Online Retail Market Basket Analysis

Semar Augusto da Cunha Mello Martins

1. Introdução:

Market Basket Analysis, também conhecido como análise de afinidade, é uma técnica usada para identificar produtos que tendem a ser comprados juntos. Essa técnica é usada para aumentar as vendas pois se for possível identificar que ao se comprar X, os clientes compram Y também, o vendedor consegue ofertar o produto Y a todos os clientes que comprarem X.

A obtenção dessa informação é extremamente útil para fazer promoções - se o preço de X cair, as vendas de Y vão tender a subir também, para a criação de combos de vários produtos, para a organização da loja de produtos que são vendidos juntos serem colocados próximos ou mesmo na mesma categoria e para controle de inventário.

Durante esse trabalho, foi feita a análise de afinidade de ações vendidas entre 01/12/2010 e 09/12/2011. A partir da análise de regras de associação entre ações, talvez seja possível fazer previsões no mercado de ações, ou seja, se uma ação X estiver valorizando e houver uma regra que os clientes que compram X compram Y, talvez seja possível em algum nível dizer que Y valorizará também. E obviamente a análise de valorização de ações no mercado financeiro é extremamente valiosa.

A limpeza e manipulação dos dados foi feita toda em R, e os algoritmos de regras de associações foram rodados tanto em R quanto no Lemonade.

2. Base de Dados:

A base de dados usada no trabalho foi “Online Retail”, fornecida no site UCI, um repositório de bases de dados grátis para o uso em trabalhos como esse.

Ela consiste, como já foi dito na introdução, em compras de ações feitas entre 01/12/2010 e 09/12/2011. Essas vendas foram feitas principalmente no Reino Unido.

Atributos:

Cada linha representa a compra de uma ação única.

- InvoiceNo: Um número que corresponde à transação em que a ação foi comprada.
- StockCode: Código da ação comprada.
- Description: Nome da ação comprada.
- Quantity: Número de ações compradas.
- InvoiceDate: Dia e hora em que a ação foi comprada
- UnitPrice: Preço da unidade da ação.
- CustomerID: Número identificador do cliente comprador da ação.
- Country: País de onde foi feita a compra da ação.

Amostra da base de dados original:

Invoice No	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom

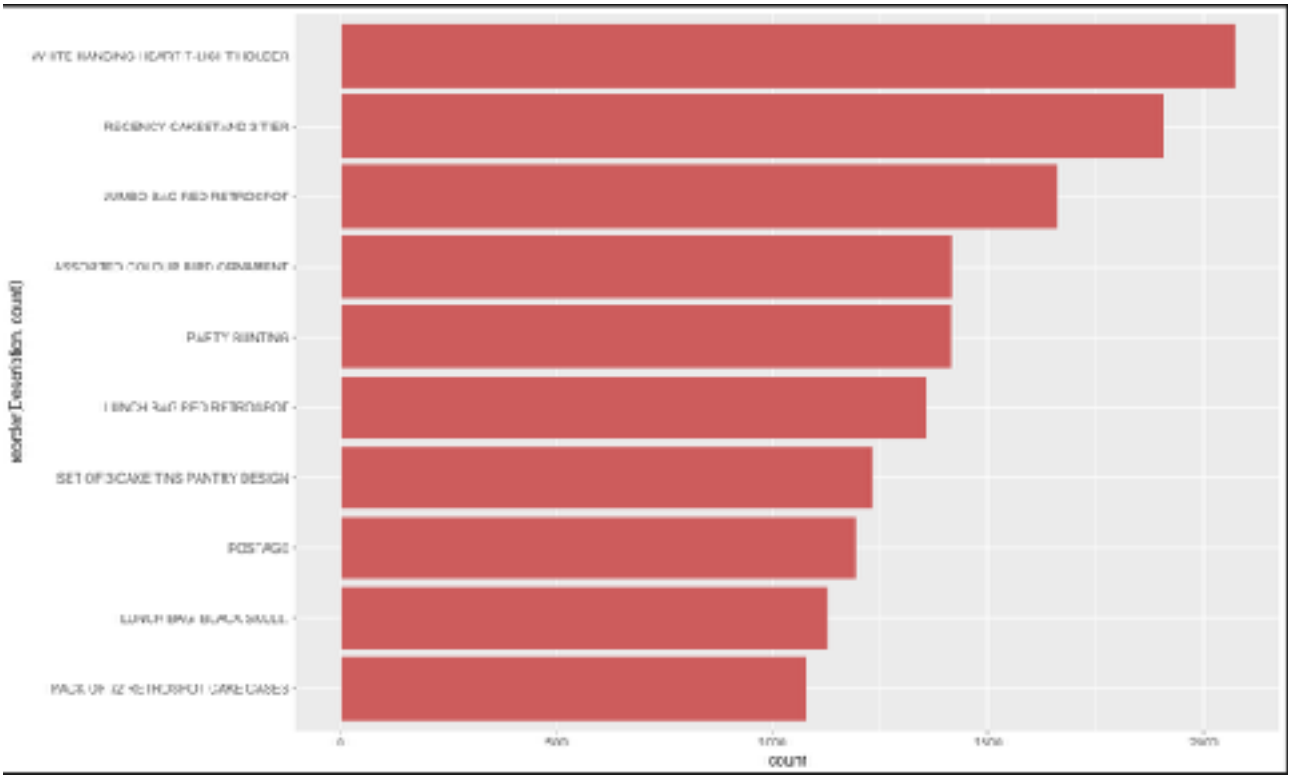


Figura 1: As 10 ações mais compradas no período.

3. Preparação dos Dados:

Para a fazer a análise de afinidade entre as ações, é necessário transformar os dados de forma que cada linha caracterize uma transação e cada coluna seja um produto comprado. Para isso, vamos retirar todas as colunas do dataset exceto "InvoiceNo" e “Descrição" pois vamos usar o nome da ação para a mineração e vamos usar InvoiceNo para agrupar todas as compras feitas numa mesma transação. Para isso, transformamos primeiro a base de dados em uma da forma:

InvoiceNo	Description
536365	WHITE METAL LANTERN
536365	CREAM CUPID HEARTS COAT HANGER

Depois agrupamos todas as descrições de mesmo InvoiceNo em uma mesma linha, ao fim dessa manipulação temos uma base de dados da forma que um algoritmo de padrões frequentes precisa.
Ao fim dessa manipulação, a base de dados ficou nesse formato:

Description
WHITE HANGING HEART T-LIGHT HOLDER,WHITE METAL LANTERN,CREAM CUPID HEARTS COAT HANGER,KNITTED UNION FLAG HOT WATER BOTTLE,RED WOOLLY HOTTIE WHITE HEART.,SET 7 BABUSHKA NESTING BOXES,GLASS STAR FROSTED T-LIGHT HOLDER
HAND WARMER UNION JACK,HAND WARMER RED POLKA DOT

Onde cada vírgula separa um item do outro.

Para o Lemonade foi feita uma etapa a mais de processamento em que foram retirados todos os itens repetidos de uma mesma transação.

3. Modelagem:

Devido à facilidade de uso, foi usada a biblioteca do R chamada “arules” (association rules). Nessa biblioteca há somente o algoritmo “apriori" para padrões frequentes. O Lemonade possui apenas o algoritmo “FPGrowth”. Infelizmente, comparar resultados entre os dois algoritmos é complicado pois o primeiro consegue gerar regras de múltiplos itens, enquanto o FPGrowth apenas gera regras de dois itens.

4. Avaliação:

4.1 - R:

O suporte escolhido para o algoritmo foi 5% e a confiança 90%. Esses valores foram escolhidos arbitrariamente devido ao algoritmo ter encontrado um número razoável de regras (165), o que possibilita uma avaliação manual. Dessas 165 regras, 31 eram redundantes (regras redundantes são regras que predizem tão bem quanto ou pior do que outra regra mais geral) e foram retiradas.

Os resultados foram:
- 25 regras de 2 itens
- 53 regras de 3 itens
- 40 regras de 4 itens
- 13 regras de 5 itens
- 3 regras de 6 itens

Medidas de interesse:

Suporte:

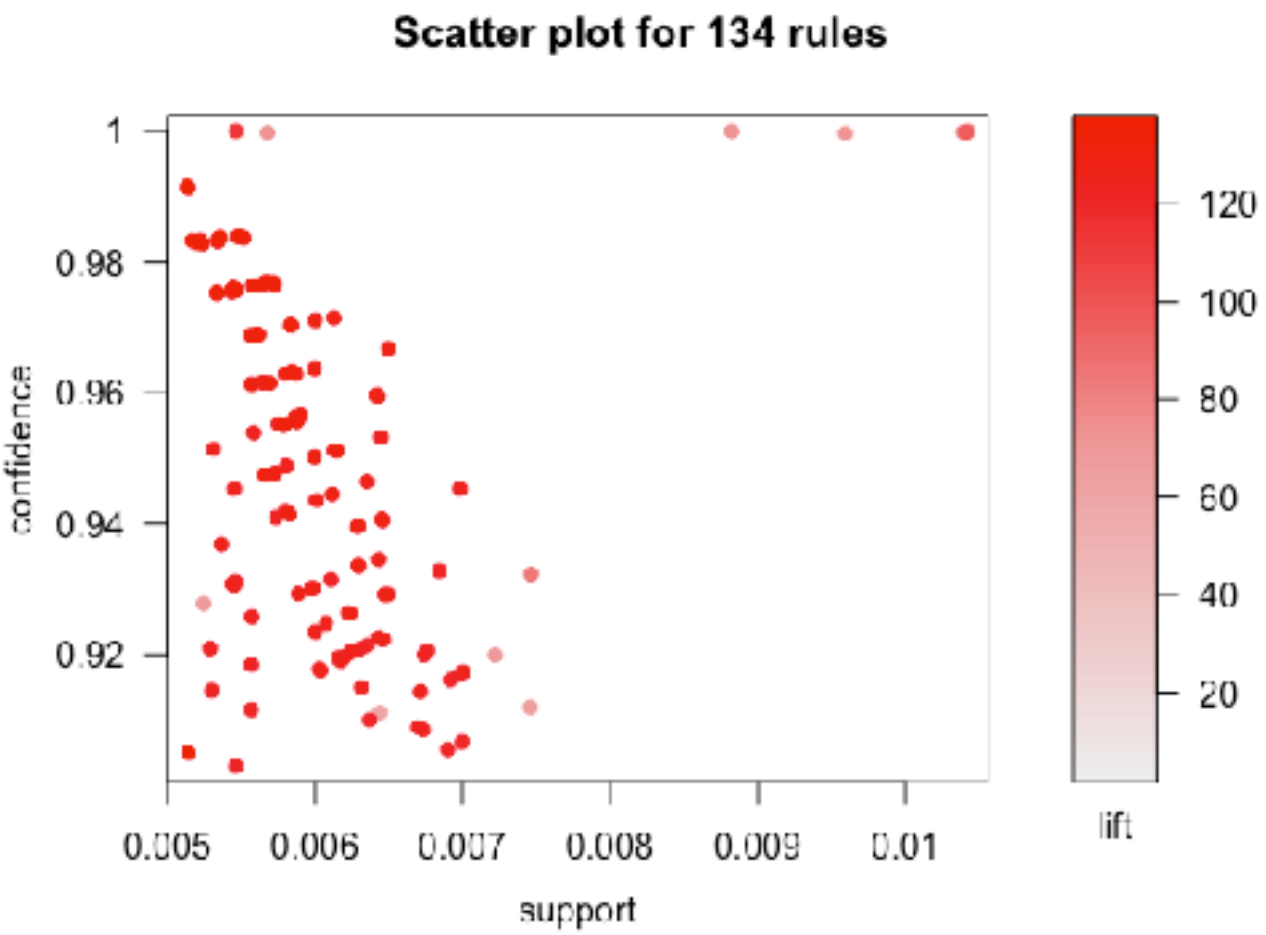
Mínimo = 5,137%, Médio = 6,148%, Máximo = 10,410%

Confiança:

Mínimo = 90,30%, Médio = 94,84%, Máximo = 100%

Lift:

Mínimo = 58.59, Médio = 121.46, Máximo = 137.52



Contudo, ao analizar as regras manualmente, percebe-se claramente que produtos “Herb Marker” são muito relacionados entre si. Ou seja, quem compra o produto “Herb Marker Basil”, também compra “Herb Marker Mint”, por exemplo. Agrupei todos os produtos “Herb Marker” em um só e rodei novamente o algoritmo para retirar todas as regras relacionadas aos produtos “Herb Marker”, uma vez que já está claro que quem compra um deles, tende a comprar os demais.
Após rodar o algoritmo novamente, exatamente da mesma maneira, mas com os produtos agrupados, foram geradas somente 13 regras com suporte maior que 5%. Dessas 13 regras, temos:

- 8 regras com 2 itens,
- 5 regras com 3 itens.

Medidas de interesse:

Confiança:

Mínimo = 5,227%, Médio = 8,080%, Máximo = 10,410%,

Confiança:

Mínimo = 91,08%, Médio = 96,95%, Máximo = 100%,

Lift:

Mínimo = 58.59, Médio = 77.31, Máximo = 113.22,

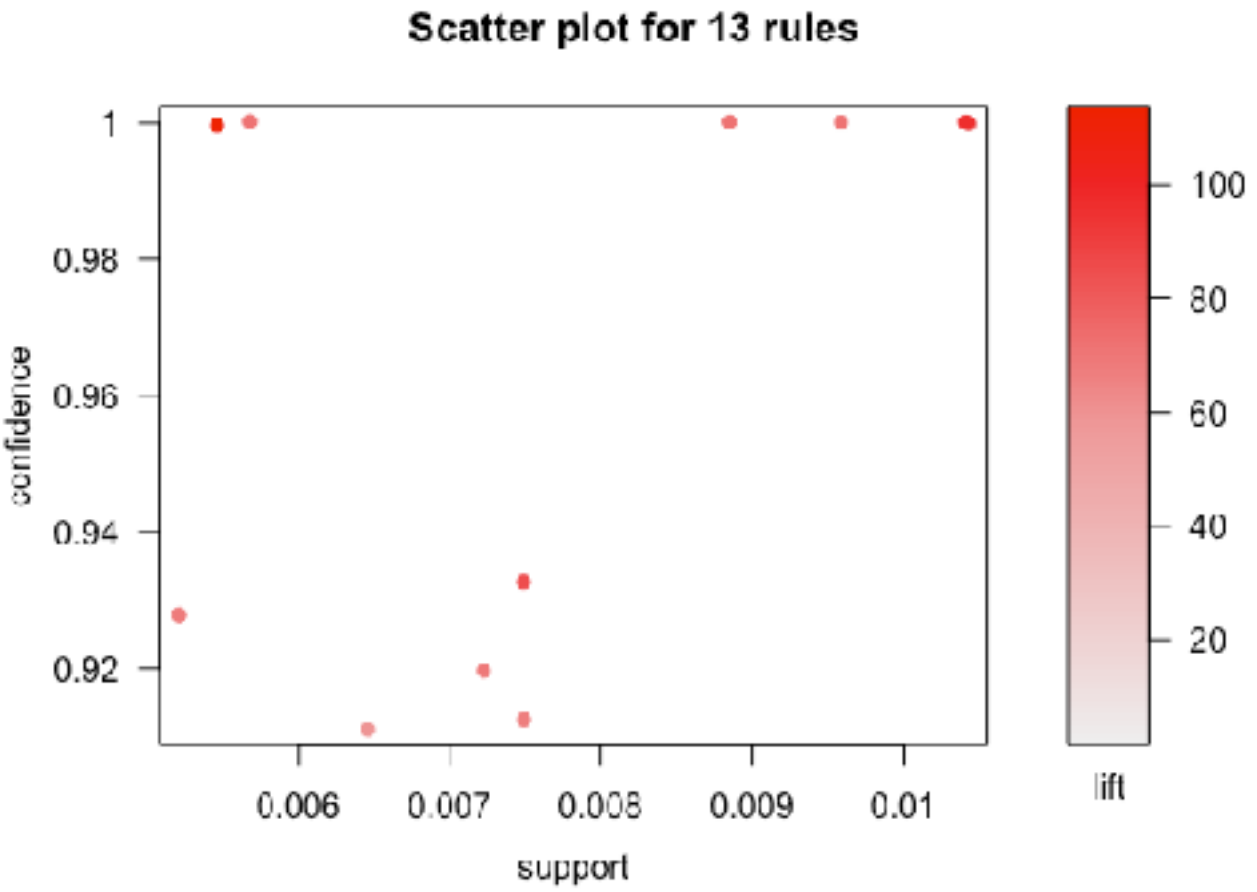


Tabela com as regras finais:

No	Quem Compra	Também Compra	Support	Confidence	Lift
1	Front Door	Key Fob	0.00567	1	71.58
2	Hot Pink	Feather Pen	0.00545	1	113.21
3	Sugar	Set of 3 Retrospot Tea	0.01040	1	96.06
4	Set of 3 Retrospot Tea	Sugar	0.01040	1	96.06
5	Sugar	Coffee	0.01040	1	69.34
6	Set of 3 Retrospot Tea	Coffee	0.01040	1	69.34
7	Back Door	Key Fob	0.00883	1	71.58
8	Shed	Key Fob	0.00959	1	71.58
9	Regency Tea Plate Pink, Regency Tea Plate Roses	Regency Tea Plate Green	0.00748	0.9325	84.12
10	Regency Tea Plate Green, Roses Regency Teacup and Saucer	Regency Tea Plate Roses	0.00522	0.9280	67.74
11	Set of 20 Red Retrospot Paper Napkins, set of 6 Red Spotty Paper Cups	Set of 6 Red Spotty Paper Plates	0.00721	0.9195	69.17
12	Regency Tea Plate Green, Regency Tea Plate PInk	Regency Tea Plate Roses	0.00748	0.9120	66.57
13	Wooden Heart Christmas Scandinavian, Wooden Tree Christmas Scandinavian	Wooden Star Christmas Scandinavian	0.00644	0.9108	58.58

Devido à escassez de regras, se a análise estivesse sendo feita com propósitos práticos, para o aumento de vendas de uma empresa, então seria vantajoso abaixar o suporte do algoritmo “apriori” para haver uma maior variedade de formas de distribuir os produtos ou colocá-los em promoção. Contudo, o baixo número de regras foi conveniente para a apresentação dos resultados, e, portando, deixarei como está.

5. Avaliação da Ferramenta Lemonade:

5.1 - Preparação dos Dados:

A parte da limpeza dos dados da ferramenta não é feita de forma intuitiva para o usuário, o que compromete demais o uso.

As funções são colocadas na aba de seleção de “caixas”, sem nenhuma explicação de como é a entrada e como é a saída. Darei um exemplo de uma situação real que precisava transformar os dados e não consegui usando a ferramenta:
Precisava transformar os dados da forma com que o dataset foi disponibilizado para a forma que os algoritmos de regras de associação precisam. Ou seja, precisava retirar todas as colunas do dataset exceto “InvoiceNo” e “Description”. Para isso é necessário somente usar a função “Projection/Select Columns” no Lemonade. Após fazer isso, contudo, seria necessário agrupar todas as linha com mesmo “InvoiceNo” e as colidir em uma só, separando elementos com descrições diferentes por vírgulas. Para isso pensei em usar a função “Transformation”, contudo, a descrição dentro dela na ferramenta não existe. Seguindo pelo tutorial disponibilizado pelo monitor Felipe, descobri qual parte da documentação do Spark tem as funções disponíveis, contudo, não consegui pensar em como usar nenhuma delas de forma a fazer o que eu precisava.

5.2 - Uso dos Algoritmos de Padrões Frequentes:

O uso dos algoritmos de padrões frequentes no Lemonade, com a ajuda de um simples tutorial se torna simples e útil, de várias formas. Contudo, obtive um problema ao tentar colocar os dados já preparados na ferramenta.

Após a transformação dos dados na forma de transações, dei upload da base na ferramenta e ela me retornou que não são permitidos elementos repetidos na mesma linha.

Decidi então remover qualquer duplicata em uma transação usando R. Para isso, usei o seguinte método:

```
dataset <- dataset[!(duplicated(dataset)),]
any(duplicated(dataset))
write.csv(dataset, "no_duplicates.csv", quote = FALSE, row.names = FALSE)
```

A saída do método usado me afirmou que não haviam mais duplicatas no dataset, então, dei upload novamente dos dados na ferramenta que me retornou o mesmo erro. Não sei o que está causando isso.

6. Conclusão:

A partir do trabalho é possível perceber alguns detalhes importantes.

A maior parte do trabalho está na parte de limpeza e análise de dados. Os algoritmos de mineração já estão implementados e são simples de serem usados.

Também percebemos que apesar de toda a limpeza feita nos dados, ainda aconteceu o caso de uma regra inútil, uma vez que as regras 9 e 12 (mostradas na tabela de regras finais) podem ser resumidas em uma só da mesma forma que foi feita com os produtos “Herb Marker”.

Sobre a análise final de resultados, percebemos que, na base de dados, o conjunto frequente {“Açúcar”, “Set of 3 Retrospot Tea”} tem confiança 100% a partir da análise das regras. O fato de termos rodados um algoritmo para mineração de regras de associação e termos conseguido derivar um itemset frequente mostra a similaridade entre os algoritmos. Em questões práticas, é possível dizer que fazer uma promoção abaixando o preço de qualquer item da coluna “Quem compra” é possivelmente interessante, uma vez que abaixar o preço desses produtos aumentará a compra deles e provavelmente de todos os produtos que são associados a ele.