Instituto Superior de Engenharia de Lisboa

AMD - Aprendizagem e Mineração de Dados

Mestrado em Engenharia Informática e Multimédia

Association Rules

Projeto Final – B

Dezembro 2023

Grupo 4
Gonçalo Silva — 47255
João Rocha — 47196
Luís Morgado — 51358

Índice

1 Analyze the Dataset

A partir do report disponibilizado pela "We-Commerce"e o dataset, podemos analisar cada um dos 13 atributos e seu significado:

tracking_record_id: É o identificador único para cada registro de evento. Garante que cada evento (ou visita) seja distintamente identificado e possa ser referenciado individualmente.

date_time: Representa o espaço temporal de quando o evento ocorreu. Este atributo é crucial para compreender o momento das visitas, analisar padrões como as horas de pico e organizar a atividade do utilizador ao longo do tempo.

user_gui: "User - Global Unique Identifier- Identifica de forma única um utilizador subscrito. Se o campo estiver vazio, significa que o visitante não é um utilizador subscrito. Este atributo é fundamental para rastrear o comportamento individual do utilizador. Permite a análise do comportamento do utilizador ao longo de várias sessões e plataformas, fornecendo uma visão abrangente da interação do utilizador com o site, independentemente do dispositivo ou browser utilizado.

campaign_id: Identifica uma campanha promocional associada ao evento. É essencial para rastrear a eficácia das campanhas de marketing, entender quais campanhas atraem mais visitantes e correlacionar campanhas com ações específicas do utilizador ou interesses de produtos.

product_gui: "Product - Global Unique Identifier- Identifica de forma única um produto que um visitante visualizou durante um evento. Este atributo é crucial para entender quais produtos estão a atrair mais interesse, rastrear visualizações de produtos e analisar tendências de compra. É necessário filtrar este atributo, pois pode conter informação irrelevante, que não representam produtos reais e podem distorcer a análise dos dados.

company: O nome da empresa que fornece o produto. Esta informação é útil para entender quais produtos de empresas são mais populares e analisar o interesse do produto por empresa.

link: O URL da página web visitada. Este atributo ajuda a entender quais páginas específicas estão a atrair tráfego, a jornada do utilizador através do website.

tracking_id, meio: Estes atributos parecem ser identificadores de rastreamento adicionais, potencialmente usados para fins de rastreamento interno ou para uma análise mais granular do comportamento do visitante e das fontes de tráfego. Parecem não conter informação relevante.

ip: O endereço de Protocolo de Internet do visitante. Embora a alocação dinâmica de IPs possa tornar isto menos útil para rastreamento a longo

prazo, é valioso para análise de geolocalização, identificação de potenciais ataques cibernéticos e compreensão da distribuição geográfica dos visitantes.

browser: Contém informações sobre o navegador do visitante, como tipo e versão. Esta informação é importante para garantir a compatibilidade do site com navegadores populares e diferentes dispositivos, bem como para entender as preferências dos utilizadores.

session_id: Este é um identificador único para a sessão de um visitante. Uma vez que um utilizador pode ter várias sessões, este atributo é crítico para rastrear visitas individuais e entender o envolvimento do utilizador ao longo do tempo.

referer: O URL de referência, indicando de onde o visitante veio antes de entrar na página atual. Isto é útil para entender as fontes de tráfego, a eficácia de links externos e os padrões de navegação do utilizador.

cookie_id: O identificador único armazenado no cookie do visitante. Este atributo é central para rastrear o comportamento individual do visitante ao longo de várias sessões, especialmente para utilizadores não subscritos.

2 Market-basket Analysis Problem

2.1 Objetivo:

The market-basket analysis goal is to, "find groups of items that tend to occur together in transactions".

2.2 Formulação do Problema:

1. **Preparação dos Dados:** O primeiro passo envolve a preparação do dataset, a limpeza dos dados e a normalização para garantir a qualidade e consistência dos dados. Cada transação (ou sessão de visitante) será representada por um *session_id* único, e dentro de cada transação, os produtos visualizados ou com os quais interagiu (identificados por *product_gui*) serão os itens de interesse. Isto forma o 'market-basket' para cada sessão.

2. **Identificação de Combinações de Produtos:** Analisar as combinações de produtos que ocorrem frequentemente juntas na mesma sessão. Isto envolve olhar para os diferentes valores de *product_gui* que são agrupados sob o mesmo *session_id*.

3. **Determinar as Association Rules:** O próximo passo é estabelecer regras de associação. Estas são regras que ajudarão a identificar se a presença de um certo produto (ou conjunto de produtos) num cesto implica a probabilidade de outro produto estar no mesmo cesto. Por exemplo, se

o produto A e o produto B são frequentemente vistos juntos, então existe uma regra: Se A, então B.

4. **Métricas para Avaliação de Regras:** Utilizar métricas como suporte, confiança e elevação (*lift*) para avaliar estas regras. O suporte mede a frequência com que um produto ou combinação de produtos aparece no conjunto de dados, a confiança mede a frequência com que as regras são consideradas verdadeiras, e a elevação mede a razão do suporte observado em relação ao esperado se os dois conjuntos fossem independentes.

5. **Análise dos Resultados:** Interpretar os resultados para entender a força e relevância destas regras de associação. Por exemplo, descobrir que dois produtos são frequentemente comprados juntos pode informar estratégias de venda cruzada ou campanhas promocionais.

6. **Valor Empresarial:** Converter estas descobertas em informação de negócios valiosa. Por exemplo, se certos produtos são frequentemente comprados juntos, eles podem ser colocados mais próximos numa loja física ou sugeridos como recomendações num mercado online.

3 Data Distribution

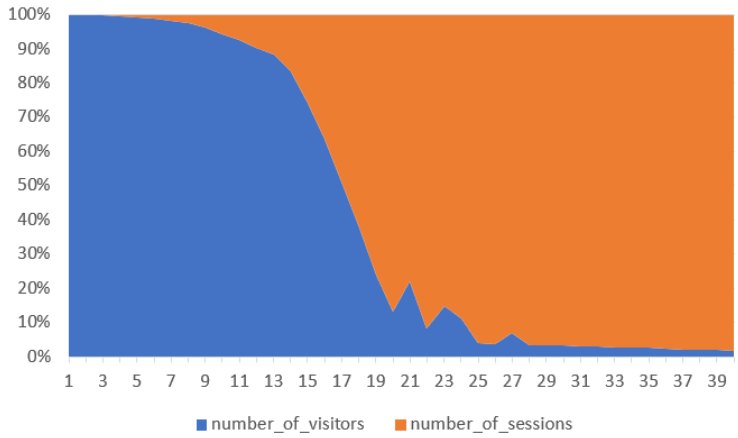
- O número total de eventos (cada linha tuplos representa um evento);

```
total_events
--------------
    415863
```

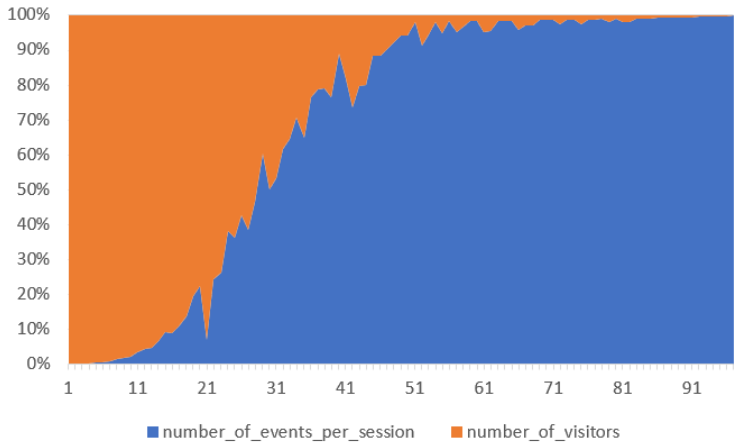
- O número total de visitantes (cookies diferentes);

```
total_number_of_cookies
------------------------
                263137
```

- A distribuição de visitantes e sessões (number-of-visitors Vs number-of-sessions);



- A distribuição de sessões e visitantes (number-of-events-per-session Vs number-of-visitors).



4 Reduce the Complexity

Para simplificar os dados no dataset e reduzir a sua complexidade, focamos a nossa análise no subconjunto que contém todos os eventos gerados por visitantes com um número de sessões dentro de um determinado intervalo de 5 e 30. Esta abordagem permite concentrar nos dados mais relevantes para o estudo dos comportamentos dos clientes, ao mesmo tempo que diminui o tempo de computação necessário para a análise, considerando os dados disponibilizados e a sua distribuição.
O dataset também foram filtradas outras informações irrelevantes do atributo **product_gui** nomeadamente páginas de login e de configuração da conta do

cliente, página principal e página do carrinho de compras.

5 Generate a Dataset file

O ficheiro com o Dataset filtrado foi exportado utilizando uma view (v_export) para o formato CSV com os atributos: **cookie_id** e **product_gui**.

```
CREATE VIEW v_export (cookie_id, session_id, product_gui)
AS
SELECT T1.cookie_id, T2.session_id, T2.product_gui
FROM
    (SELECT *
     FROM v_cookie_number_of_sessions
     WHERE number_of_sessions >= 5 AND number_of_sessions <= 30) AS T1
INNER JOIN
    track AS T2
ON T1.cookie_id = T2.cookie_id
WHERE T2.product_gui NOT IN ('open', 'home', '/customer/account/login/', '/customer/account/
AND T2.product_gui NOT LIKE '%/order_id/%' AND T2.product_gui NOT LIKE '%/account/%'
ORDER BY T1.cookie_id, T2.session_id, T2.product_gui;
```

5.1 Normalize and Prepare Data

1. **Substituição de Símbolos:** Substitui o símbolo "="por "+". O sinal de igual é substituído porque é usado no formato de "market-basket"no Orange3, e a substituição evita confusões.

2. **Eliminação de Espaços:** Remove todos os espaços em branco da string.

3. **Eliminação de Aspas:** Substitui aspas duplas (") por um símbolo de dólar ($).

4. **Remoção de Acentos:** Utiliza a função remove_accents para eliminar caracteres acentuados da string. A função está configurada para usar a codificação "iso-8859-1".

5. **Conversão para Minúsculas:** Converte todos os caracteres da string para letras minúsculas.

6. **Conversão para Basket:** O ficheiro CSV é convertido num formato de basket, adequado para análises de associação. Para cada transação identificada pelo seu ID, agrupam-se os produtos relacionados, contabilizando a frequência de cada produto dentro da mesma transação.

6 Rules and Marketing Decisions

Analisando os dados, podemos propor várias decisões de marketing:

```
Support: 0.05, Confidence: 0.8, Number of Rules: 12
--> Rule: ['lon_4004', 'lon_2125', 'lon_4508'] -> ['lon_4504'], Support: 67, Confidence: 0.9
--> Rule: ['lon_4004', 'lon_4504', 'lon_4508'] -> ['lon_2125'], Support: 67, Confidence: 0.8
--> Rule: ['lon_4004', 'lon_4504', 'lon_2125'] -> ['lon_4508'], Support: 67, Confidence: 0.9
--> Rule: ['pumpseopentoes', 'botas'] -> ['botins'], Support: 65, Confidence: 0.915492957746
--> Rule: ['lon_4504', 'lon_4508'] -> ['lon_2125'], Support: 84, Confidence: 0.84
--> Rule: ['lon_4504', 'lon_2125'] -> ['lon_4508'], Support: 84, Confidence: 0.875
--> Rule: ['lon_4004', 'lon_2125'] -> ['lon_4504'], Support: 73, Confidence: 0.8390804597701
--> Rule: ['lon_4004', 'lon_4504'] -> ['lon_2125'], Support: 73, Confidence: 0.8111111111111
--> Rule: ['lon_4504', 'lon_4508'] -> ['lon_4004'], Support: 80, Confidence: 0.8
--> Rule: ['lon_4004', 'lon_4508'] -> ['lon_4504'], Support: 80, Confidence: 0.8421052631578
--> Rule: ['lon_4004', 'lon_4504'] -> ['lon_4508'], Support: 80, Confidence: 0.8888888888888
--> Rule: ['lon_4004', 'lon_2125'] -> ['lon_4508'], Support: 74, Confidence: 0.8505747126436

Support: 0.07, Confidence: 0.7, Number of Rules: 7
--> Rule: ['lon_4508'] -> ['lon_2125'], Support: 108, Confidence: 0.7012987012987013
--> Rule: ['lon_2125'] -> ['lon_4508'], Support: 108, Confidence: 0.7346938775510204
--> Rule: ['pumpseopentoes'] -> ['botins'], Support: 96, Confidence: 0.7164179104477612
--> Rule: ['lon_4504'] -> ['lon_2125'], Support: 96, Confidence: 0.732824427480916
--> Rule: ['lon_4504'] -> ['lon_4508'], Support: 100, Confidence: 0.7633587786259542
--> Rule: ['lon_4004'] -> ['lon_4504'], Support: 90, Confidence: 0.7086614173228346
--> Rule: ['lon_4004'] -> ['lon_4508'], Support: 95, Confidence: 0.7480314960629921
```

6.1 Produtos para Manter Juntos em Uma Página Web

Os conjuntos de produtos `['lon_4004', 'lon_2125', 'lon_4508'] -> ['lon_4504']` e `['pumpseopentoes', 'botas'] -> ['botins']` apresentam alta confiança, indicando que quando os produtos do lado esquerdo são visualizados, há uma alta probabilidade de que o produto do lado direito também seja. Estes produtos devem ser destacados juntos para aumentar a chance de compra cruzada.

6.2 Produtos Raramente Visitados, Mas Frequentemente Visitados Juntos

Produtos com suportes mais baixos, mas com altas confianças, como em algumas das regras com suportes de 0.04 e 0.05, podem representar oportunidades para promoções direcionadas ou compras conjuntas.

```
['lon_4004', 'lon_2125', 'lon_4508'] -> ['lon_4504'], Support: 67, Confidence: 0.90
['lon_4004', 'lon_4504', 'lon_4508'] -> ['lon_2125'], Support: 67, Confidence: 0.83
['lon_4004', 'lon_4504', 'lon_2125'] -> ['lon_4508'], Support: 67, Confidence: 0.91
```

6.3 Produtos Mais Visitados

Produtos que aparecem frequentemente como parte de regras com suporte mais alto (como 0.07 ou 0.08), como lon_4508, lon_2125, e lon_4504, são candidatos a serem os mais visitados e podem ser usados para atrair visitantes para outras ofertas relacionadas.

Estes conhecimentos podem ajudar na formulação de estratégias eficazes de marketing e colocação de produtos, contribuindo para aumentar as vendas e melhorar a experiência dos clientes.