UNIVERSIDADE DO ESTADO DO AMAZONAS (UEA)
ESCOLA SUPERIOR DE TECNOLOGIA (EST)
CURSO DE PÓS-GRADUAÇÃO EM CIÊNCIAS DE DADOS

AQUISIÇÃO, PRÉ-PROCESSAMENTO E EXPLORAÇÃO DE DADOS – PROJETO FINAL

Professor Drº Luis Cuevas Rodriguez

Alexandre Teixeira da Silva (atds.cid25@uea.edu.br)
César Braz de Oliveira (cbdo.cid25@uea.edu.br)
Ícaro Guimarães Canto (igc.cid25@uea.edu.br)
Priscila Leylianne da Silva Gonçalves (pldsg.cid25@uea.edu.br)

Manaus/AM

2025

Análise de Recompra em Comércio Eletrônico: Uma Abordagem de Aquisição e Pré-Processamento de Dados no dataset Instacart

Dataset Utilizado: Instacart Market Basket Analysis (Disponível no Kaggle: https://www.kaggle.com/datasets/psparks/instacart-market-basket-analysis/data)

1. *Definição do Problema e das Perguntas Analíticas*

Área de Negócio Escolhida: Comércio Eletrônico de Alimentos (E-commerce de Supermercados / Mercearia Online).

Problema de Negócio: Como prever os produtos que um cliente irá recomprar em seu próximo pedido, com base em seu histórico de compras?

Perguntas Analíticas Relevantes (5 perguntas):

1. Quais produtos apresentam maior taxa de recompra entre os clientes?
    - Objetivo de Negócio: Identificar os "carros-chefes" da recorrência, permitindo ao Instacart priorizar a disponibilidade em estoque, otimizar a cadeia de suprimentos e direcionar campanhas de marketing para itens que comprovadamente fidelizam o cliente.
2. Quais horários e dias da semana concentram a maior quantidade de reorders?
    - Objetivo de Negócio: Otimizar as operações logísticas (rotas de entrega, alocação de shoppers), dimensionar a equipe de atendimento ao cliente e planejar o lançamento de promoções e notificações em momentos de maior propensão à compra.
3. Existe um padrão no número de dias entre os pedidos de um mesmo cliente?
    - Objetivo de Negócio: Segmentar a base de clientes para desenvolver estratégias de retenção personalizadas, como programas de fidelidade para clientes fiéis e ações de reengajamento para clientes com menor frequência ou maior risco de churn.
4. Quais departamentos ou corredores concentram os produtos mais frequentemente reordenados?

- Objetivo de Negócio: Entender o processo de decisão de compra do cliente. Isso pode informar a otimização da interface do usuário (UI/UX) no aplicativo, a forma como os produtos são exibidos em listas de favoritos ou sugestões, e estratégias de cross-selling ou up-selling baseadas na prioridade do cliente.

5. Como o comportamento de recompra varia entre diferentes perfis de cliente (ex: frequência de pedidos, número de itens por carrinho, horário habitual de compra)?

    - Objetivo de Negócio: Avaliar o impacto da diversificação das compras na retenção. Se clientes que experimentam mais tendem a ser mais leais, o Instacart poderia incentivar a descoberta de novos produtos. Se não, a estratégia pode ser focar na recompra dos produtos já conhecidos e amados.

2. *Coleta de Dados*

Fonte de Dados: O conjunto de dados "Instacart Market Basket Analysis" foi obtido do Kaggle (link fornecido). Esta é uma fonte pública amplamente utilizada para análises de cestas de compras, contendo dados anonimizados de mais de 3 milhões de pedidos de 200.000 usuários Instacart.

Arquivos do Dataset e Sua Relevância:

- aisles.csv: Contém aisle_id e aisle (nome do corredor/seção). Relevância: Essencial para classificar produtos em categorias mais granulares, útil para a Pergunta 1.
- departments.csv: Contém department_id e department (nome do departamento). Relevância: Permite uma classificação de produto mais ampla, também crucial para a Pergunta 1.
- products.csv: Contém product_id, product_name, aisle_id, department_id. Relevância: O coração dos dados de produto, permite vincular produtos aos seus nomes e categorias, fundamental para as Perguntas 1, 4 e 5.
- orders.csv: Contém order_id, user_id, eval_set (conjunto de dados para ML: treino, teste, prior), order_number (número do pedido do usuário), order_dow (dia da semana), order_hour_of_day (hora do dia), days_since_prior_order (dias desde o pedido anterior). Relevância: Fornece o contexto de tempo e sequência dos pedidos, crucial para as Perguntas 2, 3 e 5. O user_id é chave para rastrear o cliente.
- order_products__prior.csv: Contém order_id, product_id, add_to_cart_order (ordem de adição ao carrinho), reordered (se o produto foi recomprado neste pedido). Relevância: Detalhes dos produtos em pedidos anteriores ao conjunto de treino/teste, fundamental para todas as análises de recompra (Perguntas 1, 2, 3, 4, 5).

- order_products__train.csv: Contém as mesmas colunas que order_products__prior.csv, mas para o conjunto de treino. Relevância: Também essencial para as análises de recompra. Para análises descritivas, ambos os arquivos prior e train serão concatenados.

Avaliação da Adequação dos Dados: Os dados são altamente adequados para todas as 5 perguntas analíticas formuladas, pois fornecem as informações necessárias para:

- Identificação de Recompra (reordered): Permite quantificar a lealdade a produtos.
- Contexto Temporal (order_dow, order_hour_of_day, days_since_prior_order): Essencial para padrões de sazonalidade e frequência.
- Categorização de Produtos (aisle, department): Permite análises em um nível mais agregado que o produto.
- Sequência de Itens (add_to_cart_order): Direcciona a análise sobre a prioridade do cliente.
- Histórico do Usuário (user_id, order_number): Permite acompanhar o comportamento de cada cliente e identificar novos produtos comprados.

Limitações e Considerações:

- Anonimato do Cliente: A ausência de informações demográficas (idade, gênero, localização geográfica exata, renda) impede uma segmentação de clientes mais rica e a identificação de correlações com fatores socioeconômicos.
- Período de Tempo: Embora o order_number e days_since_prior_order deem uma sequência, a falta de datas absolutas pode dificultar a análise de tendências sazonais de longo prazo ou o impacto de eventos externos específicos no calendário.
- Ausência de preços dos produtos: O dataset não contém dados de preços. Esta é uma limitação significativa para análises de valor do carrinho ou correlações de preço com a recompra. Para a análise proposta (P2), teríamos que simular dados de preço ou adaptar a pergunta, o que é uma solução subótima em um cenário real. No contexto deste projeto, faremos uma simulação justificada para demonstrar a técnica.
- Informação agregada: Não há detalhes sobre o comportamento do cliente fora da plataforma Instacart, como compras em outros locais ou motivos de abandono.

3. *Pré-processamento dos Dados*

Esta etapa visa transformar os dados brutos em um formato limpo, consistente e adequado para a análise exploratória. Será realizada majoritariamente utilizando a biblioteca pandas em Python.

Etapas Adotadas e Justificativas Detalhadas:

1. Carregamento e Visão Geral Inicial:
    - Ação: Carregar cada arquivo CSV para um DataFrame Pandas e realizar um .info(), .head(), e .describe() para entender a estrutura, tipos de dados e estatísticas básicas.
    - Justificativa: Compreender a estrutura de cada arquivo e identificar rapidamente possíveis problemas (tipos incorretos, valores ausentes, outliers).
2. Combinação dos Dados (Merge e Concat):
    - Ação:
        - Concatenar order_products__prior.csv e order_products__train.csv em um único DataFrame order_products_combined.
        - Realizar merge entre order_products_combined e orders (usando order_id).
        - Realizar merge do resultado com products (usando product_id).
        - Realizar merge do resultado com aisles (usando aisle_id) e departments (usando department_id).
        - Serão utilizadas junções do tipo inner ou left conforme a necessidade de manter todas as informações. Um left merge começando pelo order_products_combined garantirá que todos os itens de pedido sejam mantidos e enriquecidos.
    - Justificativa: As informações estão distribuídas em várias tabelas normalizadas. Juntar essas tabelas é essencial para criar um dataset denso onde cada linha representa um item específico dentro de um pedido, com todas as informações contextuais (cliente, tempo, produto, categoria, recompra). Isso facilita análises de agrupamento e agregação.
3. Tratamento de Valores Ausentes (NaN):

- Ação: A coluna days_since_prior_order em orders.csv (e, consequentemente, no DataFrame combinado) conterá NaN para o primeiro pedido de cada user_id (order_number == 1).
    - Para análises de frequência ou intervalo, esses NaN são informativos e não devem ser simplesmente preenchidos com 0, pois 0 dias significaria "no mesmo dia".
    - Para análises onde um valor numérico é estritamente necessário (ex: cálculo de média para todos os pedidos), poderíamos considerar preencher com a média/mediana para pedidos > 1 ou criar uma flag is_first_order. Para as perguntas propostas, o NaN é interpretado como "não aplicável" (primeiro pedido).
- Justificativa: O tratamento adequado de NaN evita erros em cálculos e garante que a semântica dos dados seja mantida. A decisão de não preencher days_since_prior_order com 0 reflete sua natureza como "dias desde o pedido anterior".

4. ***Conversão de Tipos de Dados para Otimização e Análise:***
    - Ação:
        - Converter order_dow, order_hour_of_day, reordered (para booleanos ou inteiros 0/1), aisle, department, eval_set para o tipo category do Pandas.
        - Converter order_number, add_to_cart_order para int (se já não forem).
        - Manter product_id, user_id, order_id como int ou convertê-los para o tipo mais eficiente se necessário (np.int32).
    - Justificativa:
        - Otimização de Memória: O tipo category consome significativamente menos memória para colunas com valores repetidos (como nomes de departamentos), crucial para datasets grandes.
        - Desempenho: Operações de agrupamento (groupby) são mais rápidas em colunas do tipo category.
        - Consistência: Garante que os dados sejam interpretados corretamente por funções de agregação e visualização.
5. ***Criação de Novas Features (Engenharia de Features):***

- order_type (Para Pergunta 2):
    - Ação: Criar uma coluna booleana is_first_order (orders.order_number == 1) e/ou uma coluna categórica order_type ('Primeira Compra' vs. 'Recompra').
    - Justificativa: Fundamental para comparar o comportamento de novos clientes com o de clientes recorrentes.
- product_reordered_status (Para Pergunta 1):
    - Ação: Renomear reordered para product_reordered_status para maior clareza, indicando se aquele item específico foi recomprado no contexto daquele pedido.
    - Justificativa: Clareza semântica.
- order_size_category (Para Pergunta 3):
    - Ação: Calcular o número de itens por pedido (count(product_id) group by order_id) e então categorizar esses totais em faixas (ex: 'Pequeno' < 5 itens, 'Médio' 5-15 itens, 'Grande' > 15 itens).
    - Justificativa: Permite analisar se o tamanho do carrinho se correlaciona com a taxa de recompra do cliente.
- add_to_cart_order_group (Para Pergunta 4):
    - Ação: Agrupar os valores de add_to_cart_order em faixas categóricas (ex: '1º Item', '2º-5º Item', '6º-10º Item', '> 10º Item').
    - Justificativa: Facilita a visualização e análise do impacto da posição de adição no carrinho, que pode ter uma distribuição muito ampla.
- has_new_product_in_order e percent_new_products (Para Pergunta 5):
    - Ação: Para cada user_id e order_id, identificar quais produtos naquele pedido são "novos" para o usuário (i.e., nunca apareceram em pedidos anteriores daquele usuário). Calcular a proporção de produtos novos no pedido.
    - Justificativa: Permite analisar se a propensão do cliente a experimentar novos produtos em um dado pedido se correlaciona com sua lealdade ou taxa de recompra geral.

- Simulação de price_per_unit (Para a pergunta 2, dada a limitação do dataset):
    - Ação: Criar uma coluna price_per_unit simulada, atribuindo preços aleatórios por produto, talvez com faixas de preço distintas por categoria (ex: eletrônicos mais caros que alimentos).
    - Justificativa: Como mencionado nas limitações, o dataset não possui preços. Esta simulação é uma abordagem para demonstrar a metodologia de análise de correlação entre preço e recompra, reconhecendo que em um projeto real, dados de preço seriam cruciais e reais. A randomização deve ser controlada para evitar valores irrealistas e para que a hipótese de correlação inversa (menor preço -> maior recompra) possa ser demonstrada.

6. ***Verificação Final:***
    - Ação: Realizar uma última verificação .info(), .head(), describe() no DataFrame final e verificar a contagem de valores únicos para as novas colunas categóricas.
    - Justificativa: Assegurar que o pré-processamento foi concluído com sucesso e que o DataFrame está pronto para a análise exploratória.

7. ***Exploração dos Dados (Análise Exploratória)***

Esta etapa é o coração do projeto, onde geramos insights a partir dos dados pré-processados, utilizando estatísticas descritivas e visualizações. O foco é responder às 5 perguntas analíticas.

Ferramentas: Python (Pandas para manipulação, Matplotlib e Seaborn para visualização).

Passos Detalhados e Visualizações para Cada Pergunta Analítica:

1. Quais produtos e categorias/departamentos apresentam as maiores taxas de recompra?
    - Análise:
        - Taxa de Recompra por Produto: Agrupar o DataFrame combinado por product_name, calcular a soma de product_reordered_status e a contagem total de vendas para obter a taxa de recompra percentual de cada produto. Selecionar os TOP N (ex: 20) produtos.
        - Taxa de Recompra por Categoria/Departamento: Agrupar por department e aisle, calcular a taxa de recompra média para cada um.
    - Visualização:

- Gráfico de Barras Horizontais: "Top N Produtos por Taxa de Recompra".
- Gráfico de Barras Horizontais: "Taxa de Recompra Média por Departamento" e "Taxa de Recompra Média por Corredor (Aisle)".

- Insights Esperados: Produtos básicos e de consumo frequente (frutas, laticínios, ovos) e suas respectivas categorias/departamentos devem liderar as taxas de recompra, indicando sua importância na cesta de compras recorrentes.

2. Qual a sazonalidade e os horários de pico para as recompras? Existem diferenças significativas entre a primeira compra e as recompras subsequentes?
   - Análise:
     - Padrões de Recompra por Hora e Dia da Semana: Agrupar por order_dow e order_hour_of_day e contar o número de recompras. Comparar com o total de pedidos para entender a proporção.
     - Distribuição de days_since_prior_order: Analisar a distribuição dos intervalos entre os pedidos para identificar a frequência típica.
     - Comparação is_first_order vs. reorder: Agrupar por order_dow e order_hour_of_day, e filtrar para is_first_order = True e False separadamente, plotando suas distribuições.
   - Visualização:
     - Heatmap: Total de Reordens por Dia da Semana e Hora do Dia.
     - Histograms: Distribuição de days_since_prior_order.
     - Gráficos de Linha/Barra Duplos: Número de Pedidos por Hora do Dia (primeira compra vs. recompra) e Número de Pedidos por Dia da Semana (primeira compra vs. recompra).
   - Insights Esperados: Picos de atividade podem ser observados em dias úteis durante o horário comercial ou à noite, e nos fins de semana. Recompras podem ter um padrão de frequência (ex: semanal ou quinzenal) distinto das primeiras compras.
3. Existe um perfil de cliente distinto (definido por frequência de pedidos, tamanho do carrinho e/ou dias desde o pedido anterior) que demonstra maior lealdade e taxa de recompra?

- Análise:
    - Métricas por Usuário: Agrupar por user_id e calcular: total_orders, avg_items_per_order, avg_days_since_prior_order, e a user_reorder_rate (média de product_reordered_status para todos os itens do usuário).
    - Segmentação: Criar perfil_recompra (Super Fiel, Fiel, Regular, Ocasional) e frequencia_categoria (Diário, Semanal, Esporádico) com base nessas métricas agregadas por usuário.
    - Comparação de Perfis: Calcular as métricas médias (avg_items_per_order, total_orders etc.) para cada perfil_recompra ou frequencia_categoria.
- Visualização:
    - Múltiplos Gráficos de Barras: Comparando Média de Itens por Pedido, Média de Pedidos Totais e Média de Dias Desde Pedido Anterior para cada perfil_recompra e frequencia_categoria.
    - Box Plot/Violin Plot: Distribuição de user_reorder_rate por frequencia_categoria.
- Insights Esperados: Clientes com maior frequência de pedidos e maior média de itens por carrinho podem exibir taxas de recompra mais elevadas, definindo os perfis de maior valor para o negócio.

4. A ordem em que os produtos são adicionados ao carrinho (add_to_cart_order) influencia a probabilidade de recompra desses itens ou de outros itens no mesmo pedido?
    - Análise:
        - Taxa de Recompra por Posição no Carrinho: Agrupar por add_to_cart_order_group e calcular a taxa de recompra média para os produtos nessa posição.
        - Prioridade de Recompra: Analisar se produtos que foram adicionados primeiro (add_to_cart_order = 1) são mais propensos a serem recomprados em pedidos futuros.
    - Visualização:
        - Gráfico de Barras: "Taxa de Recompra por Grupo de Ordem de Adição ao Carrinho".

- Linhas: Mostrar a queda da taxa de recompra à medida que a add_to_cart_order aumenta.

- Insights Esperados: Os primeiros itens adicionados ao carrinho podem ter uma taxa de recompra mais alta, indicando que são os produtos prioritários ou os mais essenciais para o cliente.

5. Qual o comportamento de compra dos clientes em relação à experimentação de novos produtos? Clientes que experimentam mais itens novos têm uma taxa de recompra geral diferente?

    - Análise:

        - Proporção de Novos Produtos por Pedido: Calcular a percent_new_products por order_id e analisar sua distribuição.

        - Taxa de Recompra do Usuário vs. Experimentação: Agrupar usuários por média de percent_new_products ao longo de seus pedidos (ex: 'Baixa Experimentação', 'Média Experimentação', 'Alta Experimentação') e comparar suas user_reorder_rate médias.

    - Visualização:

        - Histograma: Distribuição de percent_new_products por pedido.

        - Gráfico de Barras: "Taxa de Recompra Média por Nível de Experimentação de Novos Produtos do Usuário".

    - Insights Esperados: Clientes que experimentam mas podem ser mais engajados e ter uma taxa de recompra geral mais alta (ou mais baixa, se eles estiverem sempre buscando novidades e não fidelizando produtos). Esta análise ajuda a validar a estratégia de incentivar a descoberta de produtos.

Referências:

- KAGGLE. Instacart Market Basket Analysis. Disponível em: https://www.kaggle.com/datasets/psparks/instacart-market-basket-analysis. Acesso em: junho de 2025.