BÁZICO

Relatório do Desafio Recompra Bazica

Candidato: Wesley da Fonseca Monte

Estágio em Ciência de Dados

1. Introdução

Este relatório apresenta a solução para o desafio proposto pela empresa **Bazico**, que consiste em desenvolver um modelo capaz de prever quais clientes têm maior probabilidade de fazer recompras no período de duas semanas. Para atingir esse objetivo, foi disponibilizado um conjunto de dados em formato CSV, que foi processado por meio de um código em Python.

A seguir, serão apresentadas as etapas realizadas para a criação do modelo de previsão, desde a leitura dos dados até a avaliação dos resultados obtidos. Cada etapa será detalhada para que se compreenda claramente a lógica e os métodos utilizados.

Mais adiante, é disponível um plano de ação para a empresa a partir das informações obtidas na análise, esse plano de ação serve como um possível direcionamento que a empresa deve tomar para chegar em um aumento de suas vendas e cumprir o objetivo proposto que é o retorno desses clientes.

Ao final do relatório também encontrasse uma discussão, expressando todas minhas dificuldades e aprendizados com o case.

Bibliotecas utilizadas:

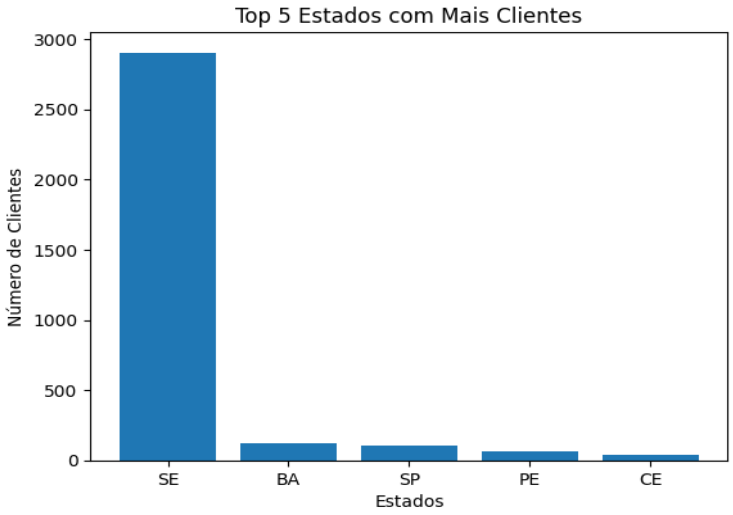
Utilizei algumas bibliotecas do Python para me ajudar na resolução do desafio as quais irei mostrar a seguir com uma breve explicação sobre cada uma delas:

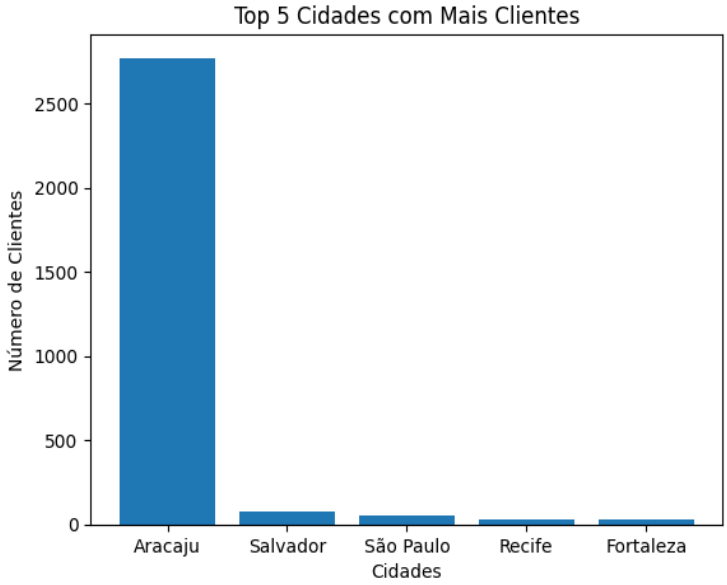
- pandas: é uma biblioteca de análise de dados que oferece estruturas de dados flexíveis e de alto desempenho para manipulação e análise de dados em Python.
- datetime: é uma biblioteca padrão do Python que fornece classes para trabalhar com datas e horários.
- matplotlib: é uma biblioteca para criação de visualizações estáticas, animadas e interativas em Python.
- numpy: é uma biblioteca para computação numérica com Python, que fornece suporte para matrizes e operações matemáticas.
- scipy.stats que realiza um teste binomial para a probabilidade de sucesso de uma população.
- Sklearn que é a biblioteca de aprendizado de máquina

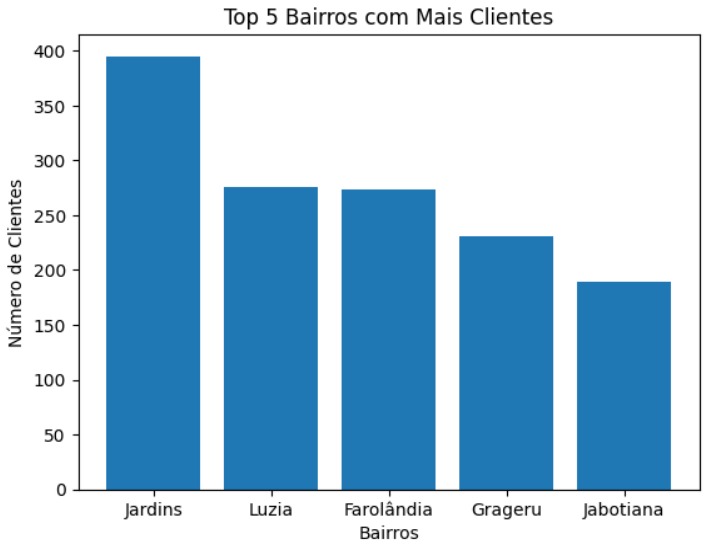
2. Metodologia

2.1 Etapa Análise

A primeira etapa do código foi para limpar os dados vazios e realizar uma análise geral nos dois Dataframes. Após a limpeza, foram verificadas informações importantes como a localização dos clientes, considerando cidade, estado e bairros. Com isso, conseguimos obter resultados interessantes, como os top 5 estados com maior número de clientes.


Top 5 Estados com Mais Clientes
Número de Clientes
3000
2500
2000
1500
1000
500
0
SE
BA
SP
PE
CE
Estados


Top 5 Cidades com Mais Clientes
Número de Clientes
2500
2000
1500
1000
500
0
Aracaju
Salvador
São Paulo
Recife
Fortaleza
Cidades


Top 5 Bairros com Mais Clientes
Número de Clientes
400
350
300
250
200
150
100
50
0
Jardins
Luzia
Farolândia
Grageru
Jabotiana
Bairros

A partir dos resultados obtidos na análise exploratória de dados, podemos ter insights valiosos sobre o perfil dos clientes da empresa **Bazico**. Sabendo onde estão concentrados os clientes, por exemplo, ao analisar os dados de cidades, estados e bairros, podemos identificar padrões e tendências que podem ser utilizados para desenvolver estratégias de marketing direcionadas para essas regiões. Dessa forma, é possível criar campanhas mais assertivas e personalizadas para cada público, aumentando as chances de conversão em vendas.

Ainda na etapa de análise foram considerados outros elementos relevantes para a criação do modelo de recompra. Dentre eles, destaca-se a análise de frequência dos clientes, em que se verificou quantos dias diferentes cada cliente retornou à empresa. Isso foi feito para entender quais clientes estão mais propensos a fazer compras novamente com um intervalo de tempo maior ou igual a 10 dias.

Esse indicador mostra quais clientes têm um maior engajamento com a empresa e podem ser mais propensos a se tornarem clientes fiéis. Além disso, a análise da frequência de compra também foi associada ao valor total gasto pelos clientes, possibilitando entender quanto cada cliente investiu na empresa e assim, direcionar campanhas de marketing e promoções específicas para os clientes que gastaram mais, incentivando-os a fazer novas compras.

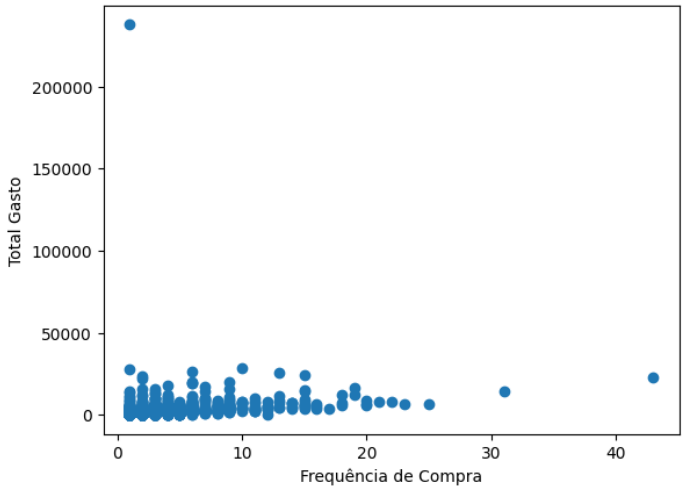
Todas essas análises prévias forneceram uma base sólida para a criação do modelo de regressão linear, que tem como objetivo prever quais clientes têm maior probabilidade de fazer recompras nos próximos 14 dias.

Imagem 1: Relação entre frequência de compra e total gasto por alguns clientes

	ID_Cliente	Freq_Compra	Total_Gasto
16	1.226840e+10	43	23123.30
21	1.238377e+10	12	5303.73
23	1.238509e+10	10	7122.57
26	1.242904e+10	22	7971.50
32	1.242915e+10	15	14777.40
...	...	...	...
1851	1.580584e+10	14	4292.00
1859	1.580671e+10	10	4264.00
1941	1.581689e+10	15	5471.20
1972	1.582084e+10	11	2727.00
2192	1.585883e+10	12	4122.00

Com base nos valores calculados na etapa de análise, foi gerado um gráfico de Dispersão para visualizar a relação entre a frequência e o total gasto por cada cliente. Além disso, a análise de frequência permite identificar quais clientes estão mais propensos a retornar à empresa, e criar estratégias para incentivar esses clientes a fazerem novas compras.

Gráfico 1: Gráfico de dispersão da frequência de compra X Total gasto



Após a análise do gráfico de dispersão, verificou-se a existência de uma correlação positiva moderada entre a frequência e o total gasto. O coeficiente de correlação de Pearson foi de 0.288, o que indica que clientes que compram com maior frequência tendem a gastar mais dinheiro na loja. Esse resultado é importante para o modelo de regressão linear, uma vez que essas variáveis podem ser utilizadas como preditores para a recompra de clientes. Além disso, essa correlação também sugere que estratégias de fidelização de clientes, como programas de recompensa ou promoções exclusivas para clientes mais frequentes, podem ser efetivas na atração e retenção de clientes na loja.

2.2 Etapa formulação do modelo e resultados dos modelos:

Nessa etapa, além da análise de correlação entre a frequência e o total gasto dos clientes, foram criados dois modelos para identificar quais clientes têm maior probabilidade de retornar à empresa. O primeiro modelo, baseado no valor gasto, listou os 100 possíveis clientes que mais gastaram e utilizou o algoritmo de Floresta Aleatória.

Já o segundo modelo, baseado na frequência de compra, foi utilizado para identificar clientes que tendem a retornar mais vezes à empresa. Esses dois modelos foram criados para realizar uma análise comparativa e avaliar qual modelo seria mais adequado para um possível plano de ação.

O modelo de floresta aleatória (Random Forest) é um algoritmo de aprendizado de máquina que utiliza um conjunto de árvores de decisão para gerar previsões. No caso desse projeto, foram utilizadas as variáveis 'Quantidade', 'Preço Unitário', 'Desconto' e 'Frete' como preditores para identificar os 100 clientes mais propensos a fazerem uma recompra

A acurácia do modelo de floresta aleatória foi de 81%. Essa métrica representa a proporção de previsões corretas em relação ao total de previsões feitas pelo modelo. Uma acurácia de 81% indica que o modelo teve um bom desempenho na classificação dos clientes que têm maior probabilidade de fazer a recompra.

Após a seleção dos 100 possíveis clientes, o modelo foi submetido a testes para verificar a precisão das previsões. Foi obtido um resultado bastante satisfatório, onde a proporção de clientes que realizaram a recompra dentre os top 100 clientes é de 99.66%. Isso indica que o modelo de floresta aleatória teve um resultado positivo em relação a previsão de recompra desses clientes, logo em seguida fiz a lista desses possíveis clientes.

Imagem 2: Listagem de alguns clientes que foram disponibilizados pelo modelo Floresta aleatória

	ID_Cliente	Data	ID_Produto	Descrição_Produto	Quantidade	Preço_Unitário	ID_Pedido	Desconto	Frete	Total_do_Pedido	recompra	recompra_predita	Categoria
28	1.597749e+10	2022-12-17	281.0	Bázica Gola C - Grafite - Air - GG	1	119.0	7851	59.5	0.0	535.5	1	1	De R$ 400 a R$ 800
1205	1.598475e+10	2022-12-24	462.0	Bázica Long - Ocean - Air - GG	1	119.0	8521	0.0	0.0	476.0	1	1	De R$ 400 a R$ 800
1285	1.560540e+10	2022-12-24	405.0	Bázica Polo - Preta - Pima - XGG	1	257.0	8569	0.0	0.0	495.0	1	1	De R$ 400 a R$ 800
1287	1.560540e+10	2022-12-24	209.0	Bázica Lord - Branca - Pima - M	1	247.0	8571	0.0	0.0	623.0	1	1	De R$ 400 a R$ 800
1288	1.560540e+10	2022-12-24	84.0	Bázica Gola C - Salmon - Air - XXGG	1	119.0	8571	0.0	0.0	623.0	1	1	De R$ 400 a R$ 800
...	...	...	...	...	...	...	...	...	...	...	...	...	...
927	1.598429e+10	2022-12-23	55.0	Bázica Gola C - Branca - Air - P	1	119.0	8366	0.0	0.0	1071.0	1	1	Acima de R$ 800
928	1.598429e+10	2022-12-23	73.0	Bázica Gola C - Bordô - Air - P	1	119.0	8366	0.0	0.0	1071.0	1	1	Acima de R$ 800
929	1.598429e+10	2022-12-23	163.0	Bázica Gola C - Bali - Air - P	1	119.0	8366	0.0	0.0	1071.0	1	1	Acima de R$ 800
930	1.598429e+10	2022-12-23	31.0	Bázica Gola C - Verde Militar - Air - P	1	119.0	8366	0.0	0.0	1071.0	1	1	Acima de R$ 800
931	1.598429e+10	2022-12-23	433.0	Bázica Gola C - Toffee - Air - P	1	119.0	8366	0.0	0.0	1071.0	1	1	Acima de R$ 800

100 rows × 13 columns

Foi realizada uma análise mais detalhada dos dados, dividindo os clientes em três categorias distintas com base na quantidade total gasta em compras na loja. A primeira categoria incluiu os clientes que gastaram até R$ 400,00 reais, a segunda categoria incluiu aqueles que gastaram entre R$ 400,00 e R$ 800,00 reais e a terceira categoria incluiu os clientes que gastaram acima de R$ 800,00 reais. Essa análise por categorias pode ser útil para destinar produtos mais caros para determinado tipo de cliente. Por exemplo, se um cliente é identificado como pertencente à categoria de gastos acima de R$ 800, pode ser uma boa estratégia oferecer a ele produtos de maior valor agregado. Além disso, essa segmentação pode ajudar na criação de campanhas de marketing mais direcionadas e personalizadas para cada categoria de cliente.

2.3 Etapa do Modelo de Regressão Linear

Utilizando a regressão linear para criar um modelo mais simples, que levou em consideração apenas a frequência e o total gasto por cada cliente, gerando uma lista de 100 possíveis clientes para recompra.

Esse modelo é útil para identificar clientes que compram com mais frequência e que podem ser direcionados para ofertas e promoções específicas, a fim de aumentar a fidelização desses clientes. A análise dos dados por frequência pode ser uma estratégia interessante para destinar produtos mais caros ou mais exclusivos para determinado tipo de cliente.

Imagem 3: Listagem de alguns clientes que foram disponibilizados pelo modelo Regressão Linear

	ID_Cliente	Freq_Compra	Total_Gasto	Recompra
16	1.226840e+10	43	23123.30	1
235	1.408880e+10	31	14518.80	1
88	1.273229e+10	25	6543.03	1
1171	1.569230e+10	23	6556.00	1
26	1.242904e+10	22	7971.50	1
1032	1.566235e+10	21	8295.40	1
309	1.446321e+10	20	9085.00	1
1776	1.578848e+10	20	6527.98	1
36	1.242935e+10	20	6177.00	1
954	1.562501e+10	19	16566.92	1
855	1.560633e+10	19	12330.00	1

Além disso o modelo gerou R2 score: 0.336 é uma medida estatística que indica o quão bem o modelo de regressão linear se ajusta aos dados observados. O valor do R2 varia de 0 a 1, sendo que um valor mais próximo de 1 indica que o modelo consegue explicar melhor a variabilidade dos dados observados.

No caso específico mencionado, um R2 score de 0.336 significa que o modelo de regressão linear explica 33.6% da variabilidade dos dados observados.

O modelo foi testado também através de uma técnica de estatística descritiva que foi o de hipótese binomial para verificar se a proporção de acertos em uma amostra é estatisticamente significativa

Nesse caso, o teste foi utilizado para verificar se a proporção de clientes que irão fazer a recompra, dentre os 100 possíveis clientes selecionados pelo modelo, é significativamente maior do que 50% (que seria o esperado caso o modelo não tivesse nenhum poder de previsão)

O resultado do teste deu que o P = 0 ou podemos rejeitar a hipótese nula com um alto grau de confiança e afirmar que o modelo baseado na frequência é melhor do que as previsões aleatórias.

2.4 Etapa final: Plano de Ação:

Com base nas informações adquiridas pelos modelos, é possível identificar padrões e tendências no comportamento dos clientes e no desempenho da empresa. Com isso, torna-se viável traçar estratégias para melhor concretizar os resultados previstos e fazer com que essas previsões se tornem reais e válidas.

Para isso, é fundamental que a empresa utilize as informações obtidas pelos modelos para guiar suas decisões e ações. Isso pode envolver o desenvolvimento de novas estratégias de marketing e vendas, a criação de programas de fidelidade, o aprimoramento do atendimento ao cliente, entre outras iniciativas.

Assim, um modelo de plano de ação com base na ferramenta 5W2H serve como uma possível estratégia para a empresa levando em consideração os resultados previstos.

Quadro 1: Plano de ação baseado na ferramenta 5W2H

O que?	Por quê?	Como?	Quando?	Onde?	Quem?	Quanto?
Segmentação de clientes em grupos e criação de campanhas de marketing e promoções específicas para cada grupo	Para aumentar as recompras de clientes e, consequentemente, as vendas da empresa Bazico nos próximos 14 dias	Criação de campanhas de marketing e promoções específicas para cada grupo de clientes identificados	Início imediato e duração de 14 dias	Em todas as regiões de atuação da empresa Bazico	Equipe de marketing e vendas da empresa Bazico	O Custo dependerá do saldo disponivel para o tráfego
Criação de um programa de recompensas para incentivar a recompra de clientes	Para aumentar as recompras de clientes e, consequentemente, as vendas da empresa Bazico nos próximos 14 dias	Criação do BaziPontos em que quando os clientes adquirice uma certa quantidade de pontos através da obtenção de produtos poderia ser convertido em algum item exclusivo da loja	Início imediato	Em todas as regiões de atuação da empresa Bazico e no site da empresa	Equipe de marketing e vendas da empresa Bazico	Será com base no valor do acumulo dos pontos
Implementação de um programa de fidelidade para os 100 clientes selecionados.	Será feito para aumentar a satisfação e a fidelidade dos clientes selecionados, gerando um aumento nas vendas e no faturamento da empresa..	O programa de fidelidade será divulgado através de e-mails, mensagens de texto e publicações nas redes sociais. Os clientes selecionados serão informados sobre as vantagens do programa e como utilizar os descontos exclusivos.	O programa será implementado nos próximos 30 dias e terá duração de 6 meses, podendo ser prorrogado de acordo com os resultados obtidos.	O programa será implementado em todas as lojas da empresa e também poderá ser utilizado no site e nas redes sociais.	A equipe de marketing será responsável por desenvolver o programa de fidelidade e a equipe de vendas será responsável por monitorar a utilização do programa pelos clientes.	O custo do programa de fidelidade dependerá das vantagens oferecidas aos clientes, mas o investimento pode ser recuperado com o aumento das vendas e da fidelidade dos clientes.

3.0 Considerações finais

O case foi uma ótima experiência, pois aprendi muito e realizei tarefas que nunca havia imaginado fazer antes. Os conceitos sobre aprendizado de máquina foram novos e desafiadores, e me motivaram ainda mais para pesquisar coisas relacionadas e aplicar todo o conhecimento que possuo nas outras bibliotecas (pandas, numpy e matplot).

Foi muito satisfatório aplicar conhecimentos que só tinha experiência através das lições e tutoriais de cursos que normalmente segue um passo a passo definido. Cada etapa do desafio me fascinou, pois pude perceber o poder da criação e a habilidade de analisar uma base de dados tão grande e manipular todas as informações com facilidade. Vi a grande diferença entre fazer uma análise em Pynthon comparado ao Excel.

Dediquei-me totalmente a cada etapa e utilizei bastante todas as ferramentas que eu tinha disponível, dentre elas, o ChatGPT, que me ajudou a encontrar um caminho significativo e corrigir os bugs que encontrei durante o processo, que a propósito não foram poucos. Acredito que cheguei a um resultado significativo, considerando o conhecimento que possuo. Fiz meu máximo, pesquisei muito para chegar em um resultado que tivesse uma lógica e contemplasse as etapas do desafio.

Foi importante buscar informações em cursos, questionar a colegas e professores se os resultados obtidos faziam sentido ou se havia algo para aperfeiçoar e poder entregar o melhor resultado possível. Fico muito feliz em ver que, mesmo nessa parte inicial do processo seletivo já aprendi muito e me identifiquei ainda mais com essa área. Hoje almejo me tornar um grande especialista e espero poder continuar a aprender e crescer na área de dados dentro da **Bazico**.