SCC-5948

Aplicando técnicas de análise de dados

A Symphony of Customer Interactions

Profª. Dra. Roseli Aparecida Romero

Julyana Flores de Prá

Thiago Rafael Mariotti Claudio

Introdução

A base de dados ***A Symphony of Customer Interactions*** apresenta um conjunto de interações e aquisições em um *e-commerce* através de diferentes atributos, entre eles estão dados do comprador, ocupação, recepção à campanhas de *marketing* e operações de compra. Dessa forma tornou-se possível um estudo sobre o comportamento humano e suas interações financeiras e como estas podem ser influenciadas.

Ao longo da análise espera-se responder à algumas hipóteses.

Tópicos abordados

01 **Definição de hipóteses**

Perguntas que pretende-se responder

02 **Exploração e transformação**

Conhecendo o *dataset*

03 **Experimentos**

Descobrindo o potencial do *dataset*

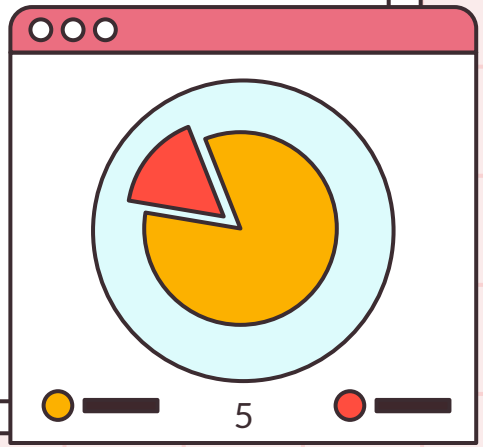
04 **Conclusão**

As hipóteses se confirmaram?

01

Definição de hipóteses

Perguntas que pretende-se responder

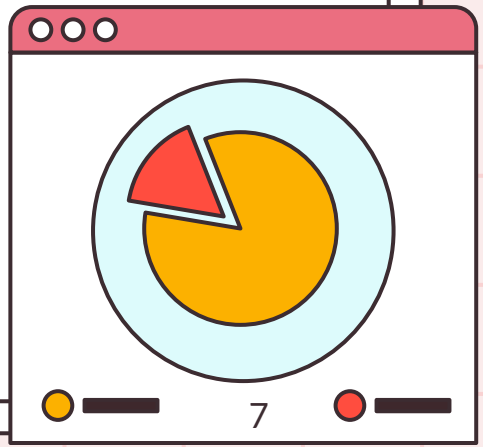


1. Existe correlação entre idade e volume de compras?
2. Existe uma relação entre a idade do cliente e a interação com as campanhas de compra?
3. Existe uma relação entre a taxa de conversão (volume de compra) e as interações no *website*?
4. A faixa socioeconômica do cliente afeta suas interações de escambo?
5. A reação à campanha de *marketing* é um bom indicativo das interações do cliente? É possível prever sua reação baseada nas interações?

02

Exploração e transformação dos dados

Conhecendo o *dataset*



Características gerais

O *dataset* apresenta:

- 32 colunas;
- 10.000 registros;
- Não há incidência de valores desconhecidos;
- Cada registro representa uma compra única dado que não há clientes repetidos de acordo com o "*Customer_ID*"

Transformações necessárias

Listas ‘Falsas’

Features que contém valores em formato de lista, mas na verdade eram uma *string* ('['not','a','list']' ≠ ['yes','a','list'])

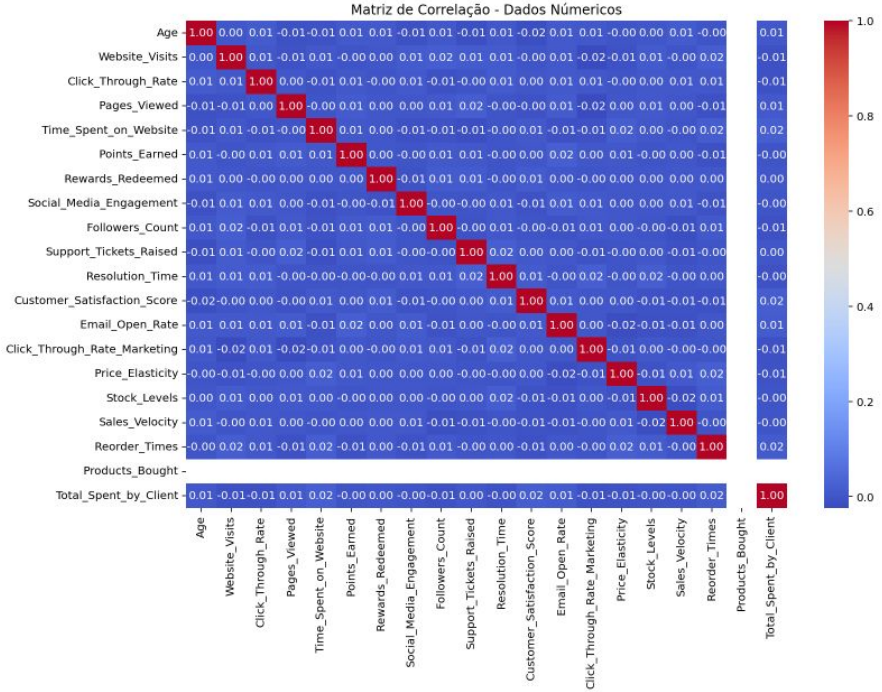
Agrupamento de informações

Uma coluna com valor total gasto seria de mais valia que um *array* de valores, então criou-se uma nova *feature*

Formatação de data

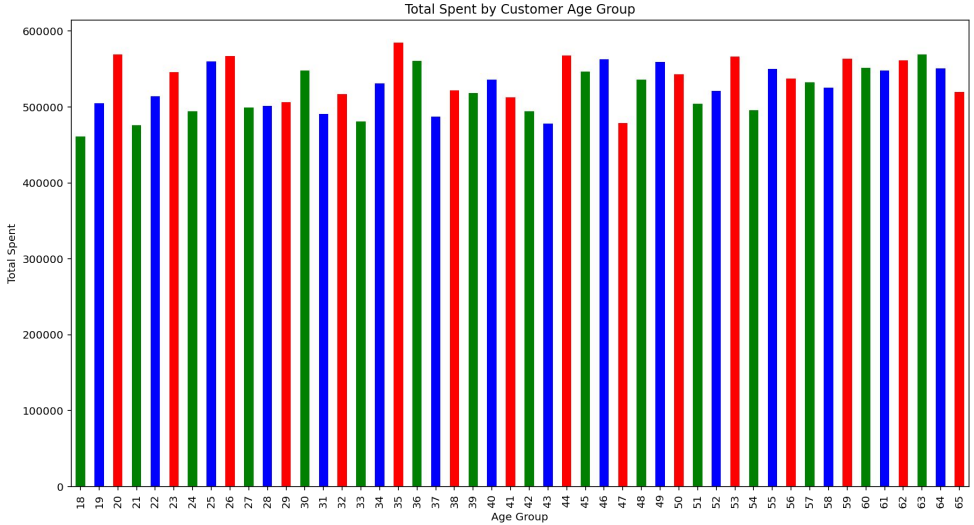
Os valores da *feature* “*Puchase_Date*” foram formatados para facilitar o processo de análise

Correlação dos dados



Exploração - Gasto por idade

Somatório dos gastos em produtos por cliente agrupados em sua respectiva faixa etária.



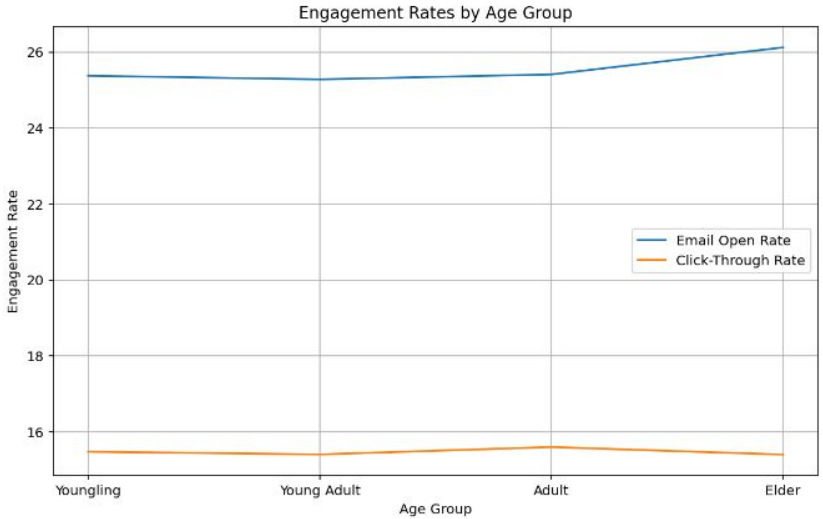
Exploração - Campanha x Idade

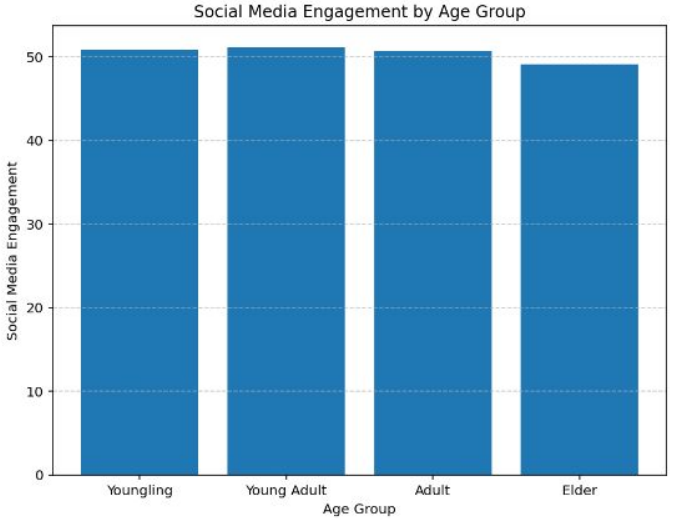
Agrupamento etário em: **Youngling (18 à 24)**, **Young Adult (25 à)**, **Adult** e **Elderly**.

Response_to_Campaign	Negative	Neutral	Positive
Age_Group			
Youngling	31.851401	33.617540	34.531060
Young Adult	31.662125	32.152589	36.185286
Adult	33.314958	34.307240	32.377802
Elder	34.505088	33.765032	31.729880

Distribuição uniforme de resposta à campanha para todas os agrupamentos etários.

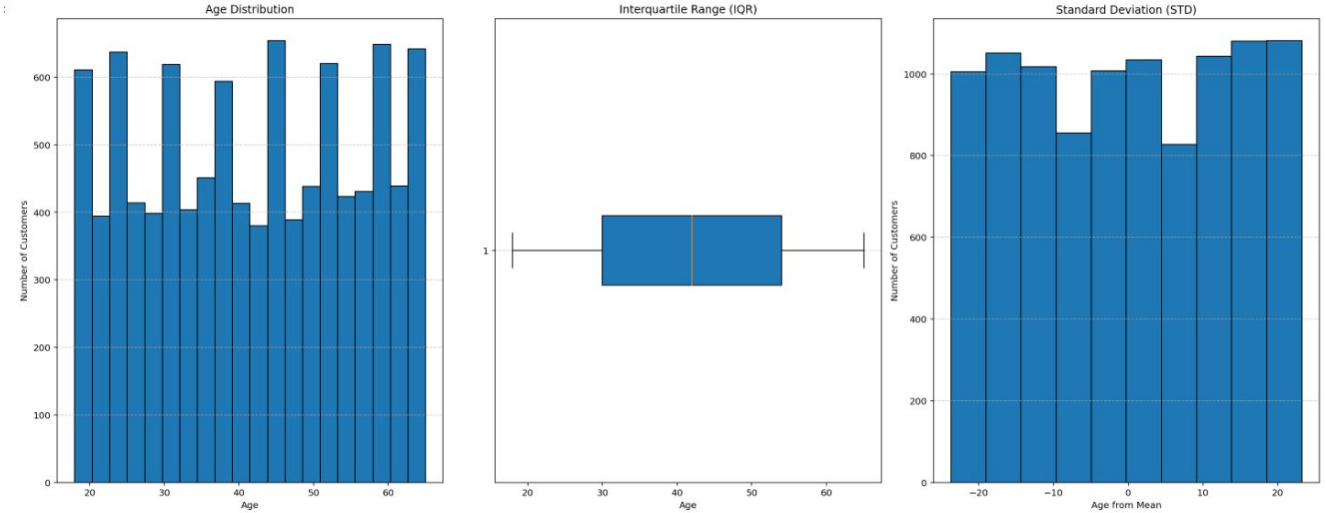
Exploração - Interação x Idade





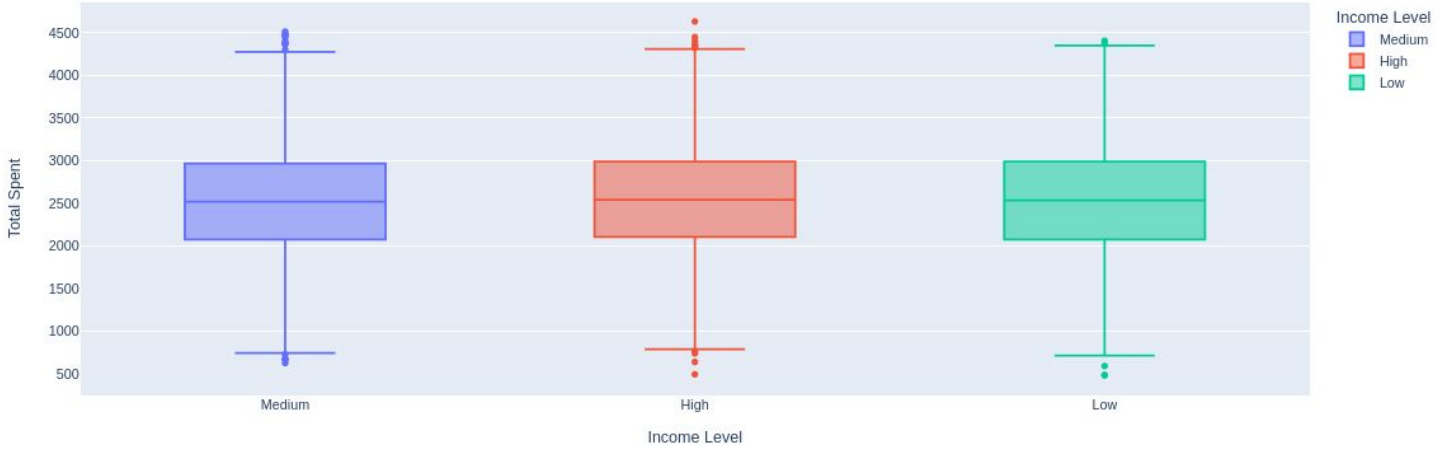
Também não observou-se nenhuma diferença entre as faixas etárias e suas interações com as redes sociais, e-mails e derivados.

Exploração - Clientes por Idade



Nota-se uma distribuição normal para as idades dos clientes. A razão para maior volume na faixa etária **Adulto** vêm do fato de que esse grupo engloba um **número maior de clientes**, e por isso dá impressão de desbalanceamento. Dito isso, não parece haver correlação entre idade e volume de compras. **Todos as faixas etárias tiveram gastos semelhantes.**

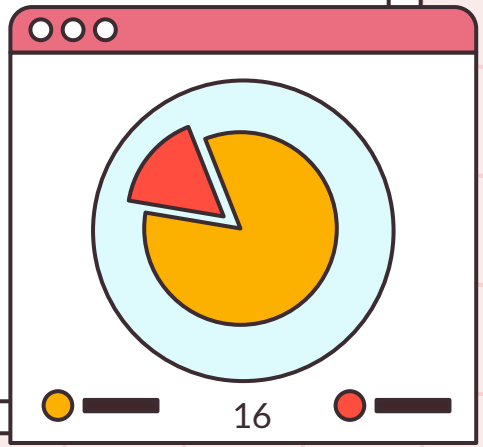
Exploração - Renda x Gasto



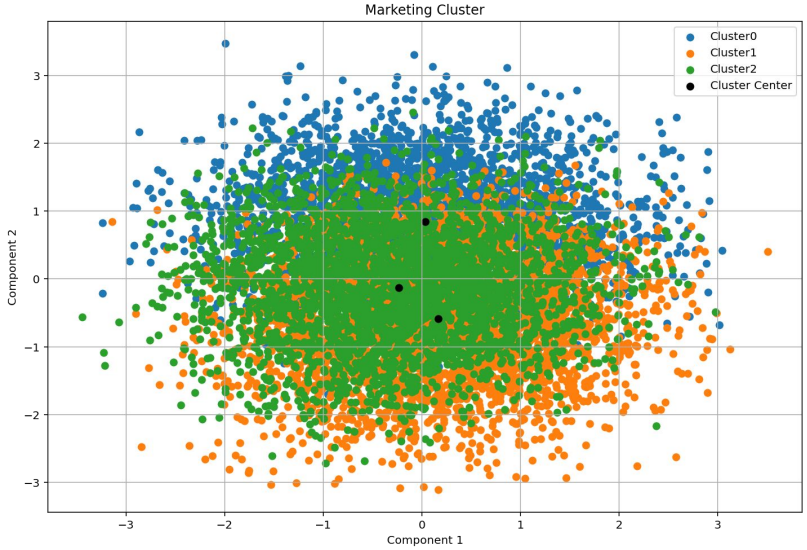
03

Experimentos

Descobrindo o potencial do *dataset*



Clusterização



Loyalty_Status	Gold	Platinum	Silver
Cluster			
0	1018	1011	1081
1	1257	1246	1276
2	1046	1028	1037

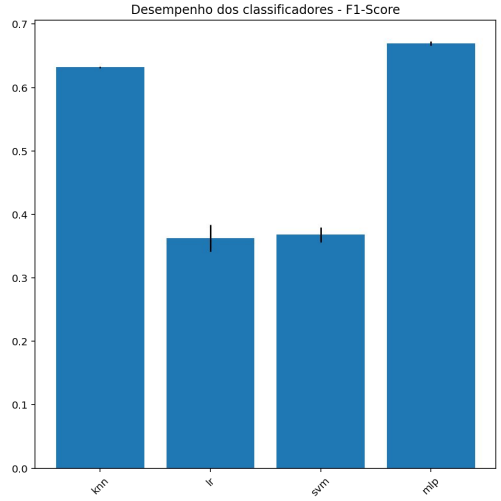
Response_to_Campaign	Negative	Neutral	Positive
Cluster			
0	1067	1005	1038
1	1242	1319	1218
2	981	1050	1080

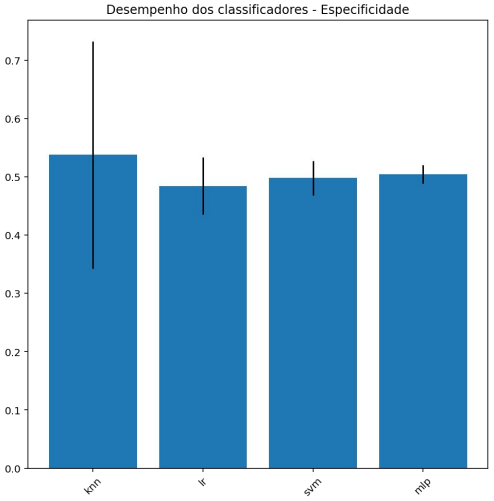
Algoritmos e Parâmetros

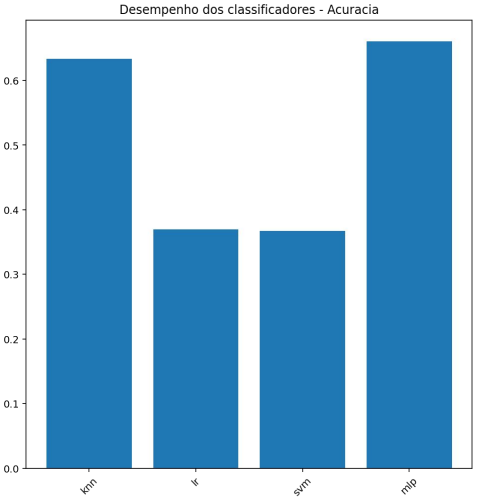
```
knn = KNeighborsClassifier(n_neighbors=3)
lr = LogisticRegression(solver='lbfgs', max_iter=1000, n_jobs=-1)
svm = SVC(kernel='linear')
mlp = MLPClassifier(solver='lbfgs', alpha=1e-5, hidden_layer_sizes=(100,), random_state=14)
```

Seed random state constate: 14

10-fold


Desempenho dos classificadores - F1-Score
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
knn
lr
svm
mlp

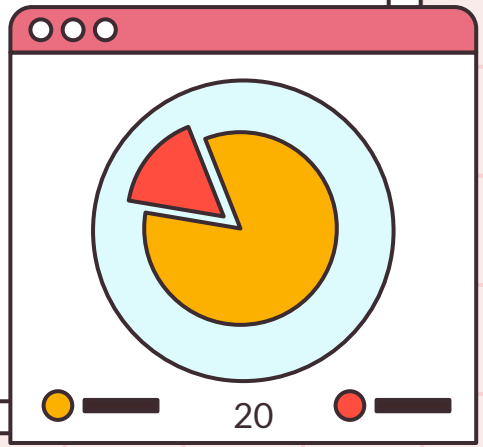

Desempenho dos classificadores - Especificidade
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
knn
lr
svm
mlp


Desempenho dos classificadores - Acuracia
0.6
0.5
0.4
0.3
0.2
0.1
0.0
knn
lr
svm
mlp

04

Conclusão

As hipóteses se confirmaram?



Do conjunto

- Conjunto de dados artificial
- Baixa Correlação e Features multi-label
- Baixo desempenho dos classificadores

Das hipoteses

1. A idade não influencia no volume de compras
2. Não existe uma faixa etária com maior tendência à interação no social do *ecommerce*
3. Clientes com maior interação no social do *ecommerce* não necessariamente possuem maior taxa de conversão ou pertecem ao grupo ‘Platinum’
4. A relação idade x renda não afeta a tendência de gastos
5. Não há relação entre resposta à campanha e interações de compras/social no geral

ICMC - USP

Obrigado!

Julyana Flores de Prá
Thiago Rafael Mariotti Claudio
Abril, 2024