Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Ana Carolina Cunha Bueno

Análise do perfil de compra dos revendedores de produtos de uma empresa de cosméticos

Curitiba
2020

Ana Carolina Cunha Bueno

Análise do perfil de compra dos revendedores de produtos de uma empresa de cosméticos

Monografia apresentada ao Programa de Especialização em Data Science e Big Data da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Abel Soares Siqueira

Curitiba

2020

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

Análise do perfil de compra dos revendedores de produtos de uma empresa de cosméticos

Ana Carolina Cunha Bueno[1]
Prof. Abel Soares Siqueira[2]

Resumo

As técnicas de mineração de texto tem como objetivo encontrar padrões, correlações ou anomalias em uma grande quantidade de dados. Esses padrões podem ser usados para definir uma estratégia do negócio ou para identificar um comportamento pouco usual. Para que a relação com os clientes seja duradoura e de confiança, é necessário conhecer o perfil do consumidor e estimular a compra, oferecendo produtos e ofertas interessantes frente à concorrência. Este trabalho tem como principal objetivo dar subsídios a uma empresa de comercialização de cosméticos para ofertar outros produtos no momento da inclusão de itens no carrinho de compras. Foram analisadas transações de revendedores desta empresa. O estudo das Regras de Associação com recurso a técnicas de Data Mining, especificamente de Market Basket Analysis foi o foco deste artigo. Este estudo proporcionou a criação de cestas de compras frequentes que podem ser usadas para sugestão/recomendação no momento da compra, através dos aplicativos e sites da empresa.
Palavras-chave: Mineração de Dados, Regras de Associação, Market Basket Analysis, Apriori.

Abstract

Techniques in Data Mining has the objective to find patterns, correlations or anomalies in a large amount of data. These patterns can be used to define a business strategy or to identify unusual behavior. For the relationship with customers to be long-lasting and trustworthy, it is necessary to know the profile of the consumer and stimulate the purchase, offering interesting products and offers ahead of the competition. This work has as main objective to give subsidies to a cosmetics commercialization company to offer other products when adding items to the market basket. Dealer transactions from this company were analyzed. The study of the Association Rules using Data Mining techniques, specifically the Market Basket Analysis was the focus of this article. This study provides the creation of frequent shopping baskets that can be used for suggestion / recommendation at the time of purchase, through the company's applications and websites.
***Keywrods**: Data Mining, Association Rules, Market Basket Analysis, Apriori.*

1 Introdução

O uso de dados no varejo tornou-se uma obrigação para quem deseja ter sucesso no relacionamento e na análise da jornada do consumidor. Pesquisas de mercado, informações do ponto de venda, clubes fidelidade, entre outros, são algumas das fontes que geram conhecimento sobre o cliente. É necessário analisar os dados e cruzar informações para tomar decisões a partir dos mesmos.

Para ter uma relação duradoura e de confiança com o consumidor, é necessário conhecê-los a fim de estimular o consumo de mais e novos produtos e serviços. Como no varejo a maior parte das compras é realizada por impulso, dar sugestões/recomendações interessantes ao cliente do que comprar pode trazer vantagens competitivas frente à concorrência. Identificar padrões de compra e conhecer o perfil do consumidor são práticas importantes para aumentar o potencial dessa recomendação.

Muitas empresas buscam por revendedores para aumentar seu lucro e diversificar as formas de venda. Os revendedores são uma forma de contato com o consumidor e são essenciais para a empresa em análise. Deste modo, este trabalho terá seu foco em gerar recomendações de compras para revendedores, através da técnica de Market Basket Analysis, no português, análise da cesta de compras.

A Market Basket Analysis é considerada uma das áreas mais antigas de Data Mining [1] a qual pretende descrever o comportamento do consumo de clientes e dessa análise retirar conclusões sobre os padrões de compra.

Esta técnica tem como objetivo descobrir combinações de itens que ocorrem com frequência acima do esperado em uma base de dados.

2 Materiais e Métodos

Nesse capítulo serão apresentadas informações úteis para pleno entendimento das análises realizadas. Será tratado desde o processo de descoberta de conhecimento a partir dos dados, até o modelo utilizado.

A metodologia empregada possui quatro grandes blocos:

- Pré-processamento dos dados;

- Clusterização de revendedores;
- Aplicação do algoritmo Apriori;
- Análise e refinamento dos resultados.

2.1 O conjunto de dados

O conjunto de dados utilizado no presente estudo foi disponibilizado por uma empresa de cosméticos que comercializa produtos como perfumes, cremes, shampoos, maquiagens, etc. Os dados são privados e não serão disponibilizados.

Os dados são referentes à compras de janeiro/2018 à janeiro/2020, possuem aproximadamente 5,5 milhões de transações e mais de 1 milhão de revendedores.

2.2 Recursos Computacionais

A linguagem SQL foi utilizada para criação da ABT (Analytical Base Table) do modelo. O software Python (versão 3.7.4) foi utilizado para análise dos dados e ajuste do modelo. Foram usados os pacotes: pandas, numpy, apyori, entre outros.

2.3 Clusterização utilizando K-Means

A tarefa de clusterização é utilizada para separar os registros de uma base de dados em subconjuntos ou clusters (agrupamentos), de tal forma que os elementos de um cluster compartilhem propriedades comuns [2]. É do conhecimento da empresa que revendedores possuem perfis de compra diferentes. Visando reduzir o viés da análise e criar combinações de itens para compra baseando-se no perfil de cada revendedor, foi necessária a criação de clusters.

O objetivo desta análise foi agrupar pontos de dados semelhantes e descobrir padrões entre os revendedores.

O algoritmo de K-Means possui classificação não-supervisionada e é um dos mais simples e populares algoritmos de aprendizado de máquinas. Os algoritmos não supervisionados fazem inferências de conjuntos de dados utilizando apenas vetores de entrada, sem se referir a resultados conhecidos ou rotulados.

O processo executado pelo K-Means é composto por quatro etapas: inicialização, atribuição ao cluster, movimentação de centroides e otimização do K-médias.

Para escolha do número de clusters foi utilizado o método de agrupamento Elbow. O método consiste em calcular a função de custo, a soma dos quadrados das distâncias internas dos clusters, e traçá-la em um gráfico. O melhor número para a quantidade de clusters é quando a adição de um novo cluster não muda significativamente a função de custo.

2.4 Market Basket Analysis

O MBA é o processo que utiliza a técnica de regras de associação para analisar hábitos de compra de clientes e encontrar associações entre os diferentes itens que os clientes colocam em sua "cesta de compra". É uma técnica matemática, frequentemente usada por profissionais de marketing, para revelar afinidades entre produtos individuais ou grupo de produtos [3].

2.5 Apriori

O algoritmo Apriori foi a primeira ferramenta para descoberta de regras de associação em bases de dados com grandes volumes.

O algoritmo identifica os itens que fazem parte do conjunto de transações em questão e em seguida determina as regras de associação entre esses itens, selecionando as que ocorrem com mais frequência.

Fracalanza (2009)[4] traz uma definição de que o método Apriori consiste em encontrar todas as regras de associação que possuam suporte e confiança maiores ou iguais a um suporte mínimo (SupMin) e uma confiança mínima (ConfMin), especificados pelo usuário.

Este algoritmo utiliza três medidas de avaliação, sendo elas:

- Suporte - porcentagem de transações da base de dados que contêm os itens de A e B (popularidade padrão de um item);
- Confiança - entre as transações que possuem os itens de A, a confiança é a porcentagem de transações que possuem também os itens de B;
- Lift - medida da eficácia do model a qual é utilizada para definir o grau de interesse de uma regra de associação e é calculada como:

$$\text{Lift} = \frac{\text{confiança}(A + B)}{\text{suporte}(B)}.$$

Os valores de suporte mínimo e confiança mínima são parâmetros que devem ser fornecidos ao algoritmo.

Somente os itens com suporte $\geq$ suporte mínimo são considerados pelo algoritmo.

3 Resultados e Discussões

O primeiro passo da análise foi a criação da base de dados para modelagem. Todos os dados ficavam em tabelas diferentes, sendo necessária a utilização da linguagem SQL para cruzar cada uma delas.

Para o presente estudo, foram selecionadas compras de revendedores dos últimos dois anos.

Primeiramente foi realizado o pré-processamento dos dados e diversos tratamentos foram realizados na base, como por exemplo:

- Filtro de compras com produtos ativos;
- Produtos com mais de 1.000 vendas no mês (retirada de itens raros);
- Cestas de compras com mais de 1 item.
- Retirada de cestas com muitos itens (com quantidade de itens outliers);

Após a filtragem da base, restaram 1.049 produtos (cosméticos) para modelagem e aproximadamente 5,5 milhões de transações.

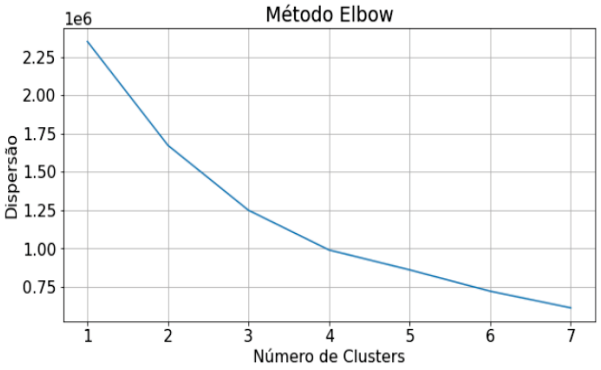
Como dito anteriormente, é do conhecimento da empresa que revendedores possuem perfis de compra diferentes. Por este motivo, uma clusterização de revendedores foi necessária.

Foram utilizadas as seguintes variáveis para encontrar os perfis distintos de revendedores: média de compras (mês), categoria de produto mais comprada (perfumaria, cuidados com a pele, maquiagem...), desconto médio, tempo como revendedor, média de itens na cesta e quantidade média de compras no mês.

Antes da aplicação do algoritmo, foi necessário o tratamento dos dados realizando uma normalização da base. Variáveis categóricas foram transformadas em dummies e variáveis numéricas foram agrupadas através de percentis e depois também transformadas em dummies.

Para escolha do número de clusters foi utilizado o método Elbow. Através do gráfico abaixo é possível concluir que o número ideal de clusters para a presente análise é 5.

Figura 1: :Método Elbow para escolha de número de clusters



Foram então gerados 5 clusters utilizando o algoritmo de K-MEANS. Analisando os centróides, concluímos que os clusters possuem os seguintes perfis:

- Revendedores que compram diversas categorias de produtos e várias vezes durante o mês;
- Revendedores que compram apenas perfumes e cuidados com a pele, várias vezes no mês;
- Revendedores que compram diversas categorias de produtos e 1 vez no mês;
- Revendedores que compram poucas categorias de produtos e 1 vez no mês;
- Revendedores que compram poucos produtos no mês (provalvemente a compra é para ele mesmo);

Para cada cluster foram selecionadas as transações dos revendedores (PEDIDO) e os códigos dos itens das cestas de compras.

Para que a aplicação do algoritmo APRIORI fosse realizada, as transações foram colocadas nas linhas e os códigos dos produtos nas colunas (sendo a coluna 1: primeiro item do carrinho e assim por diante). Caso um pedido possua menos de 10 itens, o campo relacionado é nulo.

Figura 2: :Base de entrada do algoritmo APRIORI

	ITENS									
PEDIDO	1	2	3	4	5	6	7	8	9	...
1	100823	102073	100149	109525	100871	110792	101374	110677	119918	...
2	105583	109753	105586	112540	112155	109997	100873	-	-	...
3	111565	112333	112675	100677	104982	100869	110256	100405	-	...
4	103025	101385	103111	106635	100871	-	-	-	-	...
5	100869	109620	112675	-	-	-	-	-	-	...
6	112679	100155	100871	100566	103025	100539	103109	100139	100879	...
7	111604	100256	112213	103025	112675	-	-	-	-	...
8	100406	101374	100539	111540	100869	100003	110794	104891	100512	...
...										

Sendo assim, em cada cluster foi aplicado o algoritmo APRIORI.

Duas abordagens foram utilizadas neste estudo. Primeiramente foram realizadas várias simulações utilizando diversos valores de suporte e confiança mínimos, em busca dos itens mais freqüentes, conforme a proposta original do algoritmo. A segunda abordagem tratou dos itens menos freqüentes, acrescentando ao algoritmo um limite máximo para o suporte.

A primeira constatação é que o número de regras geradas é inversamente proporcional aos valores determinados para confiança mínima e suporte mínimo, pois, à medida que estes últimos decrescem, aumenta o número de regras.

Tabela 1: Regras geradas em função de suporte e confiança mínimos.

Simulação	minsup	minconf	Nº Regras Geradas
S1	0.90	0.80	15
S2	0.60	0.70	252
S3	0.50	0.60	638
S4	0.30	0.70	1.324
S5	0.20	0.50	3.840

4 Conclusões

Para ter uma relação duradoura e de confiança como consumidor e trazer melhores resultados nas vendas, uma análise de Market Basket Analysis foi realizada numa base de uma empresa de comercialização de cosméticos.

O presente estudo teve como objetivo criar recomendações de produtos a partir da cesta de compras dos revendedores.

Por possuir um grande número de revendedores com perfis diferentes, uma clusterização foi necessária utili-

zando o algoritmo de K-MEANS. Os 5 clusters criados permitiram que as recomendações geradas fossem mais acuradas.

Após a criação dos clusters, a aplicação do algoritmo APRIORI foi realizada gerando diversas simulações e escolhendo a que possuía um bom número de regras geradas e com pouca redundância.

As recomendações serão aplicadas nos aplicativos e sites da empresa visando aumentar as vendas e antecipar as necessidades dos revendores. O próximo passo será a mensuração dos resultados e manutenção do modelo.

Como desafios futuros, poderão ser incluídas novas variáveis para clusterização dos revendedores e utilização de outras técnicas de Data Mining.

Agradecimentos

Agradeço ao Professor Abel pelo suporte e orientações durante a realização do estudo e ao meu marido Tiago Edelmo pelo apoio dado no decorrer de toda especialização.

Referências

[1] Troy Raeder and Nitesh V. Chawla. Market basket analysis with networks. *Social Network Analysis and Mining*, 1:97 – 113, 2011.

[2] Heimar de Fátima Galvão Noemi Dreyer Aan Marin. Técnica de mineração de dados: uma revisão da literatura. *Acta Paulista de Enfermagem*, 22:686 – 690, 10 2009.

[3] Micheline HAN, Jiawei KAMBER. Mineração de dados: Concepts and techniques. *USA: Morgan Kaufmann*, 2001.

[4] Troy Raeder and Nitesh V. Chawla. Market basket analysis with networks. *Social Network Analysis and Mining*, 1:97 – 113, 2011.