Teste Prático

Candidata: Anna Paula Figueiredo Gonçalves
Vaga: Cientista de Dados Júnior

Agenda

- Descrição do Problema
- Ferramentas Utilizadas
- Pré processamento dos dados
  - Base de Dados
- Regras de Associação
  - Algoritmo Apriori
- Resultados

Descrição do Problema

A partir de um arquivo **Access**, com os dados de venda de um mercado. O objetivo é encontrar itens com **potencial de venda conjunta**.

Por exemplo, identificar que compradores de cerveja têm **maior probabilidade** de comprar fraldas, com isso seria sugerido deixar estes itens na mesma prateleira.

A análise pode ser realizada na ferramenta que você achar mais pertinente. Além disso, uma apresentação descrevendo a lógica e os tratamentos dos dados também deverá ser entregue.

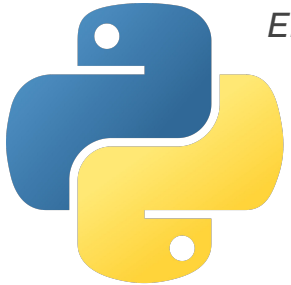
Ferramentas e Bibliotecas Utilizadas

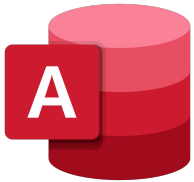
Access database Engineer

Python 3.9

- *Pandas*
- *Pyodbc*
- *Plotly*
- *Mlxtend*

Visual Studio Code





Pré-processamento dos dados

Leitura

Conexão com o banco através da biblioteca ***pyodbc.***

Manipulação

Transformou-se as tabelas do **SQL** para um *data frame* da biblioteca ***pandas.***

Qualidade dos Dados

Verificação de dados duplicados, dados ausentes, máximos e mínimos e valores únicos.

Base de Dados

tabela item			
codItem	descrição	marca	tipo

tabela consultas			
codItem	descrição	marca	tipo

tabela transações		
IDTransação	valorTotal	tipo pagamento

tabela item transação	
IDTransação	item

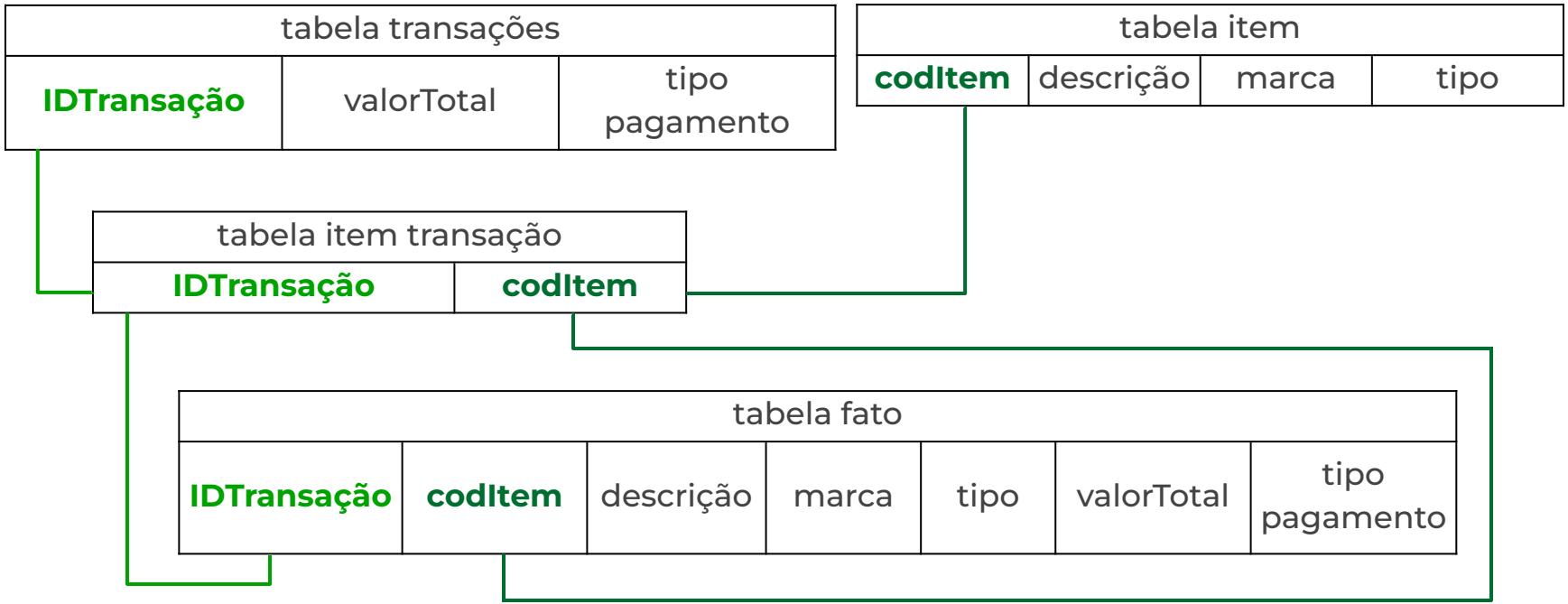
A base de dados é composta por 4 tabelas: itens, consultas, transações e a tabela item transação. Esse tipo de tabela pode ser chamada de dimensão, pois nela encontramos uma chave primária e atributos relacionados somente ao escopo definido pelo nome da tabela.

Foi verificado que a tabela item e consulta eram iguais então a tabela consulta foi inutilizada. Além disso, 7 ids presentes na tabela transações, não tiveram correspondência na tabela item transação, impossibilitando o relacionamento dessas transações com a venda de um produto.

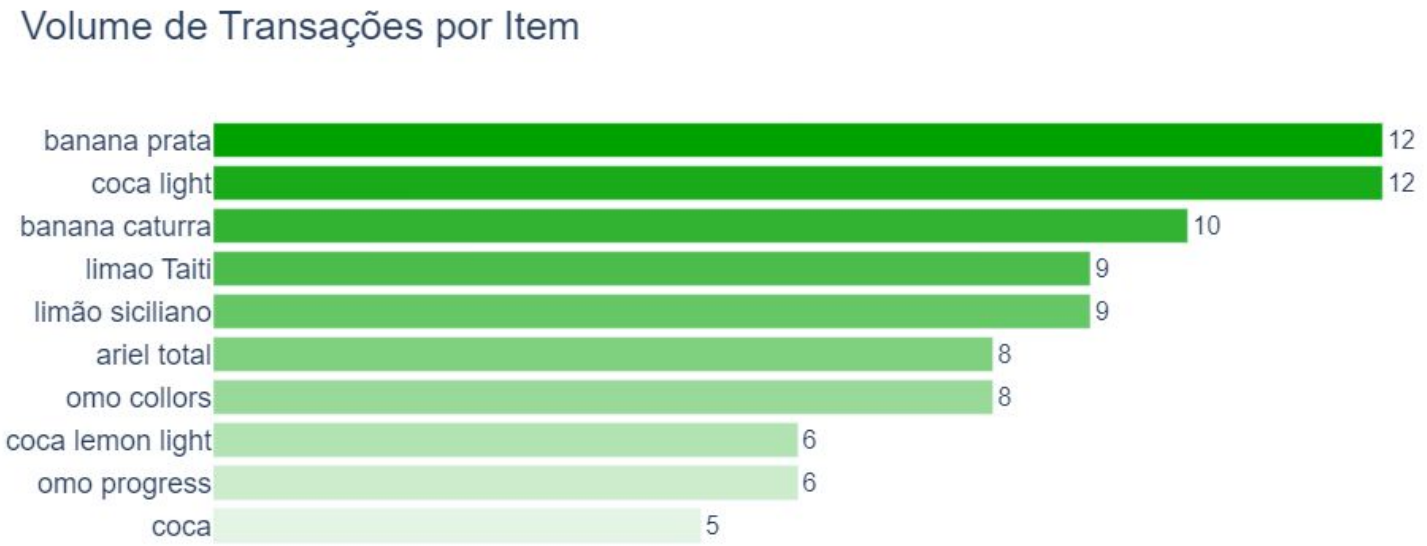
A partir disso foi construído a tabela fato, onde concentra-se todas as operações realizadas. Ela foi construída através de joins com as chaves primárias das tabelas dimensões. O modelo ***star schema*** é apresentado na próxima página.

Tabela Fato

A tabela fato é considerada a tabela principal desta análise, pois condensa todas as informações das vendas.



Saída de Itens



Observações

Na etapa de **pré-processamento** dos dados, foi identificado possíveis incongruências na base de dados. (escrita incorreta)

codItem	descrição	marca	tipo
3	limao Taiti	limao	fruta
4	limão siciliano	limão	fruta

codItem	descrição	marca	tipo
5	coca	coca	refrigerante
6	coca light	coca	refrigerante
7	coca lemon light	coca	refirgerante

Itens de código distintos
Escrita incorreta

Tais fatores devem ser avaliados com o cliente, para verificar a legalidade dos fatos. Entretanto, geralmente o padrão é cada produto ter um identificador único, a princípio foi respeitada essa ordem.

Regras de Associação

As Regras de Associação são muito utilizadas no campo de **Mineração de Dados** e **Machine Learning** quando o objetivo é associar a saída de um item com outro item, seja em: cesta de compras, sequência de ações realizadas em sites, compras em supermercado, recomendação de séries e filmes, entre outras aplicações.

Seu conceito básico é identificar elementos ou ações que implicam em outro elemento ou ação. Como, por exemplo:

Clientes que compram leite e pão, acabam comprando manteiga.

{leite, pão} → {manteiga}

Clientes de comprarem carne e carvão, acabam comprando cerveja e refrigerante.

{carne, carvão} → {cerveja, refrigerante}

Algoritmo Apriori

O **Algoritmo Apriori** é uma referência em **Regras de Associação**. Seu objetivo é identificar regras, itens, que quando comprados solo ou em conjunto, implicam na compra de um ou mais produtos específicos.

Para a aplicação desse algoritmo é necessária uma **base histórica de transações,** que contém o **conjunto de itens** comprados.

O algoritmo recebe como entrada uma tabela onde **cada linha representa uma transação** e **cada coluna representa os itens presentes na base de dados**. Os valores das linhas é representado por **zero quando o item não foi comprado**, e **um quando o item foi comprado**, conforme o exemplo abaixo:

idTransação	descrição_ariel total	descrição_banana caturra	descrição_banana prata	descrição_coca	descrição_coca lemon light	descrição_coca light
1	1	0	1	0	0	0
2	1	0	0	0	0	0
3	1	1	0	0	0	0

Algoritmo Apriori

1° calcula-se a frequência dos elementos na base de dados.

Todo item do conjunto de transação, deve obedecer um **suporte mínimo** na base de dados, isto é:

Suporte mínimo: a proporção de transações que contém um item em específico.

supp(item) = (n° de transações que o item aparece) / (n° total de transações)

Exemplo:

transação	pão	leite	manteiga
t1	1	1	1
t2	1	0	0
t3	1	0	1

supp(pão) = 3/3 = 1
supp(leite) = 1/3= 0.33
supp(manteiga) = 2/3 = 0.66

Algoritmo Apriori

Para definir uma porcentagem ideal de saída de um produto é interessante estar alinhado com o objetivo do negócio. Neste aspecto é importante definir qual a porcentagem mínima (**suporte mínimo**) de saída de um produto é ideal para escalar as vendas ?

Para este estudo, o suporte mínimo foi reduzido à **0.1**. Isto é, o item fica elegível à regra se ele tiver **saído em pelo menos 10% das compras**. Neste caso, todos os 10 itens são elegíveis ao suporte mínimo.

Para identificar esses itens elegíveis ao suporte mínimo foi utilizado a função **Apriori** da biblioteca **Mlxtend.**

Observações: Foram realizados testes com outros *thresholds* para o suporte mínimo, mas não foi possível gerar regras a partir deles, então optou-se pelo **0.1.**

Algoritmo Apriori

2° geração das regras de associação.

Para gerar as regras de associação, é importante termos como mensurar o desempenho, o quanto estão válidas ou não.

Por isso têm-se duas métricas muito importantes, que são a **confiança** e o **lift.**

A **confiança** mensura a **frequência com que aquele fato (regra) é observado no conjunto de dados**. Em outras palavras, quantas vezes aquele conjunto de itens ou seja quem compra pão e compra manteiga, aparece no conjunto total de transações.

conf(item a → item b) = (item a U item b) / (n° total de transações)

Algoritmo Apriori

1ª Fase:
Aplicando o suporte mínimo de 0.10, todos os itens ficam elegíveis à fase 2.

transação	pão	leite	manteiga
t1	1	1	1
t2	1	0	0
t3	1	0	1

supp(pão) = 3/3 = 1
supp(leite) = 1/3= 0.33
supp(manteiga) = 2/3 = 0.66

2ª Fase:
Gera todas as combinações de itens e calcula a confiança.

combinações de itens	confiança
{pão} ➔ {leite}	⅓ = 0.33
{pão} ➔ {manteiga}	⅔ = 0.66
{manteiga} ➔ {leite}	⅓ = 0.33
{pão, leite} ➔ {manteiga}	⅓ = 0.33
{leite, manteiga} ➔ {pão}	⅓ = 0.33
{pão, manteiga} ➔ {leite}	⅓ = 0.33

Se aplicado uma confiança mínima de 0.5 isto é, a regra ter aparecido em pelo menos 50% das transações, somente a regra {pão ➔ manteiga} fica válida.

Algoritmo Apriori

O **lift** é uma métrica que calcula a **possibilidade de um item ser comprado em relação ao outro item**, e também considera a popularidade de cada item individualmente. Essa métrica mede a força da relação dos itens, se eles estão correlacionados, ou atuam de maneira independente.

lift(item 1 → item 2) = supp(item 1 U item 2) / supp(item 1) * supp(item 2)

Tem-se então:

Se lift(item 1 → item 2) >= 1 , existe uma correlação clara entre os produtos. Caso contrário, não existe correlação clara entre os dois produtos.

Para a geração de regras, e cálculos das métricas, foi utilizada a função **association rules** biblioteca **Mlxtend.**

Foi passado como threshold para **confiança**, o valor de **0.5**. Isso significa que em pelo menos 50% dos casos existe aquela regra, ou seja, a compra de um item implica na compra de outro item seja verdade. Ademais, foi aplicado um filtro para retornar somente os com Lift maior ou iguais à 1.

Resultados

Resultados: Regras Geradas

Com o suporte mínimo de 10% e a confiança de 50%, foram geradas 18 regras.

Isto é sempre que um item ou um conjunto de itens forem comprados, implicará também na compra de outro item.

Como, por exemplo: Os clientes que compram banana caturra e coca light, também compram limão siciliano 100% das vezes(confiança).

compra de itens	suporte	confiança
{banana caturra, coca light} → {limão siciliano}	0,14	1,0
{banana caturra, limão siciliano} → {coca light}	0,14	1,0
{coca light, limão siciliano} → {banana caturra}	0,14	1,0
{omo progress, limao Taiti} → {coca light}	0,11	1,0
{omo progress} → {coca light}	0,18	0,83
{banana prata, limao Taiti} → {coca light}	0,11	0,75
{coca light, limao Taiti} → {banana prata}	0,11	0,75
{coca light, limao Taiti} → {omo progress}	0,11	0,75
{coca} → {banana caturra}	0,11	0,60
{banana prata, coca light} → {limao Taiti}	0,11	0,60
{omo progress, coca light} → {limao Taiti}	0,11	0,60
{limão siciliano} → {banana prata}	0,18	0,55
{ariel total} → {limao Taiti}	0,14	0,50
{omo collors} → {banana prata}	0,14	0,50
{omo collors} → {coca light}	0,14	0,50
{coca lemon light} → {limao Taiti}	0,11	0,50
{omo progress} → {limao Taiti}	0,11	0,50
{omo progress} → {coca light, limao Taiti}	0,11	0,50

Resultados: Regras Geradas

Note que sempre que a confiança for igual a um. Significa que sempre que a regra criada sempre vai ocorrer na base de transações.

Por isso, se o objetivo for facilitar a vida do cliente, a indicação é colocar especialmente os itens de confiança igual a um, próximos.

Por outro lado, se o objetivo for vender mais, ou seja, aumentar o número de transações, de tal modo a induzir a compra, é indicado uma análise mais aprofundada desses itens, inclusive apresentando os valores de cada produto.

Entretanto, para fins de maior exploração da base ofertada, foi estimado a aplicação dessas regras geradas ao nível de transação. Esses resultados são apresentados nas próximas páginas.

Potencial de Transações por Regra Aplicada

Para estimar o potencial de transações dado a disposição dos produtos na prateleira de acordo com as regras, foi feito o cálculo da seguinte maneira:

Dado a regra {limão siciliano} → {banana prata} de exemplo.

O total de transações antecedentes, é representado por todas as transações em que houve a compra de limão siciliano.

O potencial de transações é representado por todas as transações que ocorreram compra do limão siciliano e **não ocorreu** a venda de banana prata. Isto é: as transações que se o limão fosse colocado ao lado da banana prata, provavelmente o cliente levaria também a banana prata.

Logo, se a regra fosse aplicada e em todos os casos os clientes que comprassem limão siciliano também comprasse banana prata, teríamos um aumento de 44% no volume de transações que envolvem o limão siciliano.

regras sobre itens	total de transações em antecedentes	potencial de transações	taxa de aumento
{limão siciliano} → {banana prata}	9	+4	44%

Potencial de Transações por Regra Aplicada

A tabela ao lado apresenta o potencial de transações caso fosse aplicado às regras, isto é, a disposição dos itens lado a lado e tivesse de fato induzido o cliente à compra.

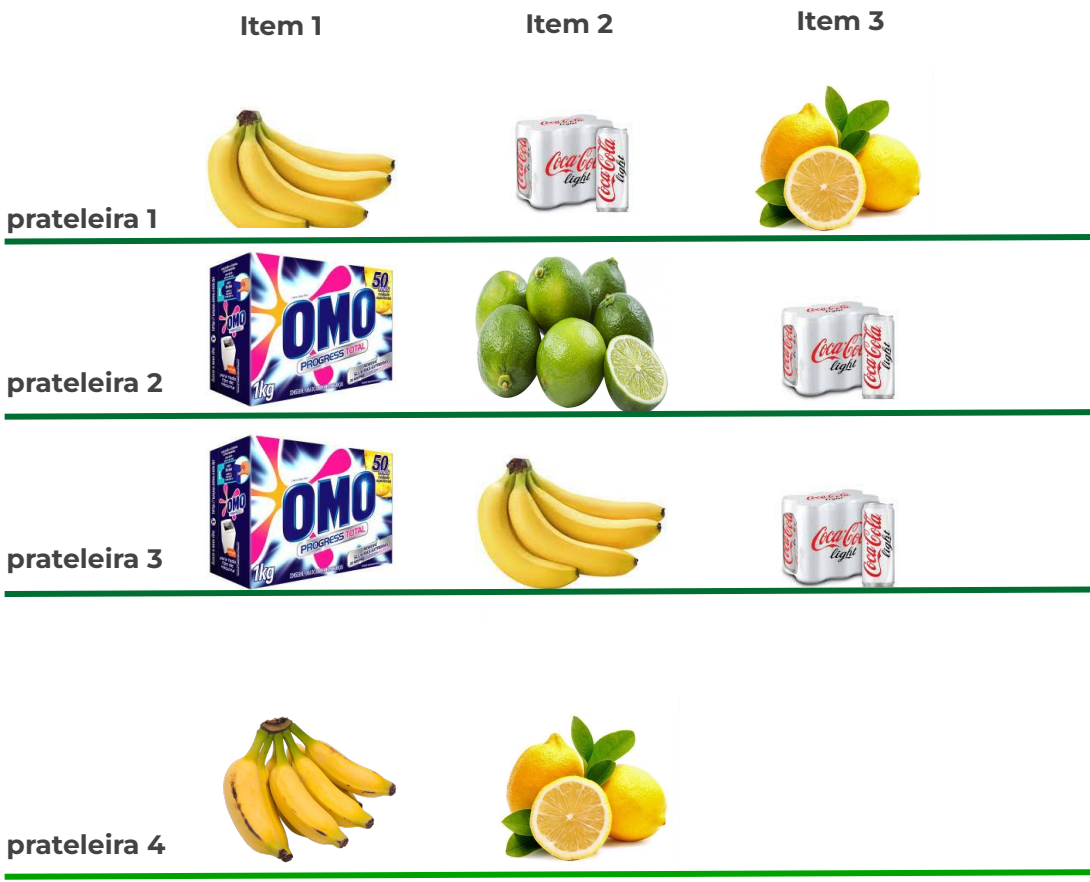
regras sobre itens	total de transações em antecedentes	potencial de transações	taxa de aumento
{limão siciliano} → {banana prata}	9	+4	44%
{ariel total} → {limao Taiti}	8	+4	50%
{omo collors} → {banana prata}	8	+4	50%
{omo collors} → {coca light}	8	+4	50%
{coca lemon light} → {limao Taiti}	6	+3	50%
{omo progress} → {limao Taiti}	6	+3	50%
{coca} → {banana caturra}	5	+2	40%
{banana prata, coca light} → {limao Taiti}	5	+2	40%
{coca light, omo progress} → {limao Taiti}	5	+2	40%
{omo progress} → {coca light}	6	+1	17%
{banana prata, limao Taiti} → {coca light}	4	+1	25%
{coca light, limao Taiti} → {banana prata}	4	+1	25%
{coca light, limao Taiti} → {omo progress}	4	+1	25%
{omo progress} → {coca light, limao Taiti}	6	+1	17%

Disposição Final dos Produtos

Produtos na Prateleira

Nas prateleiras de 1 a 3 é possível identificar os produtos que sairiam 100% das vezes, isto é: sempre que o cliente comprar o item 1 e o item 2, ele também compraria o item 3.

Já na prateleira 4, cujo objetivo é induzir a venda onde:
{limão siciliano} → {banana prata}. Inverte-se a ordem do produto para que os clientes que irão comprar o limão siciliano, passe pela banana prata pelo menos 2x, uma na ida e outra na volta.



Obrigada!

Dúvidas, sugestões, feedbacks:


anna.figueiredo@aluno.ufop.edu.br


AnnaPaulaFigueiredo


anna-paula-figueiredo

Medium @annapaulafigueiredoo