Desafio Técnico - Hotmart

Caio Gabriel Barreto Balieiro

5 de agosto de 2022

Conteúdo

Introdução

Conjunto de dados

- Neste desafio técnico foi explorado o conjunto de dados referente a uma amostra de vendas da Hotmart de janeiro a junho de 2016.
- No que se refere aos dados, tem-se mais de 1,5 milhões de registros de compras de produtos da Hotmart.
- Foram removidos 38192 valores discrepantes utilizando a regra da padronização z score, valores maiores 3 e menores que -3 são considerados valores discrepantes.
- No decorrer das análises foi apresentado uma ferramenta utilizando auxílio do Streamlit para responder as perguntas propostas neste desafio técnico.
- Os códigos em python utilizados neste desafio são os seguintes: desafio.ipynb e ETL.ipynb e no deploy utilizando o streamlit.
- O conjunto de dados apresentado, tem as seguintes features:

Conjunto de dados

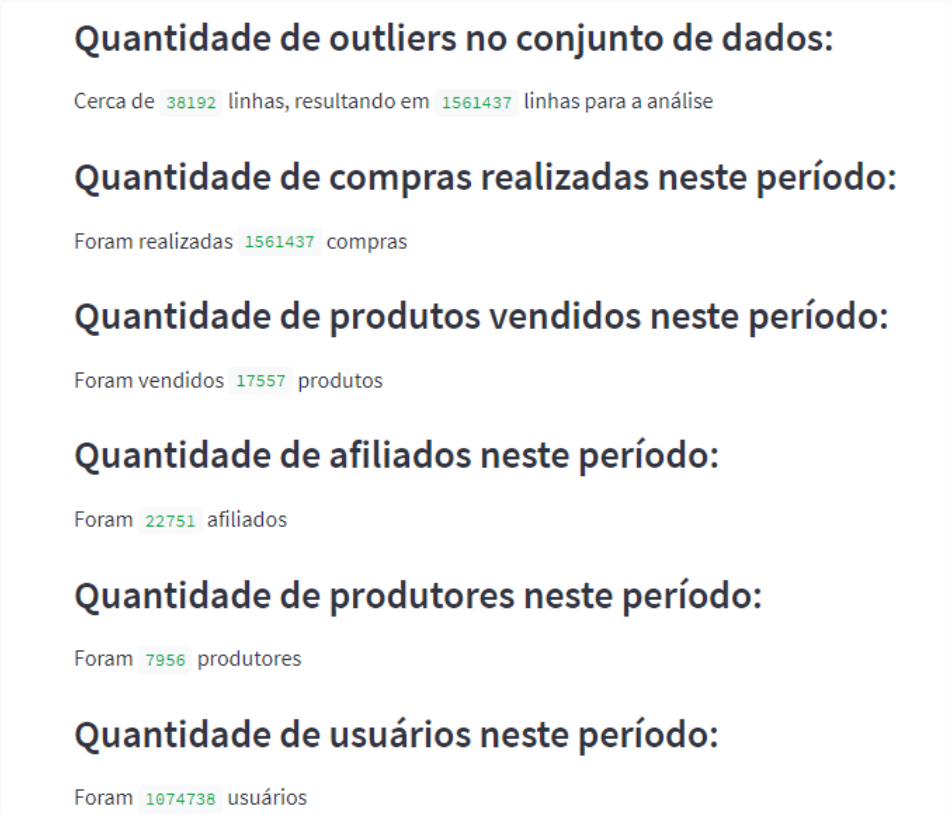
- Descrição das features:

  - **purchase_id**: Identificação da compra na Hotmart;
  - **product_id**: Identificação do produto na Hotmart;
  - **affiliate_id**: Identificação do afiliado na Hotmart;
  - **producer_id**: Identificação do produtor na Hotmart;
  - **buyer_id**: Identificação do comprador na Hotmart;
  - **purchase_date**: Data e hora em que a compra foi realizada;
  - **product_creation_date**: Data e hora em que o produto foi criado na Hotmart;
  - **product_category**: categoria do produto na Hotmart. Exemplo: e-book, software, curso online, e-tickets, etc.;
  - **product_niche**: nicho de mercado que o produto faz parte. Exemplo: educação, saúde e bem-estar, sexualidade, etc.;
  - **purchase_value**: valor da compra. Esse dado, assim como nicho e categoria foi codificado para manter a confidencialidade. O valor apresentado no dataset é o z-score do valor real;
  - **affiliate_commission_percentual**: percentual de comissão que o afiliado receberá da compra;
  - **purchase_device**: tipo de dispositivo utilizado no momento da compra, como: Desktop, Mobile, Tablet, ou Outros;
  - **purchase_origin**: endereço do site do qual a pessoa veio antes da compra. Por exemplo, se uma pessoa veio do Facebook, Youtube, ou até mesmo de outra página no site oficial do produto;
  - **is_origin_page_social_network:** informa se essa compra veio de uma URL do Facebook, Youtube, Instagram, Pinterest, ou Twitter

Mineração e Análise exploratória dos dados

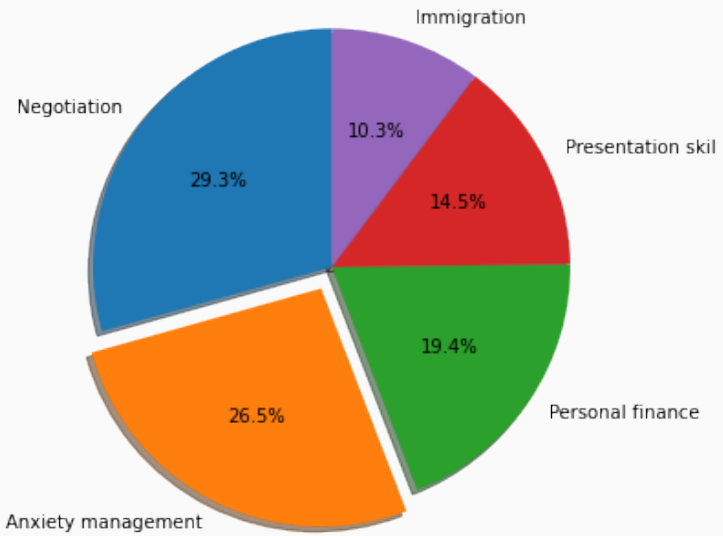
Mineração dos dados

- O conjunto de dados foi coletado utilizando o banco de dados MySQL utilizando bibliotecas em python, detalhes desse código é encontrado no arquivo ETL.ipynb.
- O conjunto de dados apresenta as seguintes características:



Análise Descritiva

- Em relação a todas as compras realizadas na Hotmart de janeiro a junho de 2016, temos que o top 5 das product niche é:



- Note que, Negotiation e Anxiety management contemplam mais de 50% dos classificações de product niche, e a menor é Immigration com aproximadamente 10%.

Análise Descritiva

- Na tabela abaixo tem-se que o top 5 da product category apresenta phisical book e podcast com aproximadamente 95%.

	top 5 product_category	product_category	Porcentagem
0	Phisical book	1298492	83.385960
1	Podcast	216222	13.885245
2	Workshop	36568	2.348307
3	eBook	3955	0.253980
4	Subscription	1970	0.126509

- Enquanto que Subscription representa apenas 0,12% das compras.

Análise Descritiva

- Nota-se abaixo que os valores do z score do valor da compra do produto varia de -0.54 até no máximo 3, por conta do ponto de corte para remoção dos valores discrepantes.

mean	-0.11
std	0.58
min	-0.54
25%	-0.46
50%	-0.36
75%	0.02
max	3.00

- Enquanto que a média é -0.11, isto indica que grande parte dos valores de compra são menores que o valor da média do valor de compra.

Perguntas Hotmart

Pergunta 1 - A empresa depende dos maiores produtores da plataforma? Ou seja, os produtores que mais vendem são responsáveis pela maior parte do faturamento da empresa?

- No que se refere a essa pergunta, foi decido pelo autor trabalhar com a uma transformação da escala original da valor de compra (z score).
- Isto é, foi quebrado a variável purchase value em um score variando de 1 a 5, isto é feito quando trabalha-se com modelos de segmentação de RFM.
- Utilizando essa técnica pode-se obter um valor melhor que o faturamento baseado na feature purchase value (que permite valores negativos).
- Outra técnica foi criar uma nova feature definindo como valores abaixo da média se purchase value for menor ou igual a 0 e maior que a média se valores fossem positivo.

Pergunta 1 - A empresa depende dos maiores produtores da plataforma? Ou seja, os produtores que mais vendem são responsáveis pela maior parte do faturamento da empresa?

- NOTE que, caso tivessem os resultados das compras sem o z score o faturamento poderia ser estimado com precisão, no entanto neste caso foi considerado o score gerado pela divisão do quintil da feature purchase value.
- A quantidade de produtores nessa amostra foi de 7936, então para responder essa pergunta, vamos fixar o top 500 dos produtores e verificar qual o percentual de uso em relação a produção em geral.

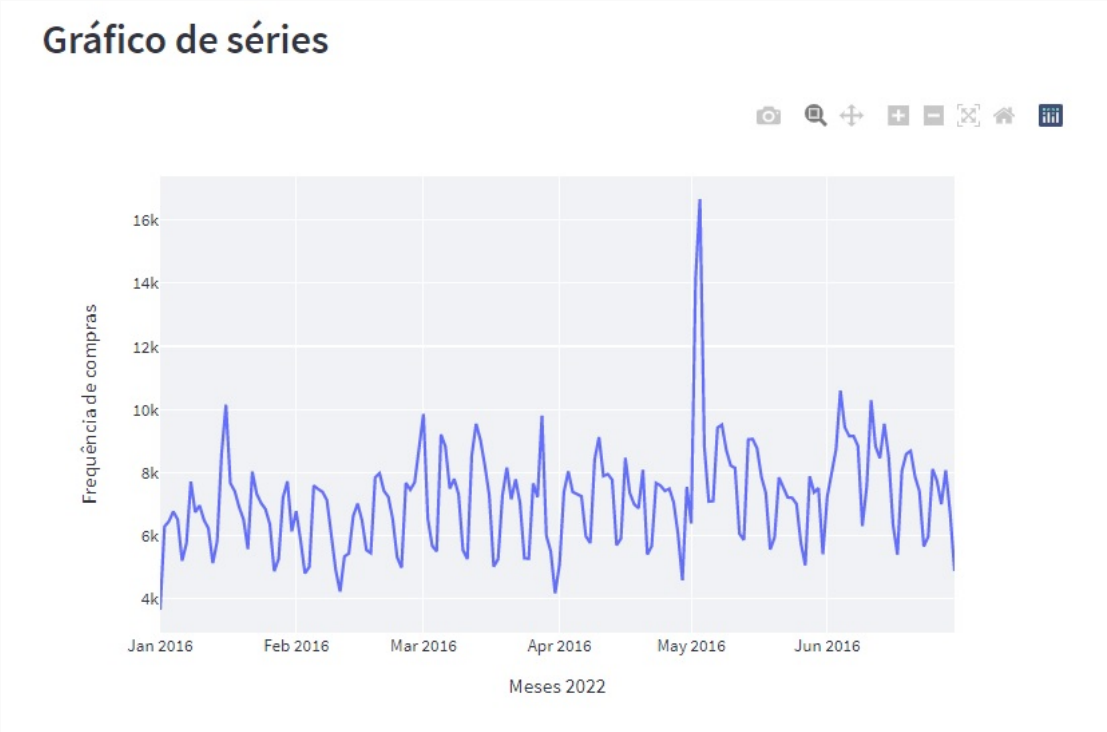
- Observa-se que os 500 melhores produtores apresentam mais de 79% do faturamento da Hotmart, mostrando que os mesmos impactam positivamente a plataforma.

Pergunta 2 - Existe algum padrão ou tendência relevante nos dados?

- Para responder essa pergunta, foi utilizado a feature product category, e foi considerado os resultados de quantidade de vendas por dia de janeiro até junho de 2016.
- Através desses dados, foi construido um gráfico de séries temporais, de autocorrelação e aplicação do teste de hipótese de Dickey-Fuller Aumentado para avaliar se o número de compras por dia relacionado a cada product category tem algum padrão de sazonalidade e de tendência.
- Considerando Phisical book, temos que o gráfico de compras apresenta comportamento de onda, assim como o de autocorrelação indicando que há uma componente de sazonalidade, ou seja, os picos de vendas ocorrem em dias repetidos em cada semana. Além disso o valor-p também indica que a série não é estacionaria (Tem possível comportamento de tendência e sazonalidade).

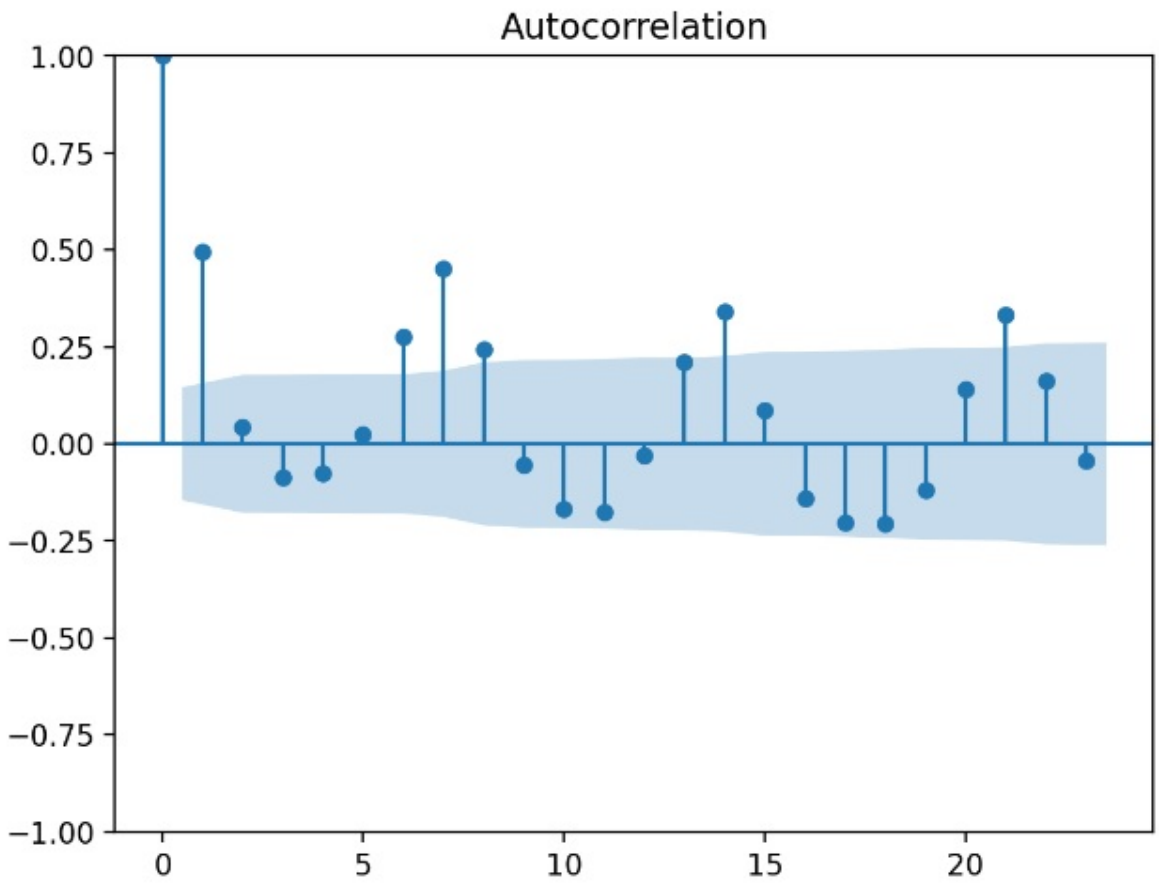
Pergunta 2 - Existe algum padrão ou tendência relevante nos dados?

- Se observamos temos que o gráfico de série temporal tende a crescer os valores quando comparados do inicio ao fim, isto indica uma tendência de crescimento de vendas de janeiro a junho de 2016. Pode-se fazer isto com as demais variáveis utilizando o app desenvolvido em Streamlit.



Pergunta 2 - Existe algum padrão ou tendência relevante nos dados?

Gráfico de autocorrelação



Pergunta 3 - É possível segmentar os usuários com base em suas características(faturamento, nicho de produto, etc)?

- A resposta é sim! Basta utilizar o tratamento adequado aos id dos usuários e utilizando técnicas, tais como (One-Hot encoding) e utilizando um algoritmo de segmentação.
- Para este desafio foi adotado o modelo Kmeans. No entanto, vale ressaltar que um modelo de segmentação de RFM também poderia ser aplicado, utilizando as features buyer id, purchase date, purchase value (sem z-score) e construindo uma nova feature relacionada a número de transações de cada buyer id.

Pergunta 3 - É possível segmentar os usuários com base em suas características(faturamento, nicho de produto, etc)?

- Utilizando o algoritmo de Kmeans e considerando o gráfico de Elbow (gráfico de cotovelo) foi encontrado 7 cluster para essa amostra, cujo os tamanhos é apresentado abaixo.

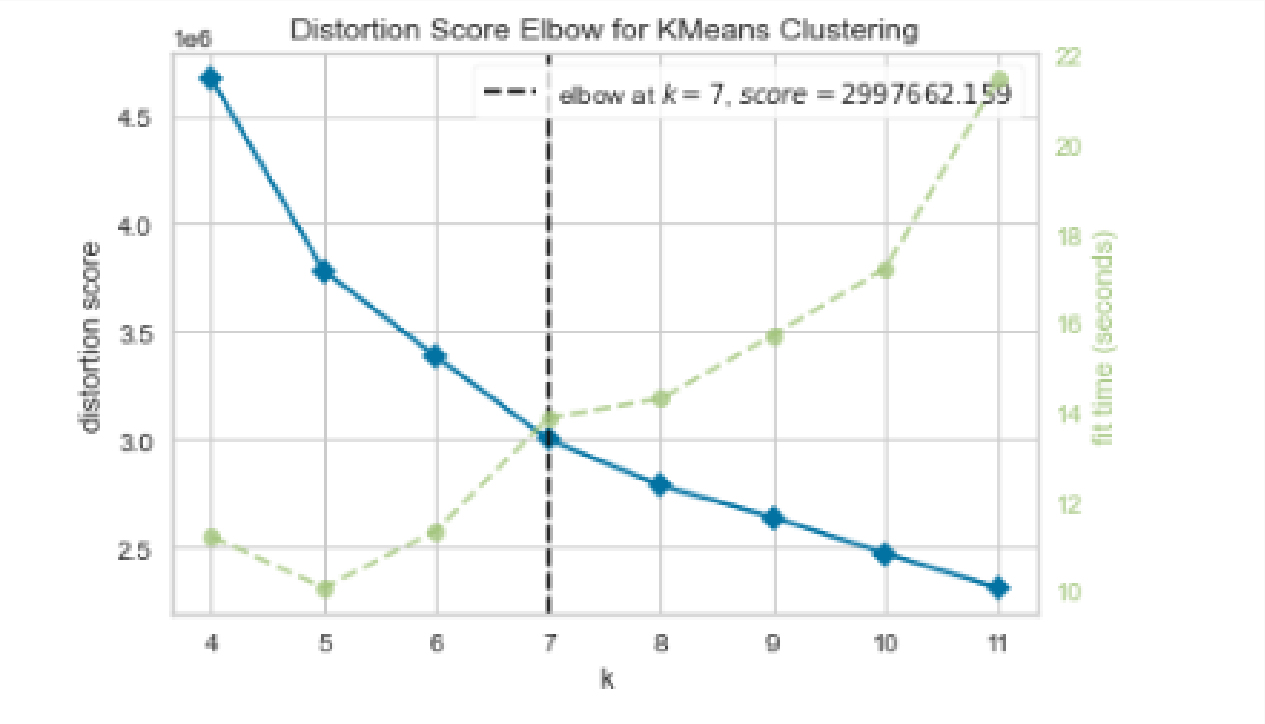
Descritiva dos cluster encontrados pelo Kmeans:

Quantidade de usuários em cada cluster:

	cluster
0	343550
1	36841
2	317756
3	80521
4	656
5	6350
6	289064

Pergunta 3 - É possível segmentar os usuários com base em suas características(faturamento, nicho de produto, etc)?

- Gráfico de cotovelo para escolher o melhor valor de k para segmentar os clientes.



Pergunta 3 - É possível segmentar os usuários com base em suas características(faturamento, nicho de produto, etc)?

- Para melhorar ainda mais as análises foi considerado abaixo uma descritiva sobre os cluster em relação ao faturamento (utilizando o score) e observa-se por exemplo que os grupos com maiores faturamento são os cluster 0 e 2.

Faturamento (score de faturamento) da Hotmart de usuários em cada cluster:

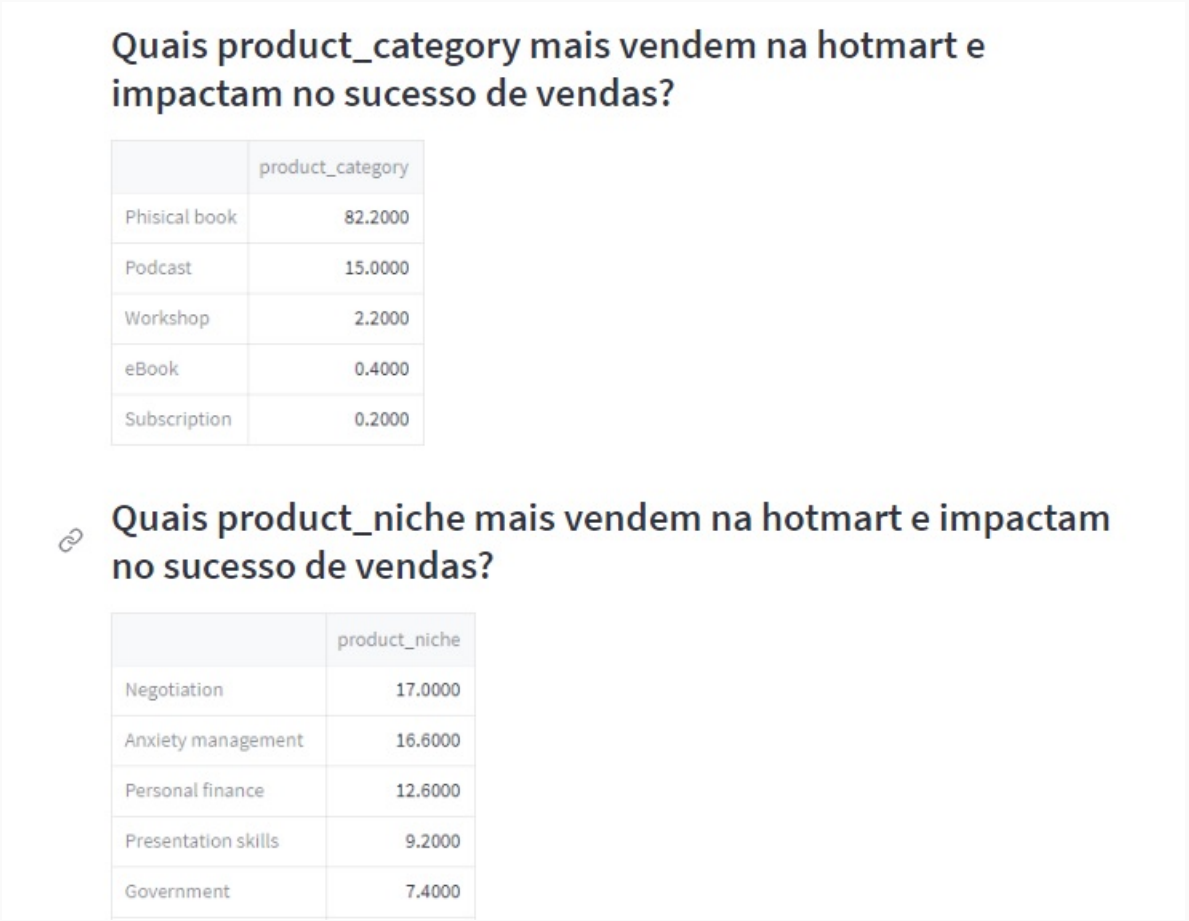
	score_faturamento
0	1665243
1	543853
2	1101175
3	694033
4	37689
5	173193
6	455470

Pergunta 4 - Quais características mais impactam no sucesso de um produto? Ou seja, o que faz um produto vender mais?

- Na Figura 1, tem-se que, considerando o top 500 de vendas da hotmart, tem-se que a product category que mais vendem são Phisical Book e Podcast contemplado mais de 97% deles.
- Enquanto que considerando product niche, tem-se que Negotiation, Anxiety managemen, Personal finance e Presentation skills contemplam cerca de 55,4% dos produtos que mais são vendidos.

Pergunta 4 - Quais características mais impactam no sucesso de um produto? Ou seja, o que faz um produto vender mais?

Figura 1: Resultados das características que mais impactam no sucesso do produto



Quais product_category mais vendem na hotmart e impactam no sucesso de vendas?

	product_category
Phisical book	82.2000
Podcast	15.0000
Workshop	2.2000
eBook	0.4000
Subscription	0.2000

Quais product_niche mais vendem na hotmart e impactam no sucesso de vendas?

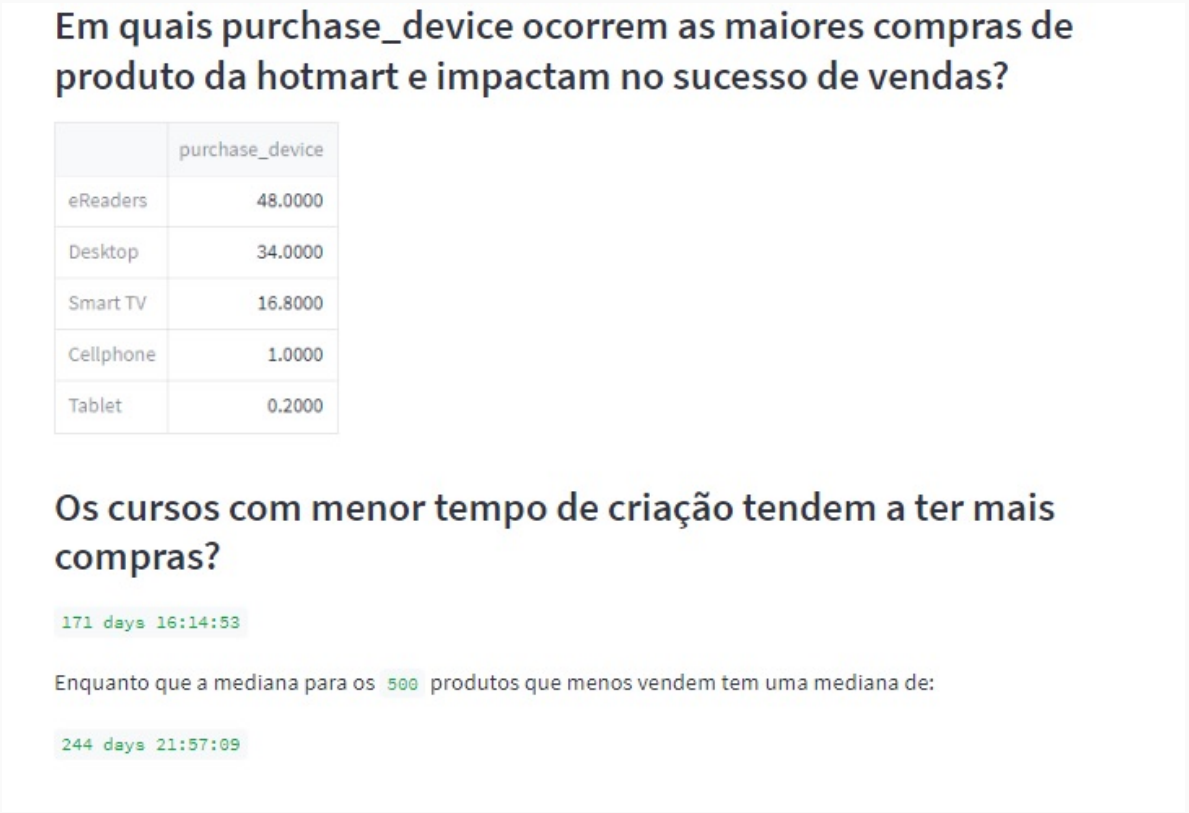
	product_niche
Negotiation	17.0000
Anxiety management	16.6000
Personal finance	12.6000
Presentation skills	9.2000
Government	7.4000

Pergunta 4 - Quais características mais impactam no sucesso de um produto? Ou seja, o que faz um produto vender mais?

- Na Figura 2, tem-se que os purchase device que ocorrem maiores compras de produtos da hotmart são eReaders, Desktop e SmartTV que contemplam mais de 98%.
- Além disso, quando comparados o tempo de compra com o tempo de criação do produto, tem-se que com os produtos que vende mais o tempo de criação e de venda é 171 dias em mediana, e os 500 piores vendem produtos com mais de 244 dias de vida.

Pergunta 4 - Quais características mais impactam no sucesso de um produto? Ou seja, o que faz um produto vender mais?

Figura 2: Resultados das características que mais impactam no sucesso do produto



Em quais purchase_device ocorrem as maiores compras de produto da hotmart e impactam no sucesso de vendas?

	purchase_device
eReaders	48.0000
Desktop	34.0000
Smart TV	16.8000
Cellphone	1.0000
Tablet	0.2000

Os cursos com menor tempo de criação tendem a ter mais compras?

`171 days 16:14:53`

Enquanto que a mediana para os `500` produtos que menos vendem tem uma mediana de:

`244 days 21:57:09`

Pergunta 5 - É possível estimar quanto de faturamento a Hotmart irá fazer nos próximos três meses a partir do último mês mostrado no dataset?

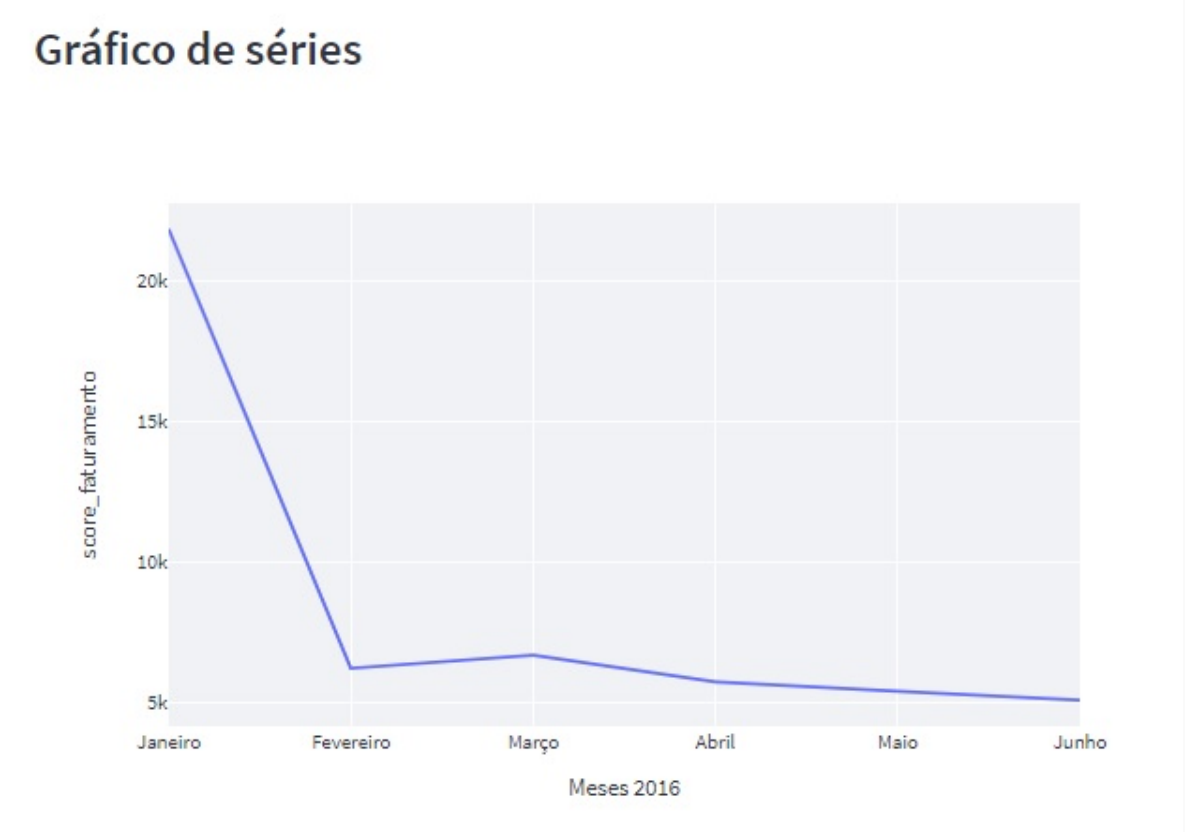
- A reposta para essa pergunta é sim (do ponto de vista de codar é possível), entretanto, é necessário avaliar se a quantidade de dados é adequada, para este desafio foi dada apenas uma amostra cujo a variação era de janeiro a junho de 2016 (6 valores do passado), são poucos meses observados para que o modelo possa aprender (como as observações eram apenas 6, nem deu para separar em treino e teste). Por essa razão o modelo apresentou previsões semelhantes para os demais três meses.
- Solução para isto: 1 - Aumentar a quantidade de anos para que a previsão do modelo faça sentido, e a Hotmart possa usufruir dos resultados do modelo. Isto é fácil de resolver pois a Hotmart tem dados suficiente para melhorar e aumentar a quantidade de meses observados. 2 - Utilizar o auxílio do Pyspark para trabalhar com esse aumento de dados

Pergunta 5 - É possível estimar quanto de faturamento a Hotmart irá fazer nos próximos três meses a partir do último mês mostrado no dataset?

- Na Figura 3 tem-se o gráfico de faturamento (score de faturamento) e na Figura 4 tem-se o resultado estimado utilizando o modelo Sarimax.
- Observa-se que os valores apontados pela previsão são iguais, isto mostra que o modelo não é adequado para o conjunto de dados.
- Uma explicação prática é que tem poucos dados observados para que o modelo posso aprender a ponto de fazer boas previsões. Com poucos meses observados inclusive, fica inviável dividir a base em treino e teste.
- Caso o número de meses fosse maior o modelo Sarimax seria um ótimo candidato a estimação tal como modelos de aprendizagem de máquina como o Random Forest que é bastante utilizado, assim como as redes neurais.

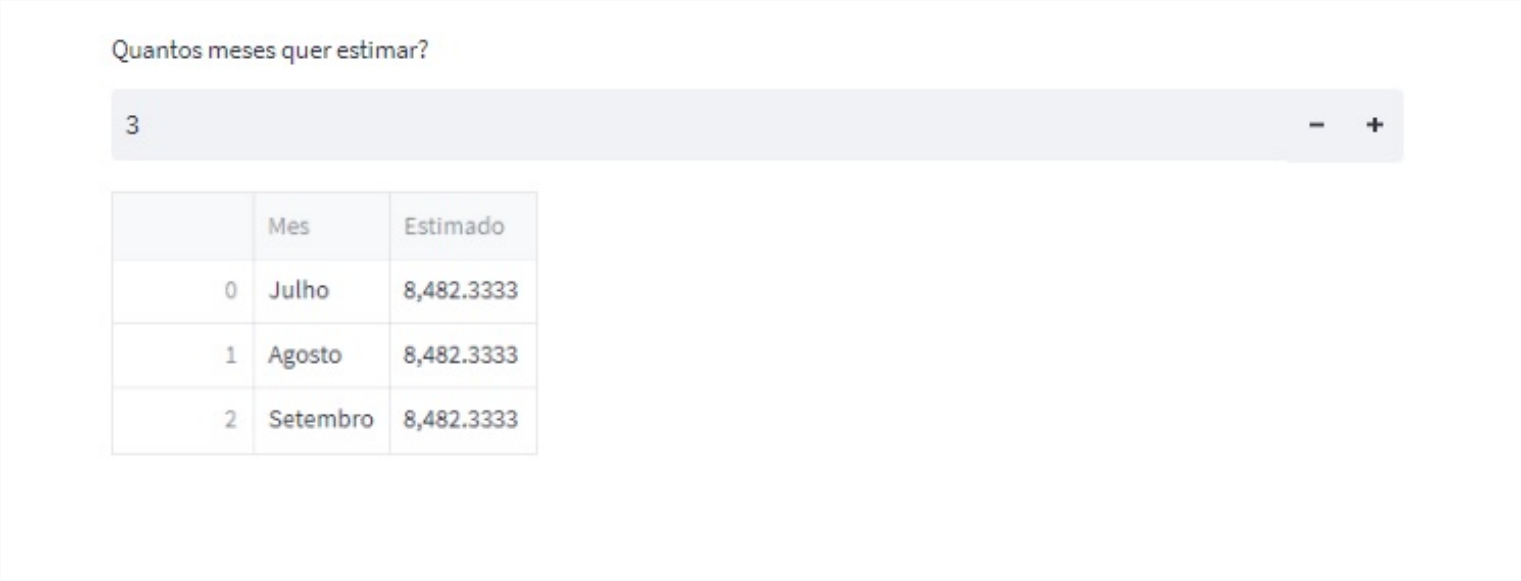
Pergunta 5 - É possível estimar quanto de faturamento a Hotmart irá fazer nos próximos três meses a partir do último mês mostrado no dataset?

Figura 3: Resultados das características que mais impactam no sucesso do produto



Pergunta 5 - É possível estimar quanto de faturamento a Hotmart irá fazer nos próximos três meses a partir do último mês mostrado no dataset?

Figura 4: Resultados das características que mais impactam no sucesso do produto



Modelos de Aprendizagem de Máquina

Modelos de Classificação

- Os modelos que foram utilizados para classificar se uma compra tem valor abaixo ou acima do valor médio de todos as compras foram: Modelo de regressão logística, Árvore de decisão e *Random forest*.
- As fetures considerada para o modelo foram: product category, product niche, purchase device, affiliate commission percentual e a variável target: class faturamento(acima da média, abaixo da média).
- Após a definição dos modelos a serem utilizados, foi separado a amostra de vendas da Hotmart em uma amostra treino (80%) e amostra teste (20%).
- Com esta partição dos dados os modelos citados acima foram treinados e testados e os resultados das métricas de Acurácia e F1-score são dadas na Tabela 1.

Modelos de Classificação

- É importante destacar que a variável target é desbalanceada, isto é o número de classificação abaixo da média é maior do que acima da média.
- Neste sentido, foi utilizado técnicas de undersampling, isto é, reduz a variável com maior quantidade de classificações para menor, escolhendo as linhas de forma aleatória.
- Logo, o conjunto de dados apresenta dados balanceados e pode ser utilizado para o processo de aprendizagem.

Modelos de Classificação

Tabela 1: Modelos de Aprendizagem de máquinas utilizados para predição classificação de faturamento na Hotmart.

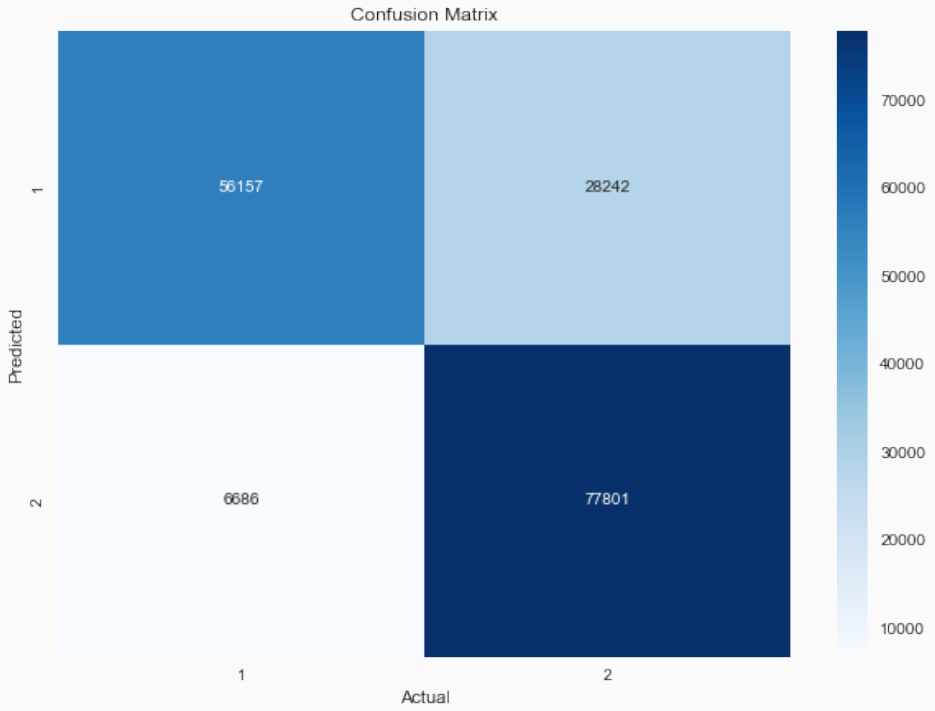
Modelos	Aucrácia	F1-Score
Regressão Logística	0,913	0,906
Decision Tree	0,893	0,888
Random Forest	**0,923**	**0,920**

Modelos de Classificação

- Com base na Tabela 1, tem-se que o modelo que obteve as maiores métricas de acurácia e f1-score, foi o modelo Random Forest.
- Portanto, o modelo Random Forest é o melhor modelo dentre os escolhidos.
- É importante destacar que o modelo ainda poderia passar por uma processo de tunning dos parâmetros para melhorar ainda mais a precisão em relação as classificações.
- Por fim, um gráfico de matriz de confusão foi apresentado na Figura 5, mostrando o potencial de acerto do modelo escolhido.
- O deploy deste modelo foi realizado via streamlit, cujo o mesmo considero o exemplo de 100 usuários e o modelo realizando a predição se a compra realizada está abaixo ou acima da média geral dos valores de compras.

Modelos de Classificação

Figura 5: Matriz de confusão para o modelo Random forest



Resultados e discussões

Resultados e discussões

- Neste desafio foram apresentadas 5 perguntas inicias e um conjunto de dados contendo mais de 1,5 milhões de linhas.
- Neste sentido, para responder essas perguntas de maneira sucinta e bem fácil, foi criado um app utilizando o streamlit para visualização dos dados, e das respostas apresentadas no desafio.
- As respostas das perguntas estão no decorrer da desta apresentação assim como no app Hotmart App. Além disso os códigos utilizando a linguaguem Python foram disponibilizados para a consulta sendo eles: desafio.ipynb e ETL.ipynb.
- Além das perguntas foi apresentado um modelo de classificação para detectar se uma determinada compra tem um valor acima ou abaixo da média geral, desta forma a Hotmart pode verificar se os produtos que estão sendo vendidos são maiores ou menores que o valor médio de compras.

Resultados e discussões

- Vale ressaltar que com a base com o faturamento original, pode-se obter de maneira direta as respostas para diversos itens neste desafio, bastante apenas fazer uma leve modificação no código, e o mesmo obtém as estatísticas com mais precisão.
- Um ponto interessante é que a amostra que foi fornecida apresentou apenas meses de janeiro a junho de 2016. No entanto, se estivesse disponível mais meses modelos de predição/séries temporais poderiam ter melhores performance para retirada de insights para a área de negócio da Hotmart.