – Entrega (Mar/2022)

1 – Entendendo e Manipulando os Dados

O dataset é constituído por 2240 registros simulados de respostas dos clientes a pesquisa de marketing. Possui 29 variáveis, sendo 15 numéricas e 14 categóricas.

Para desenvolvimento do projeto, algumas transformações foram necessárias a fim de criar features utéis à modelagem:

- As variáveis Year_Birth e Dt_Customer foram usadas para o cálculo da idade na data do registro (Age), e então removidas;

- Assumiu-se que as variáveis MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts referem-se ao valor mensal gasto pelo cliente com produtos das respectivas categorias. Essas variáveis foram somadas (variável TotalProducts) e transformadas em valores relativos percentuais referentes a esses gastos, a fim de tentar identificar os tipos de produto mais consumidos pelo cliente, independentemente de seu gasto mensal. Adicionou-se o prefixo “Rlt_” aos nomes originais. Os dados originais não foram preservados.

- Da mesma forma, as variáveis NumWebPurchases, NumCatalogPurchases, NumStorePurchases também foram somadas (TotalPurchases) e transformadas em valores relativos percentuais, referentes a preferência do consumidor pelo canal de vendas, independentemente do número de compras. Adicionou-se o prefixo “Rlt_” aos nomes originais. Os dados originais não foram preservados.

- Assumiu-se que as variáveis AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4 e AcceptedCmp5 referem-se ao histórico de resposta do cliente a 5 campanhas de marketing anteriores (codificação one-hot). Para resumir esses dados, as variáveis foram somadas (AcpCamps). Os dados originais foram mantidos.

- A variável Education foi codificada ordinalmente: 1 = Basic; 2 = 2n Cycle; 3 = Graduation; 4 = Master, 5 = PhD.

- Os valores “Alone”, “Absurd” e “YOLO” em Marital_Status foram alterados para “Single”;

- A variável Marital_Status foi codificada com o método one-hot (um campo para cada valor, com 1 = correspondente e 0 = não correspondente);

- Z_CostContact e Z_Revenue foram removidas, vez que apresentam valores constantes em todo o dataset (não é útil para machine learning);

- As variáveis ID e Recency também foram removidas, por não serem pertinentes na presente análise;

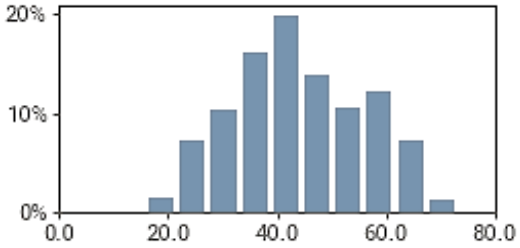
A análise exploratória preliminar foi realizada com auxílio da biblioteca Sweetviz do Python (vide arquivo report1_sweetviz.html). Com base nessa análise, as seguintes manipulações foram realizadas no dataset original:

- Imputação dos valores ausentes em Income pela mediana, e remoção do outlier (666666);
- Remoção de outliers em Age > 100;

A biblioteca Sweetviz foi novamente usada para geração de novo relatório automático (report2_sweetviz.html). Também foi gerado um heatmap de correlações, com auxílio do pacote Seaborn, cujos resultados encontra-se analisados na seção 3.

2 – Perfil dos Clientes[1]

Idade



A idade média é de 44 anos, com mínimo em 16 e máximo em 73 anos[2]

60,3% dos clientes possui idade entre 34 e 56 anos

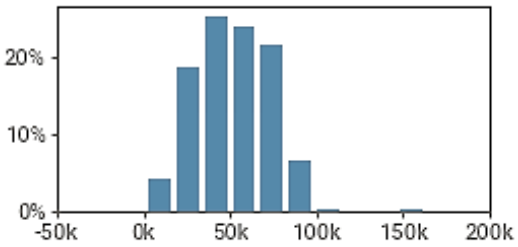
19% com menos de 34 anos, e 20,7% com mais de 56 anos

Renda

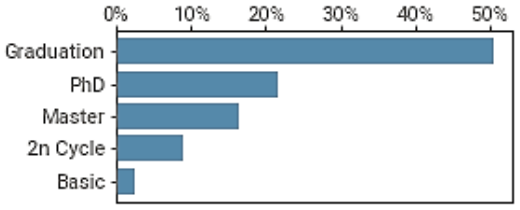
A renda média é de $ 51.958,81

73,55% dos clientes possui renda entre $30.000 e $80.000

Apenas 0,5% dos clientes possui renda superior a $100.000



Escolaridade



50,4% dos clientes possui Graduação

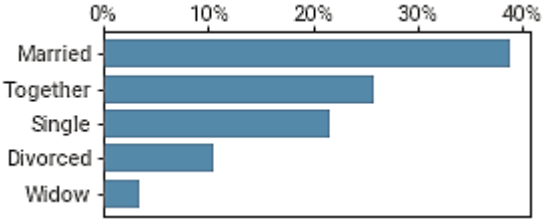
33,2% possuem PhD ou Master

11,5% não possuem ensino superior

Estado Civil

64,5% são casados ou vivem em união estável

21,6% são solteiros, enquanto divorciados e viúvos somam 13,8%



[1] Estatísticas referentes à amostra em estudo.
[2] Foram descartados 3 prováveis *outliers* com idades de 113, 114 e 121 anos.

3 – Correlações

A análise preliminar de correlações entre variáveis permitiu algumas constatações interessantes:

- Clientes com maior renda:
    - Compram mais (tanto em quantidade de compras, como em volume das compras e produtos por compra);
    - Visitam menos o Website da empresa (correlação forte);
    - Geralmente possuem poucas crianças em casa (correlação forte);
    - Compram mais produtos Gold e por catálogo;

- Clientes com crianças:
    - Geram menos vendas;
    - Mais acessos ao website são observados nesse grupo (correlação moderada);
    - Compram menos produtos Gold e por Catálogo;

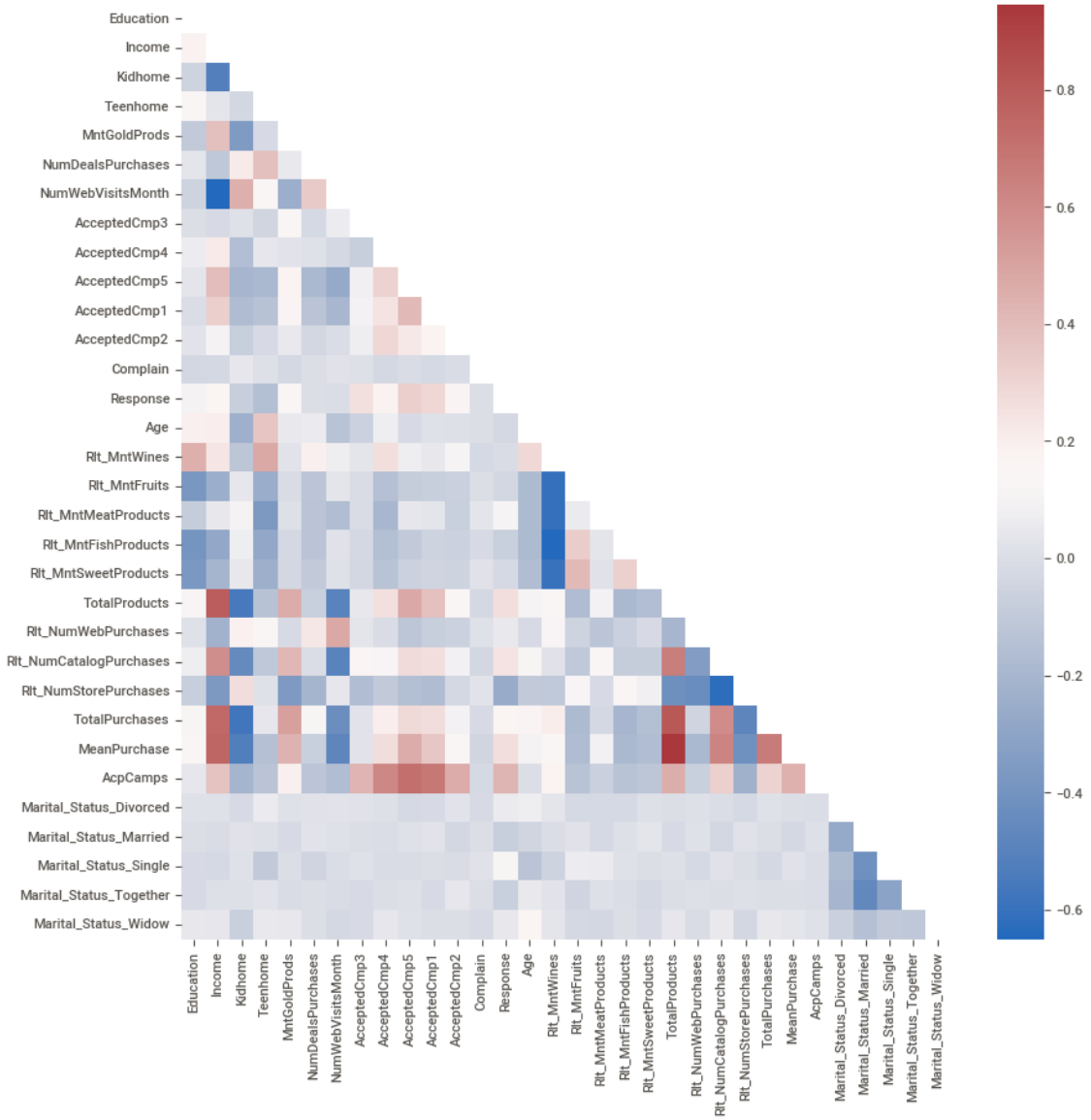
- Clientes com maior compra de vinho compraram menos outras modalidades de produtos da empresa (correlação forte);

- Vendas por catálogo geram mais vendas e maior relação gasto por compra (correlação forte).

- Clientes que compram por catálogo compram menos na loja física (correlação forte).

- Clientes que compram na loja física compram menos, e com uma relação gasto/compra (correlação moderada).

Correlações



4 – Segmentação de Clientes

4.1. Segmentação por Nível de Consumo

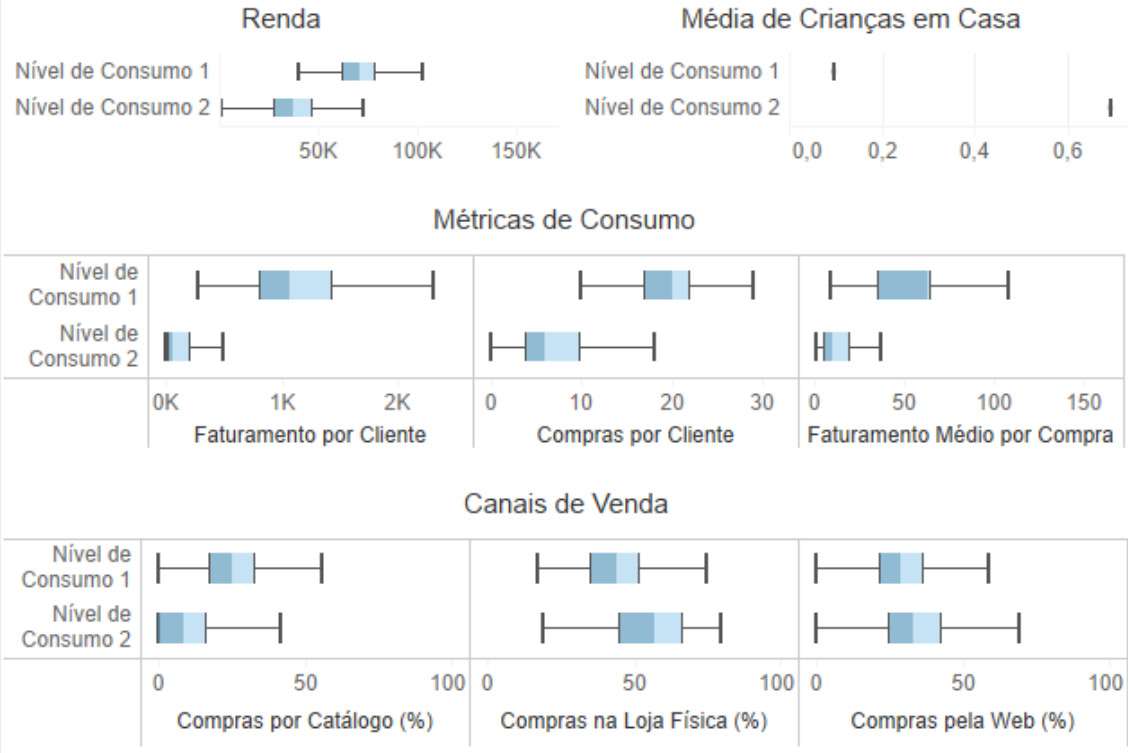
A fim de identificar os clientes com maior padrão de consumo, treinou-se o algoritmo K-Means com as seguintes variáveis de entrada: "TotalProducts", "TotalPurchases", "MeanPurchase". Os dados foram padronizados pelo método Standart Scaler (scikit learn).

O número de cluster foi definido em 2, uma vez que o objetivo é obter uma classificação binária. Além disso, a clusterização em 2 grupos apresentou o maior Silhouette Score[3] (0,62), tendo sido testado com até 10 grupos, o que demonstra a adequação do agrupamento. A variável gerada, contendo as predições foi nomeada como “ConsumptionLevel”.

Os dados foram analisados com os programas Orange Data Mining[4] e Tableau[5]. As principais diferenças observadas encontram-se descritas a seguir:

Principais Diferenças entre os Grupos de Nível de Consumo

	Faturamento Médio por Compra	Faturamento Médio por Cliente	Número de Compras Médias por Cliente	Renda Média	Numero de Clientes no Grupo
Nível de Consumo 1	57	1.147	20	70.951	929
Nível de Consumo 2	13	140	7	38.207	1.283



Esses números demonstram significativa diferença no padrão de consumo dos grupos de clientes identificados, dados que **os 929 consumidores do nível 1 compram**

[3] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html
[4] https://orangedatamining.com/
[5] https://www.tableau.com/

em média 8,2 vezes mais produtos e apresentam renda média cerca de 1,8 vezes maior que os 1.283 clientes do nível 2.

Além disso, clientes do nível 1 compram mais por catálogo, e possuem geralmente poucas crianças em casa.

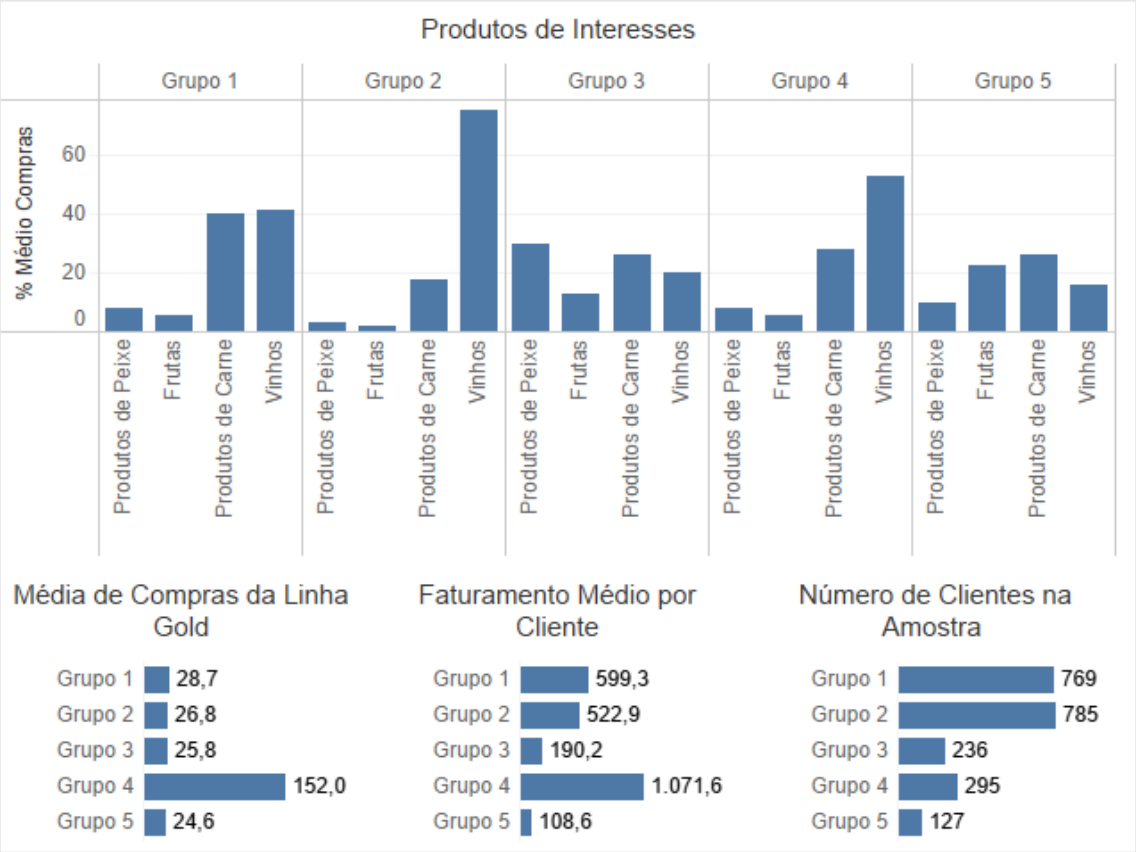
4.2. Segmentação por Produtos de Interesse

Para segmentação por produto de interesse, treinou-se o algoritmo K-Means com as seguintes variáveis de entrada: "Rlt_MntWines", "Rlt_MntFruits", "Rlt_MntMeatProducts", "Rlt_MntFishProducts", "Rlt_MntSweetProducts", "MntGoldProds". Os dados foram padronizados pelo método Standart Scaler (scikit learn), pois a variável MntGoldProds apresenta valores em escala diferente (não é percentual).

Os métodos Silhouette Score e Elbow Curve foram avaliados para determinação do número ideal de clusters, que resultou em 6 para o Elbow Curve e 5 para Silhouette Score. Optou-se pela utilização do método Silhouette Score, vez que alcanço uma boa métrica de separação (Silhouette Score=0,31 para k=5), com um grande decaimento do índice para 6 clusters (Silhouette Score=0,24 para k=6).

Os grupos gerados, “InterestGroup”, foram analisados com os programas Orange Data Mining e Tableau. As seguintes características foram observadas:

Características por Grupo de Interesses



Pode-se resumir da seguinte forma:

- **GRUPO 1**: Interesse principal por 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;

Esses resultados indicam que **as campanhas de marketing devem concentrar seus esforços principalmente nos segmentos de Vinhos e Produtos de Carne**,

assim como **ampliar a carteira de clientes com interesse nos produtos da linha Gold (grupo 4)**.

5 – Predição do Perfil do Consumidor

5.1. Modelos para Prospecção de Novos Clientes

A fim de gerar um modelo que permita identificar o perfil de clientes potenciais (ou seja, que não possuem padrão de consumo conhecido), diversos modelos foram treinados usando as variáveis Education, Income, Kidhome, Teenhome, Age e Material_Status como entrada, e ConsumptionGroup e InterestGroup como alvo, quais: Naive Bayes, Random Forest, Gradient Boosting, kNN, Tree, AdaBoost e SVM.

Para otimizar a realização desse processo e comparação dos resultados, essa etapa foi inteiramente desenvolvida no Orange Data Mining[6]. Foi empregada validação cruzada K-Fold, com k=5, para avaliação dos modelos, analisando-se as métricas AUC, Acurácia (CA), F1, Precisão e Recall. Os metaparâmetros utilizados encontram-se acostados no Anexo 1. Seguem os resultados:

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

[6] O arquivo modelos.ows foi usado como molde para todos os treinamentos de modelos preditivos realizados, alterando-se apenas as Features e Target em “Select Columns” e usando o widget “Test and Score” para gerar as métricas de desempenhos.

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

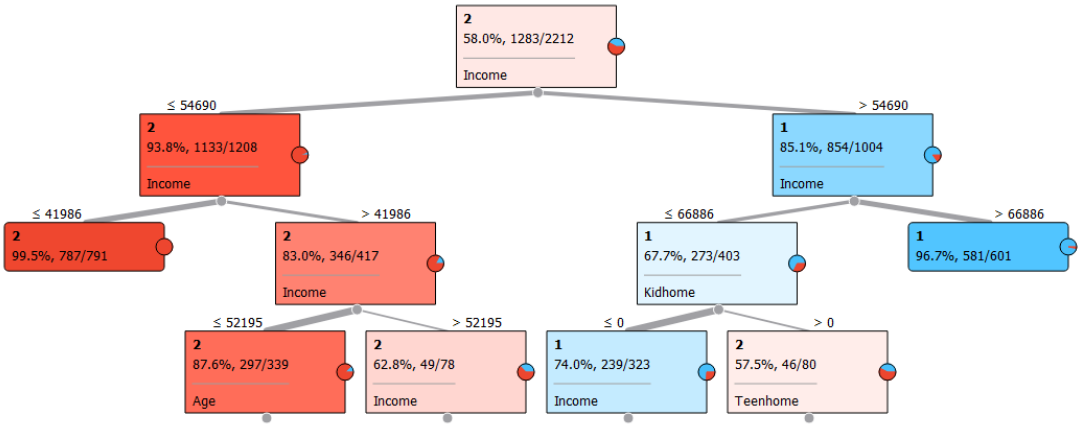
Como esperado, modelos para precisão do nível de consumo desempenharam melhor do que para grupos de interesse, pois o número de grupos a serem classificados são apenas 2 (e não 5), o que facilita a tarefa de classificação. Ainda assim, os modelos para determinação do grupo de interesses atingiram métricas aceitáveis para tarefa de orientação de anúncios específicos para determinado perfil de consumidor, uma vez que não se trata de uma tarefa crítica, na qual erros de classificação podem comprometer seriamente o resultado do negócio. Um banco de dados com mais informações, como, por exemplo, sexo e endereço do cliente, histórico das compras por data (pode existir sazonalidade nos padrões de compra) etc, poderia melhorar significativamente o desempenho dos modelos gerados.

Dentre os testes realizados, o Gradiente Boosting apresentou o melhor desempenho AUC (área sobre a curva ROC)[7] em ambos os alvos. Trata-se de um modelo composto por diversas árvores de decisões sequenciais, em que cada árvore é treinada para "corrigir" sua antecessora[8]. É um modelo robusto, adequado para utilização automática, mas cuja lógica interna é de difícil representação e interpretação humana. Em contraponto, o modelo Tree (Árvore de Decisão), pode ser facilmente visualizado e interpretado, mas, por sua simplicidade, costuma não desempenhar tão bem.

A imagem a seguir apresenta os 4 primeiros níveis do modelo de árvore de decisão treinado para predição do grupo de consumo:

[7] https://medium.com/kunumi/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-em-machine-learning-classifica%C3%A7%C3%A3o-49340dcdb198
[8] https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab

Árvore de Decisão para Predição do Grupo de Consumo (4 primeiros níveis)



Verifica-se que a renda é o critério mais determinante para essa aferição, vez que, logo no primeiro nível, **rendas acima de $54.690 permitem selecionar 85,1% dos clientes do grupo de consumo 1, ao passo que abaixo desse valor estão 93,8% dos clientes do grupo de consumo 2**.

5.2. Modelos para Marketing Direcionado aos Clientes Castrados

O cadastro de clientes e histórico de compras podem ser usados para criação de campanhas personalizadas aos grupos de interesses e consumo. Como as informações sobre o perfil de consumo estão disponíveis, a grande maioria dos modelo de Machine Lerning deverá retornar previsões bastante confiáveis sobre quais grupos o cliente se enquadra, podendo inclusive atualizar essas predições a cada nova compra, pois os dados de entrada serão os mesmos que foram utilizados para a definição dos clusteres pelo K-Means.

Apenas a título de ilustração, a tabela a seguir apresenta o resultado do treinamento dos mesmos modelos da etapa anterior, para predição do grupo de interesse, tendo como variáveis de entrada as mesmas utilizadas para o K-Means: *Rlt_MntWines, Rlt_MntFruits, Rlt_MntMeatProducts, Rlt_MntFishProducts, Rlt_MntSweetProducts, MntGoldProds:*

Modelos Avaliados para Predição do Grupo de Interesses
(Mesmas Variáveis de Entrada Usadas no K-Means, Média sobre as Classes)

Model	AUC	CA	F1	Precision	Recall
SVM	1,000	0,977	0,977	0,977	0,977
Gradient Boosting	0,999	0,968	0,968	0,968	0,968
Random Forest	0,999	0,963	0,962	0,963	0,963
kNN	0,993	0,947	0,946	0,947	0,947
Naive Bayes	0,965	0,796	0,793	0,798	0,796
AdaBoost	0,963	0,947	0,947	0,947	0,947
Tree	0,957	0,933	0,932	0,932	0,933

6 – Avaliação da Campanha de Marketing

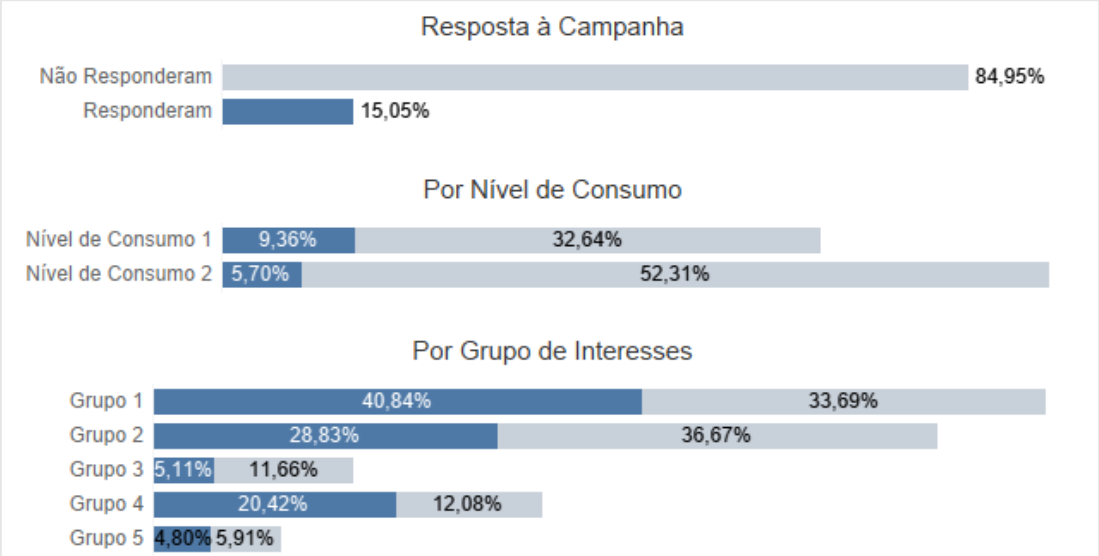
6.1. Métricas da Campanha Piloto

Conforme informações, o custo de contato foi $3 por cliente (total de $6.720 para os 2.240 clientes). Responderam a campanha 337 clientes (15,05 %). O lucro obtido por cliente que respondeu a campanha foi de $11 ($3.707). O prejuízo total foi de $3.013.

6.2. Perfil dos Clientes que Responderam à Campanha

A resposta à campanha foi avaliada com o Software Tableau. Seguem as principais observações:

Perfil dos Clientes que Responderam à Campanha



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

Observa-se que a maior parte dos consumidores que responderam a campanha enquadram-se nos níveis de consumo 1 e grupos de interesses 1, 2 e 4, portanto **a campanha está alinhada ao perfil dos clientes com maiores padrões de consumo**.

Pode-se contatar essa afirmação também pela comparação das métricas de consumo entre os clientes que responderam e não responderam à campanha: diferenças percentuais de 85,2%, 69,0% e 25,0% para faturamento médio por cliente, faturamento médio por compra e número de compras médio por clientes, respectivamente.

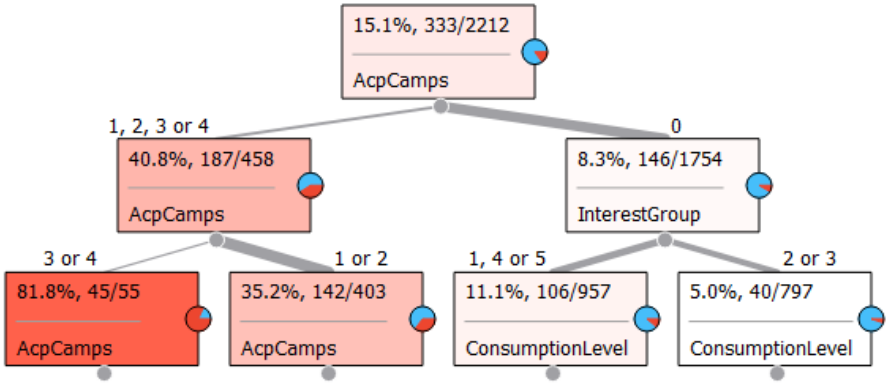
6.3. Modelo para Maximização do Lucro da Campanha e Projeções

Para maximizar o direcionamento e retorno financeiro da campanha, foi treinado um modelo de árvore de decisão com variáveis de entrada InterestGroup, ConsuptionLevel e AcpCamps. A árvore de decisão com poucas variáveis foi escolhida buscando-se criar um modelo de baixa complexidade, humanamente explicável, e de alto poder de generalização, mas também foram testados modelos mais robustos e com mais features (idade, renda, grau de instrução, etc), alguns que, inclusive,

desempenharam ligeiramente melhor. Apesar disso, a árvore de decisão treinada a partir de 3 features foi o modelo de escolha, em razão de sua fácil interpretabilidade.

A imagem a seguir representa os 3 primeiros níveis de uma das árvores geradas[9]:

Árvore de Decisão para Resposta à Campanha (3 primeiros níveis)



Observa-se que 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.

O modelo foi gerado com Python, visando facilitar o ajuste do cutoff e modelagem do problema de maximização do lucro. O único parâmetro passado para o DecisionTreeClassifier (scikit learn) foi max_depth = 10.

Para cada ensaio, o treino e teste foram realizados em subamostras distintas, na proporção de 80% treino, 20% teste.

Algumas métricas foram criadas especificamente para o problema de maximização do lucro da campanha, levado em consideração os dados de custo e faturamento declinados na seção 8.1. São elas:

- Eficiência (%):

$$\frac{[\text{lucro obtido}]}{[\text{lucro máximo possível, considerando acerto de 100\% das predições}]} * 100$$

[9] A árvore inteira pode ser visualizada na pasta gráficos/arvore2.png que acompanha esse relatório.

- Lucro por Disparo ($):

$$\frac{[\text{lucro obtido}]}{[\text{número de clientes alvo da campanha}]}$$

- Proporção de Disparos (%):

$$\frac{[\text{número de clientes alvo da campanha}]}{[\text{número de clientes na base dados}]} * 100$$

Também foram avaliadas as seguintes métricas-padrão de classificação: Acurácia, Precisão e Recall.

Testaram-se diferentes cutoffs (pontos de corte) sobre as predições de probabilidade, variando de 0,15 a 0,70, de 0,05 em 0,05. Para cada cutoff, o treino foi repetido 100 vezes, a fim de aferir-se as médias e desvios-padrão de cada métrica em diferentes amostras de treino/teste. Os resultados estão apresentados na tabela a seguir:

Métricas do Modelo de Árvore de Decisão para Otimização do Direcionamento da Campanha

	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

Esses dados foram usados para modelagem de 3 cenários de desempenho da campanha, considerando sua aplicação em uma base de dados de 10.000 clientes cadastrados: pessimista, mais provável e otimista. As fórmulas foram as seguintes:

- Previsão Média (mais provável):

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

- Previsão Pessimista:

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

- Previsão Otimista:

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

As variações de 2 desvios-padrão empregadas nas fórmulas representam o nível de confiança de 95% para um resultado real entre a previsão otimista e pessimista.

Seguem resultados:

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

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

Consta-se que o cutoff em 0,3 apresenta as maiores previsões média e pessimista, ao passo que 0,25 traz uma previsão otimista 17,8% maior, porem com previsão pessimista 52,2% inferior. Assim, **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), porem com perspectiva de lucro significativamente menores.

Com a segmentação da campanha com o modelo treinado, com cutoff ajustado em 0,3, espera-se que a campanha atinja 11,67 ± 1,79 % dos clientes da base de dados, com taxa de retorno[10] de 52 ± 7 %, e um lucro por disparo de $ 2,71±0,75, o que resulta em um **lucro total previsto de aproximadamente $ 3.162,57 para cada dez mil clientes na base de dados**. **Essas métricas são muito superiores às obtidas com a campanha piloto**!

Ressalte-se que modelos mais robustos e com mais variáveis podem desempenhar ainda melhor, entretanto maiores estudos seriam demandados.

7 – Recomendações e Ações Possíveis

7.1. Campanha de Marketing Direto

Os dados demonstram que o escopo da campanha está bem focado aos clientes com maior perfil de consumo. O modelo desenvolvido na seção 6 permitiu uma seleção refinada de clientes com maiores chances de responder a ação de marketing, o que gerou uma excelente projeção de vendas.

Assim, 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.

7.2. Segmentação dos Clientes

Foram desenvolvidos modelos para segmentação dos clientes por grupos de interesse e perfil de consumo. Ambos os modelos dependem tão somente dos históricos de compras, não demandando nova pesquisa. Esses dados podem ser usados para

[10] Percentual de clientes que responderão a campanha, equivalente a métrica “precisão”.

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.

7.3. Prospecção de Novos Clientes

A empresa deve buscar principalmente por clientes com renda acima de $ 54.690 e que não possuem crianças (perfil correspondente 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.

7.4. 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.

Anexo 1 - Metaparâmetros Utilizados nos Modelos das Seções 5.1 e 5.2

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