



EDA Data Project



Introdução

APLICAÇÃO DE DATA ANALYSIS

Realizar a análise de uma base de dados concedida pelo iFood com informações sobre campanhas passadas

O desafio é de entender os dados, encontrar possíveis valores e previsões para o negócio e propor ações que sejam úteis para as campanhas da empresa

O objetivo final é avaliar as habilidades com dados a partir da base provida

SOBRE A BASE DE DADOS

Dados referentes a uma empresa que atua no setor do varejo, com vendas a partir de campanhas e operando entregas dos produtos

A empresa tem cinco categorias majoritárias de produtos, que são: vinhos, carnes, frutas, peixes e doces.

Os pedidos podem ser feitos em três canais, sendo eles lojas, catálogos e site da empresa.

O departamento de marketing está realizando estudos a partir de uma campanha piloto realizada com 2240 clientes, onde seis disparos de ofertas foram realizados.

A ideia é avaliar como foi a resposta dos clientes nesses disparos e realizar uma previsão para futuras campanhas

Olá!

Sou Daidson Alves

Você pode me encontrar
em:

<https://daidson.github.io>





1

Análise Exploratória

EDA - Observando os componentes

“

“The goal is to turn data into information, and information into insight.”

Carly Fiorina

Principais informações

- 2240 clientes;
- 29 tópicos de informações (features);
- Poucas features com valores nulos;
- Informações desde estado civil do cliente até a renda do mesmo.

Principais informações



Total number of missing values	Missing proportion in Data	Missing >= 5%
--------------------------------	----------------------------	---------------

Income	24	0.010714	False
--------	----	----------	-------

*Feature que apresentou
valores nulos (renda)*

*Informações
encontradas na base de
dados e quantos valores
únicos cada uma tem*

ID	2240
Year_Birth	59
Education	5
Marital_Status	8
Income	1974
Kidhome	3
Teenhome	3
Dt_Customer	663
Recency	100
MntWines	776
MntFruits	158
MntMeatProducts	558
MntFishProducts	182
MntSweetProducts	177
MntGoldProds	213
NumDealsPurchases	15
NumWebPurchases	15
NumCatalogPurchases	14
NumStorePurchases	14
NumWebVisitsMonth	16
AcceptedCmp3	2
AcceptedCmp4	2
AcceptedCmp5	2
AcceptedCmp1	2
AcceptedCmp2	2
Complain	2
Z_CostContact	1
Z_Revenue	1
Response	2

Disposição dos Dados

- 15 variáveis numéricas;
- 8 variáveis categóricas;
- 1 variável alvo;
- 5 variáveis com informações a serem tratadas;
- O alvo é a variável ‘Resposta’, que constata se o cliente respondeu positivamente ou não ao último disparo de oferta de marketing.

Disposição dos Dados

```
#dividing the data into numeric, categorical, target and other features
numeric_features = data[["Income", "Kidhome", "Teenhome", "Recency",
                        "MntWines", "MntFruits", "MntMeatProducts",
                        "MntFishProducts", "MntSweetProducts", "MntGoldProds",
                        "NumDealsPurchases", "NumWebPurchases", "NumCatalogPurchases",
                        "NumStorePurchases", "NumWebVisitsMonth"
                       ]]

categorical_features = data[["Education", "Marital_Status", "AcceptedCmp1",
                            "AcceptedCmp2", "AcceptedCmp3", "AcceptedCmp4",
                            "AcceptedCmp5", "Complain"
                           ]]

target = data[["Response"]
              ]

other_features = data[["ID", "Year_Birth", "Dt_Customer", "Z_CostContact", "Z_Revenue"]]
```

Classificação das variáveis da base de dados

Disposição dos Dados

- As variáveis 'Z_CostContact' e 'Z_Revenue' não mostraram significância nos dados pois exprimem apenas um valor único, logo serão descartadas;
- Iremos indexar as variáveis 'Dt_customer' e 'Year_birth' em nossos dados.

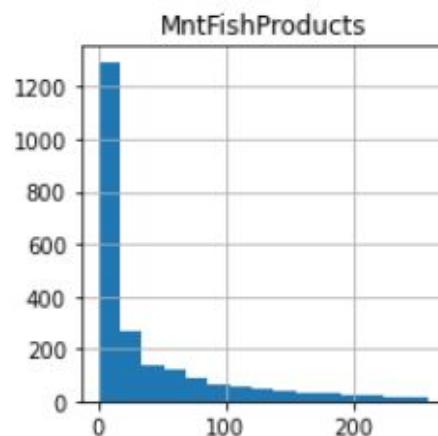
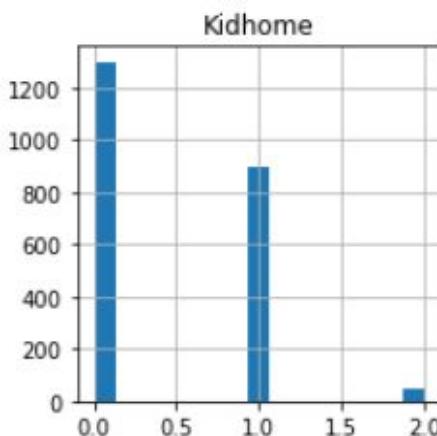
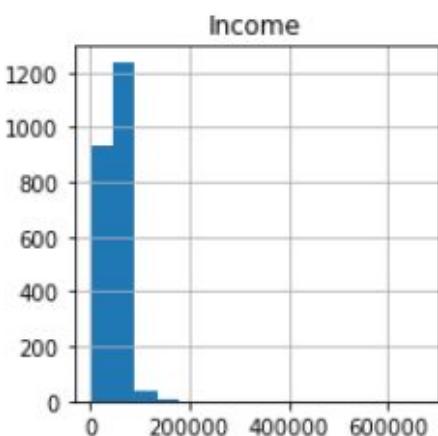
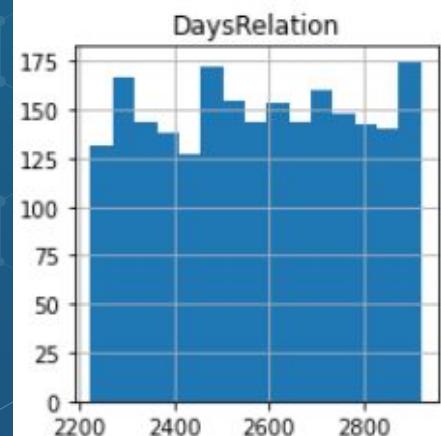


2

Segmentação de Clientes

Insights sobre nossos dados

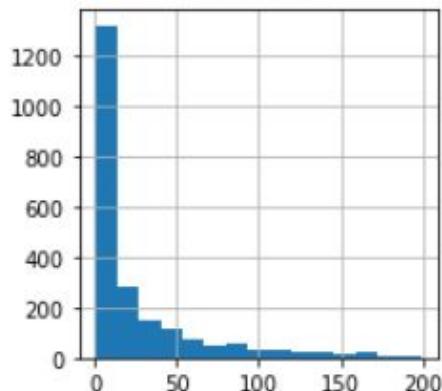
Distribuição das variáveis numéricas



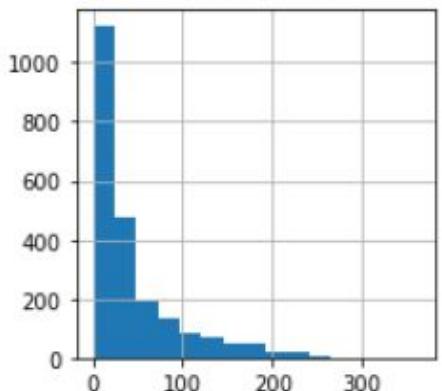
Linha horizontal representa valores distintos e vertical, quantidade

Distribuição das variáveis numéricas

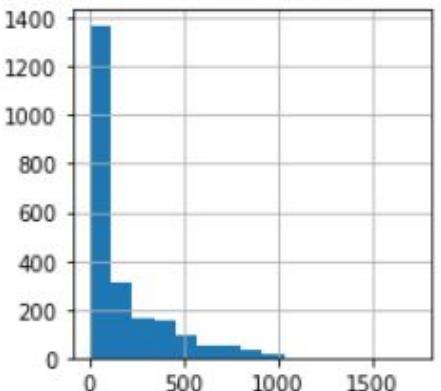
MntFruits



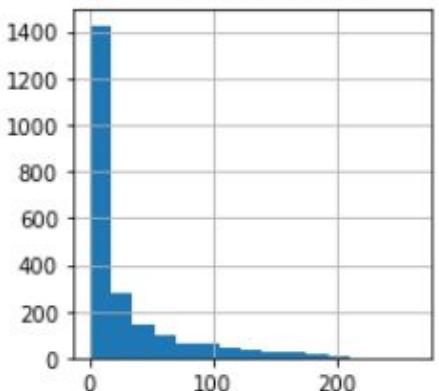
MntGoldProds



MntMeatProducts



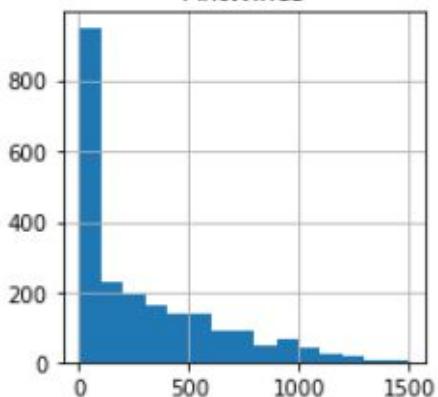
MntSweetProducts



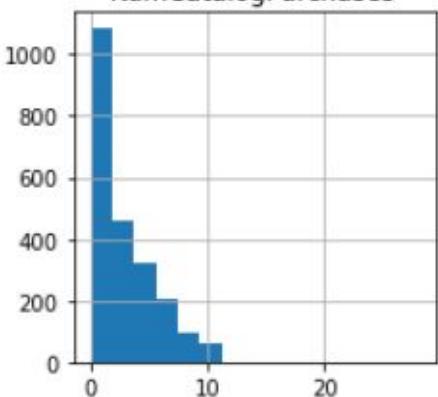
Linha horizontal representa valores distintos e vertical, quantidade

Distribuição das variáveis numéricas

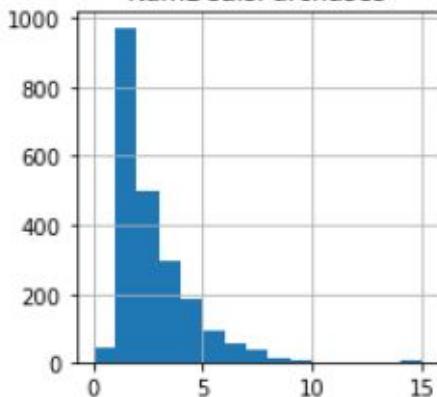
MntWines



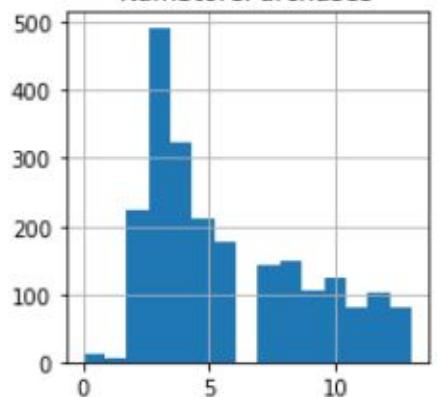
NumCatalogPurchases



NumDealsPurchases

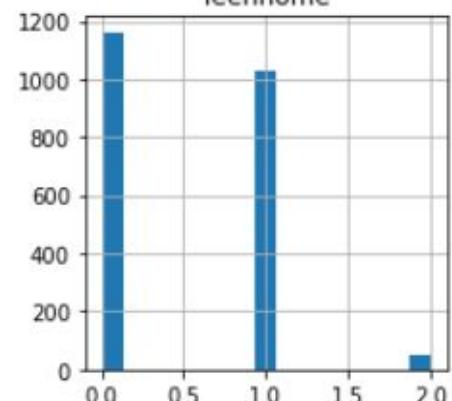
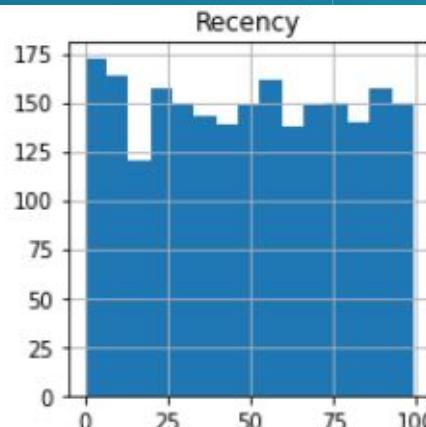
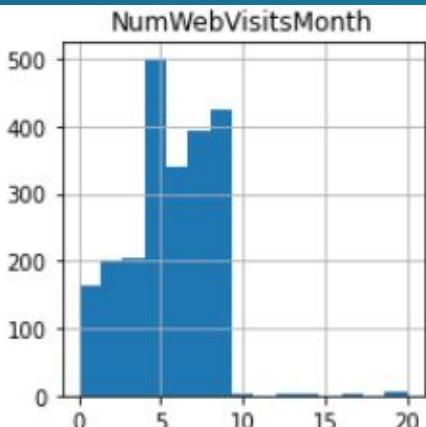
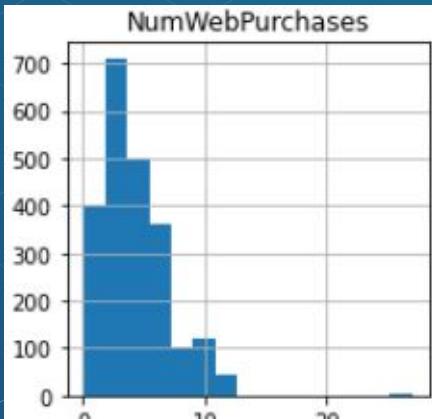


NumStorePurchases



Linha horizontal representa valores distintos e vertical, quantidade

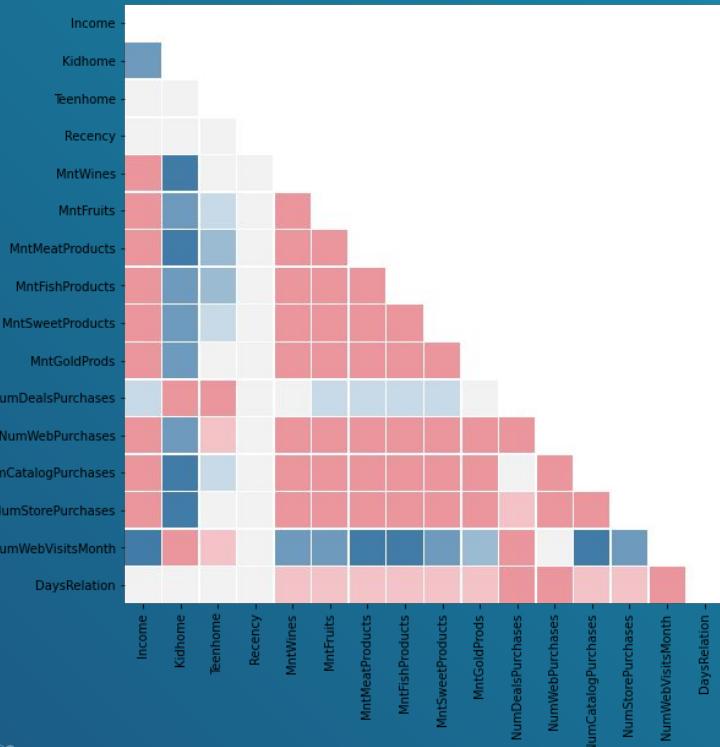
Distribuição das variáveis numéricas



Linha horizontal representa valores distintos e vertical, quantidade

Correlação das variáveis numéricas

Quão mais
correlacionadas as
variáveis são, maior o
grau no índice da legenda

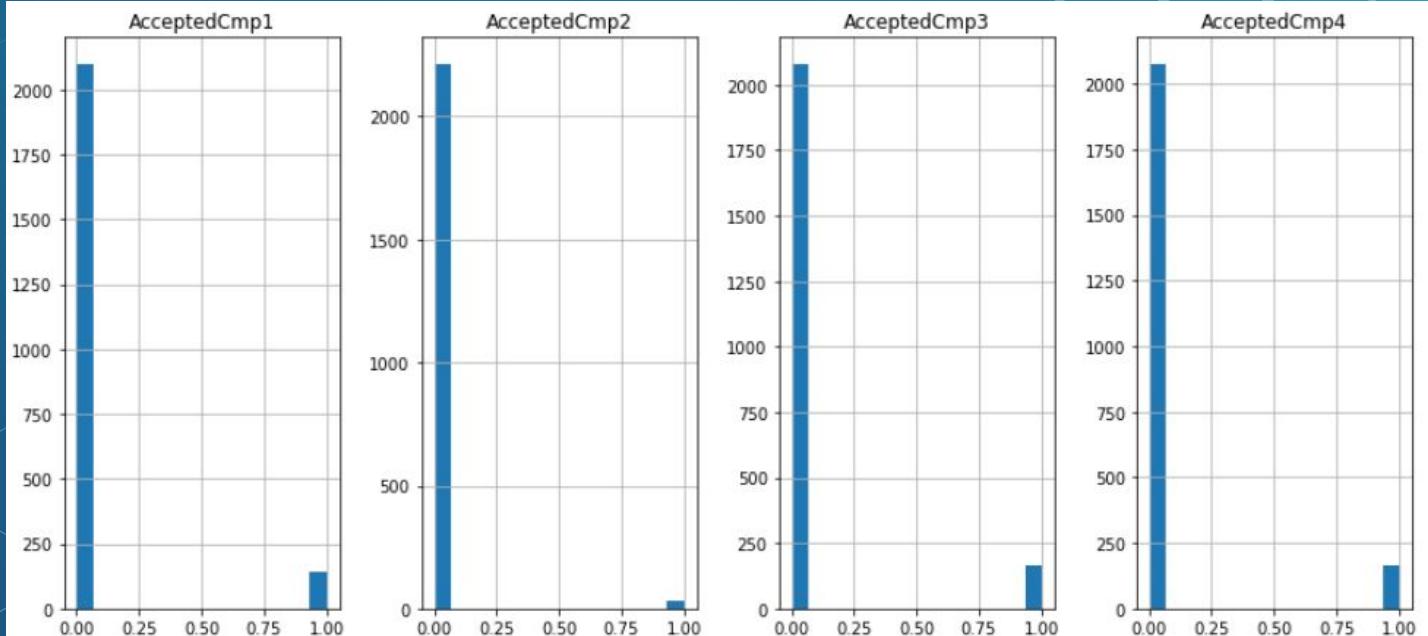




Insights das variáveis numéricas

- Maior parte dos dados trata de clientes em seus 40 e 50 anos;
- Há um número maior de compras feitas diretamente em lojas do que pelo catálogo ou site;
- Não é grande o número de compras feitas quando se há um desconto. Assim sendo, pode-se inferir que um desconto não é o fator principal que levará a uma compra nas nossas futuras campanhas direcionadas;
- Apesar de grande o número de visitas no site, as vendas do mesmo estão próximas da metade do número de vendas realizadas diretamente nas lojas.

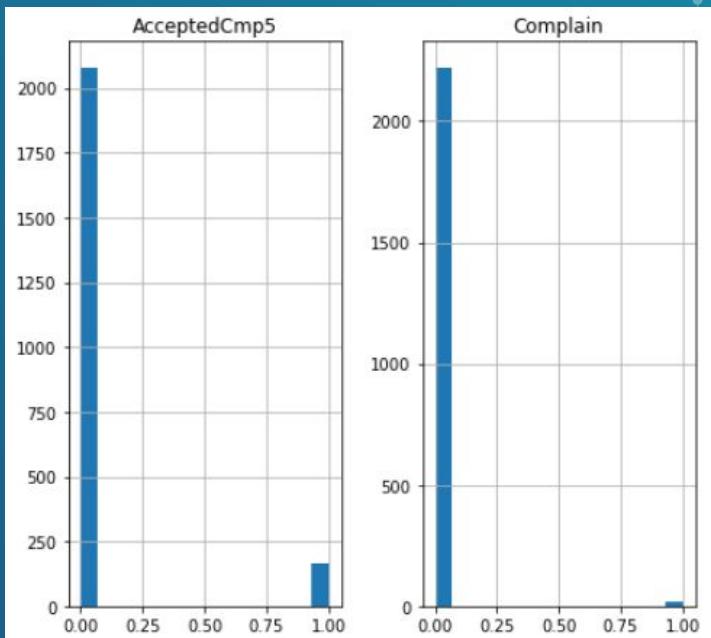
Distribuição das variáveis categóricas



Linha horizontal representa valores distintos e vertical, quantidade



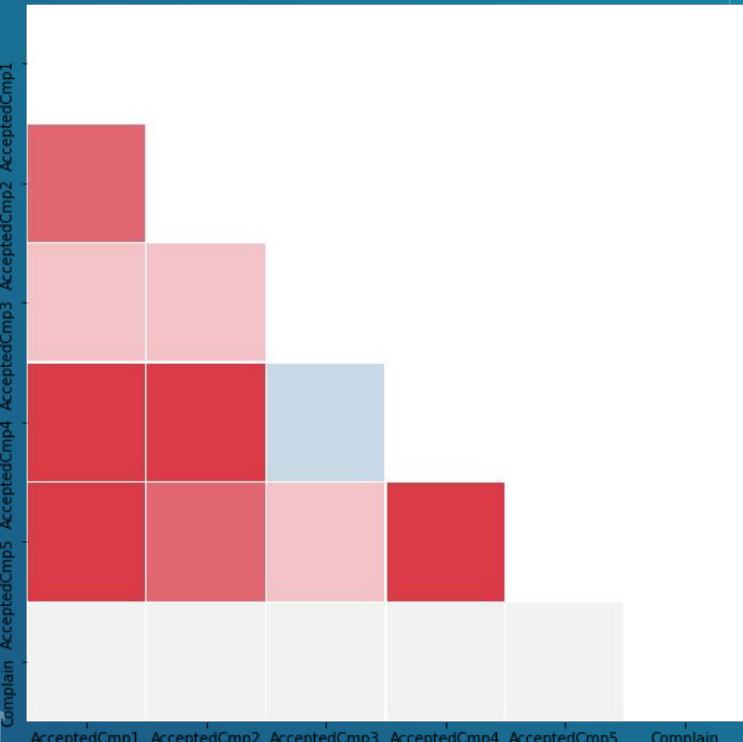
Distribuição das variáveis categóricas



Linha horizontal representa valores distintos e vertical, quantidade

Correlação das variáveis categóricas

Quão mais correlacionadas as variáveis são, maior o grau no índice da legenda. Educação e Estado Civil foram retiradas da análise por não apresentarem valores bitwise (0 ou 1)



Relação com o alvo

Relação das variáveis categóricas com o alvo. Limiar de ocorrências da variável foi de 50 (terceira coluna), e limiar de relação com o alvo foi de 15%, pois esse é o valor da média do alvo.

		Occurrences	Target relation	>= 50 occurrences?
Education	2n Cycle	203	0.108374	True
	Basic	54	0.037037	True
	Graduation	1127	0.134871	True
	Master	370	0.154054	True
	PhD	486	0.207819	True
	Absurd	2	0.500000	False
Marital_Status	Alone	3	0.333333	False
	Divorced	232	0.206897	True
	Married	864	0.113426	True
	Single	480	0.220833	True
	Together	580	0.103448	True
	Widow	77	0.246753	True
AcceptedCmp1	YOLO	2	0.500000	False
	0	2096	0.121660	True
AcceptedCmp2	1	144	0.548611	True
	0	2210	0.142081	True
AcceptedCmp3	1	30	0.666667	False
	0	2077	0.123736	True
AcceptedCmp4	1	163	0.472393	True
	0	2073	0.131211	True
AcceptedCmp5	1	167	0.371257	True
	0	2077	0.116514	True
Complain	1	163	0.564417	True
	0	2219	0.149166	True
	1	21	0.142857	False



Insights das variáveis categóricas

- Nossa limiar de ocorrências foi de 50, o que representa 2% do total possível da base (2% de 2240);
- Relação com a variável alvo foi de 15% pois esse valor foi o valor médio do alvo (Response);
- Não houveram muitas reclamações nos disparos da campanha, logo nossos consumidores não se sentiram incomodados com o serviço, levando a crer que possivelmente não se sentirão chateados num próximo disparo.

Os únicos tipos de cliente que não apresentaram boa relação com o alvo foram os que se encontravam nas categorias 'Casados' e 'Juntos' da variável Estado Civil; já as categorias 'PhD' e 'Mestre' estão acima do valor de relação com a variável alvo.



3

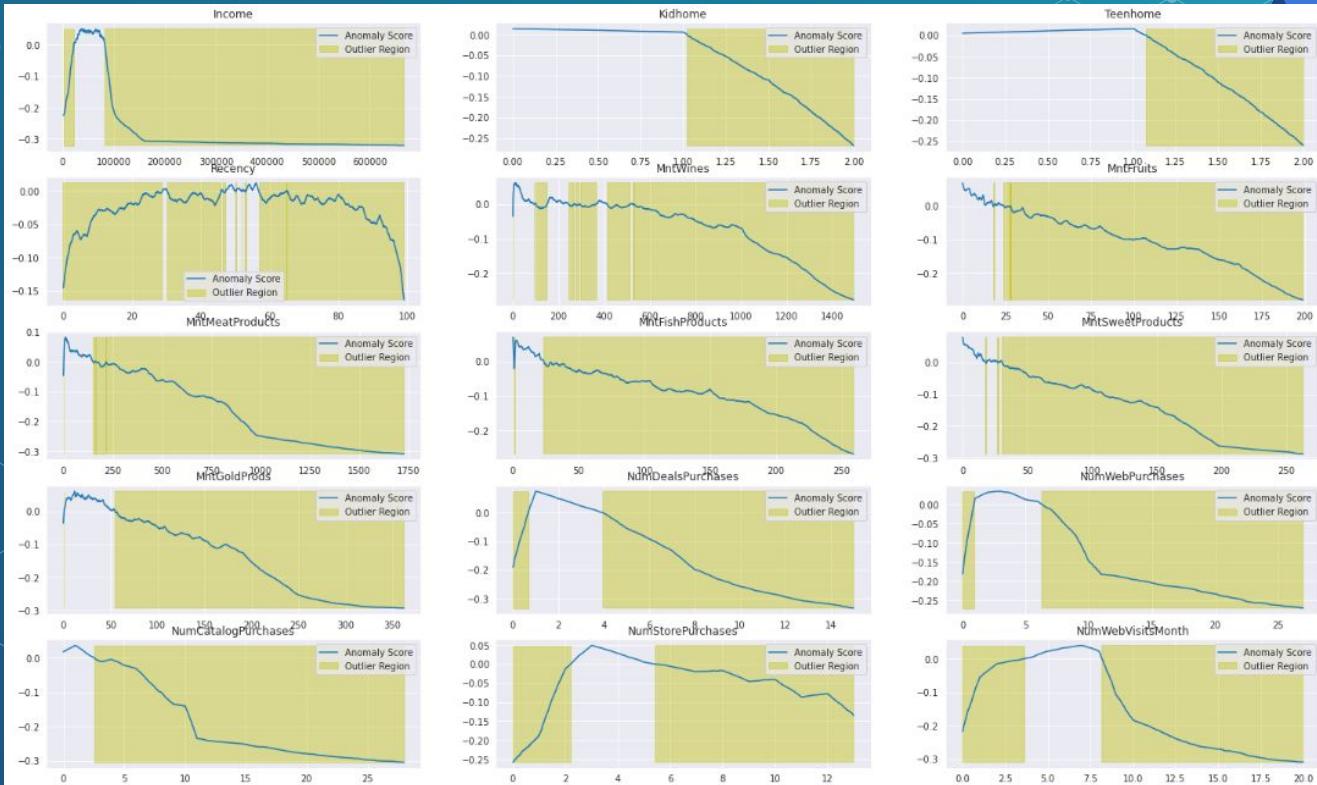
Modelo de Classificação

Criar um modelo preditivo usando nossos dados

O que fazer com os dados agora?

- Primeiro, vamos detectar as anomalias que os dados podem ter e para isso utilizamos o algoritmo Isolation Forest;
- Completamos dados nulos a partir de regressão linear;
- Vemos quais clientes apresentam valor de idade fora do comum (maior que 90 ou 100 anos);
- Analizamos outliers utilizando o método de Isolation Forest.

Gráficos de análise dos outliers



Análise das
variáveis
numéricas
procurando por
anomalias.

Dividindo bases para teste e treino

- 40% da base será utilizada para testes;
- Consequentemente, 60% irá para o treino do modelo;
- Iremos criar novas features para ajudar em nossa avaliação:
 - Percentual de produtos Gold vendidos;
 - Total e porcentagem de campanhas aceitas;
 - Porcentagem gasto em vinhos, frutas, carnes e peixes;
 - Valor gasto total;
 - Potencial de consumo (que é valor gasto dividido pela renda);
 - Frequência de compras;
 - RFM, que é Recência, Frequência e Valor Monetário;

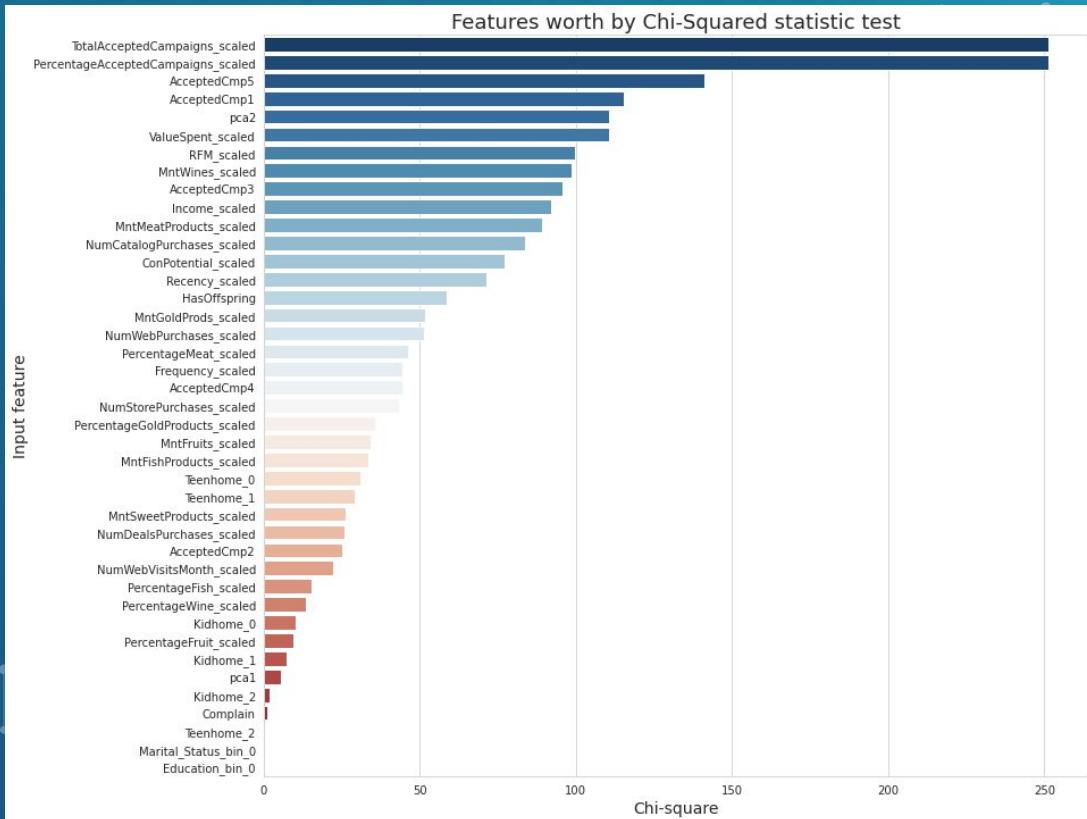


Dividindo bases para teste e treino

- A partir disso, normalizamos os dados usando MinMax Scaler;
- Utilizamos a transformação de Box-Cox para analisar a melhor transformação para cada feature;
- Com isso, criamos uma representação de quais variáveis tiveram maior valor na nossa análise, assim como seu uso para próximas campanhas.
- O método utilizado para verificar o valor das variáveis foi o Chi-Squared.

Valor das variáveis no modelo

Na vertical, das variáveis mais importantes em cima para as menos na região de baixo.





Criando um modelo de previsão

- Nossa método escolhido será o Multilayer Perceptron (MLP);
- Ele foi utilizado por causa das percepções entre diferentes camadas e níveis com maior assertividade dos resultados;
- Assim, criamos uma matriz de confusão com o treino para explanar a veracidade do modelo e analisar o negócio.

Criando um modelo de previsão

Representação do código e do resultado da MLP



Resultados da Matriz de Confusão

		Positive	Negative
Positive	772	0	
	0	124	

O valor de precisão do resultado foi de 1, o que leva a crer que ou o modelo teve um excelente resultado ou o treino teve overfit (treino excessivo) de variáveis.

4

Apresentação de Negócios

Marketing and Business Case

Considerações e resultados

- A base de dados tinha valores consistentes, sem dados confusos;
- Para que os melhores resultados possíveis fossem obtidos, novas variáveis foram criadas. Logo, conseguimos ver que aqueles que responderam a campanhas anteriores e os que mais gastaram nelas são os que têm mais chance de continuar com respostas satisfatórias;
- A eliminação de clientes foi mínima a fim de preservar os dados originais ao máximo. Uma próxima avaliação pode ser feita removendo-se os clientes que não responderam a nenhuma campanha;
Criar e analisar RFM (Recência, Frequência e Valor Monetário) e o número de campanhas aceitas foi um método de engenharia de variáveis para agregar valor ao modelo.

Considerações e resultados

- A proporção de dinheiro gasto em cada produto pode nos ajudar direcionando e apontando as campanhas para apenas aos que reagiram às últimas campanhas e direcionando os produtos já comprados;
- Quando observando frequência, não foi possível verificar se nossos clientes comprariam mais de acordo com de acordo com essa variável;
- Ter crianças e/ou adolescentes em casa não teve impacto significante em responder satisfatoriamente às campanhas passadas, ou seja, não é um fator que delimita quem se deve procurar em próximas campanhas;

Considerações e resultados

- Reclamações não estiveram relacionadas com o fato de não responder às campanhas passadas. Na verdade, não alterou significativamente o valor das variáveis;
- Vinhos e carnes foram os produtos com maior valor em nosso modelo, então as próximas campanhas podem oferecer descontos neles e/ou dar atenção a eles;
- Ainda que descontos tenham sido dados, esse fator não foi majoritário no aceite de campanhas já que as pessoas tendiam a responder às campanhas passadas independente de descontos;

OBRIGADO!

Dúvidas?

Contato:

- ❖ daidson.alves@gmail.com
- ❖ <https://bit.ly/3hXaHjr>

