Instacart Market Basket Analysis

Jean Cardoso, Ph.D.

22 de julho de 2024

Resumo

Este documento descreve a metodologia e as conclusões do projeto do Kaggle "Instacart Market Basket Analysis", que visa prever quais são os produtos que os clientes mais provavelmente comprarão em seus pedidos futuros. O projeto abrange análise exploratória de dados, pré-processamento, modelagem e teste de previsão, bem como a interpretação dos resultados alcançados.

1 Introdução

A competição do Kaggle "Instacart Market Basket Analysis" visa prever quais produtos da loja os consumidores comprarão novamente em suas compras futuras. Este é um dos desafios de grande interesse nos sistemas de recomendação, uma vez que a capacidade de fornecer tal predição ao cliente auxilia na personalização da recomendação de produtos ao cliente, melhorando a experiência do comprador. Para obter essa previsão, primeiro é feita uma análise exploratória de dados, a fim de compreendermos o banco de dados. Nessa fase, já é possível observar o comportamento de compra dos clientes e oferecer insights para campanhas de marketing e fidelização de clientes, por exemplo. É nesse momento também que podem surgir perguntas como a influência de determinadas variáveis a algum comportamento de compra interessante. Uma das perguntas respondidas nesse trabalho é: existe influência dos corredores e departamentos na ordem de adição dos produtos ao carrinho? Novamente, a resposta pode proporcionar ideias para o aumento de venda. Depois da análise de dados, é feita a preparação dos dados para a modelagem. Após feita a mescla dos bancos de dados disponíveis, utilizam-se os dados históricos para criar novas features de interesse. Finalmente, é feita a modelagem utilzando-se cinco métodos de modelagem e a partir de métricas adequadas, obtém-se o melhor modelo.

2 Análise Exploratória de Dados (EDA)

Nossa análise começa pela compreensão dos nossos dados. Nas tabelas abaixo, apresentamos um recorte dos bancos de dados utilizados.

aisle_id	aisle
1	prepared soups salads
2	specialty cheeses
3	energy granola bars

Tabela 1: Banco de dados Aisles

Significado das colunas Aisles

- aisle_id: Identificador único do corredor.
- aisle: Nome do corredor.

department_id	department
1	frozen
2	other
3	bakery

Tabela 2: Banco de dados Departments

Significado das colunas do banco de dados Departments:

- department_id: Identificador único do departamento.
- department: Nome do departamento.

product_id	product_name	aisle_id	department_id
1	Chocolate Sandwich Cookies	61	19
2	All-Seasons Salt	104	13
3	Robust Golden Unsweetened Oolong Tea	94	7

Tabela 3: Banco de dados Products

Significado das colunas do banco de dados Products

- product_id: Identificador único do produto.
- product_name: Nome do produto.
- aisle_id: Identificador do corredor ao qual o produto pertence.
- department_id: Identificador do departamento ao qual o produto pertence.

order_id	user_id	eval_set	order_number	order_dow	order_hour_of_day	days_since_prior_order
2539329	1	prior	1	2	8	NaN
2398795	1	prior	2	3	7	15.0
473747	1	prior	3	3	12	21.0

Tabela 4: Banco de dados Orders

Significado das colunas do banco de dados Orders

- order_id: Identificador único do pedido.
- user_id: Identificador único do usuário (cliente).
- eval_set: Conjunto de avaliação do pedido (prior para pedidos anteriores, train para treinamento, test para teste).
- order_number: Número do pedido (sequência de pedidos feitos pelo usuário).
- order_dow: Dia da semana em que o pedido foi feito (0 = domingo, 1 = segunda-feira, etc.).
- order_hour_of_day: Hora do dia em que o pedido foi feito (0-23).
- days_since_prior_order: Dias desde o último pedido (NaN para o primeiro pedido).

Significado das colunas do banco de dados Order Products Prior

- order_id: Identificador único do pedido.

order_id	product_id	add_to_cart_order	reordered
2	33120	1	1
2	28985	2	1
2	9327	3	0

Tabela 5: Banco de dados Order Products Prior

- product_id: Identificador único do produto.
- add_to_cart_order: Indica a ordem em que os produtos foram adicionados ao carrinho.
- reordered: O número 1 Indica que o produto foi recomprado (já havia sido comprado anteriormente pelo usuário). O número 0 indica que o produto foi comprado pela primeira vez pelo usuário.

order_id	product_id	add_to_cart_order	reordered
1	49302	1	1
1	11109	2	1
1	10246	3	0

Tabela 6: Banco de dados Order Products Train

Significado das colunas do banco de dados Order Products Train

- order_id: Identificador único do pedido.
- product_id: Identificador único do produto.
- add_to_cart_order: Indica a ordem em que os produtos foram adicionados ao carrinho
- reordered: 1 Indica que o produto foi recomprado (já havia sido comprado anteriormente pelo usuário). 0 Indica que o produto foi comprado pela primeira vez pelo usuário.

2.1 Mesclando dados aisles, departments e products

A partir da mescla dos bancos de dados, podemos reunir todas as informações e fazer uma análise de dados completa e compreeder o que podemos inferir a partir dos dados.

Quantidade	Total
Produtos	49688
Corredores	134
Departamentos	21

Tabela 7: Quantidades totais

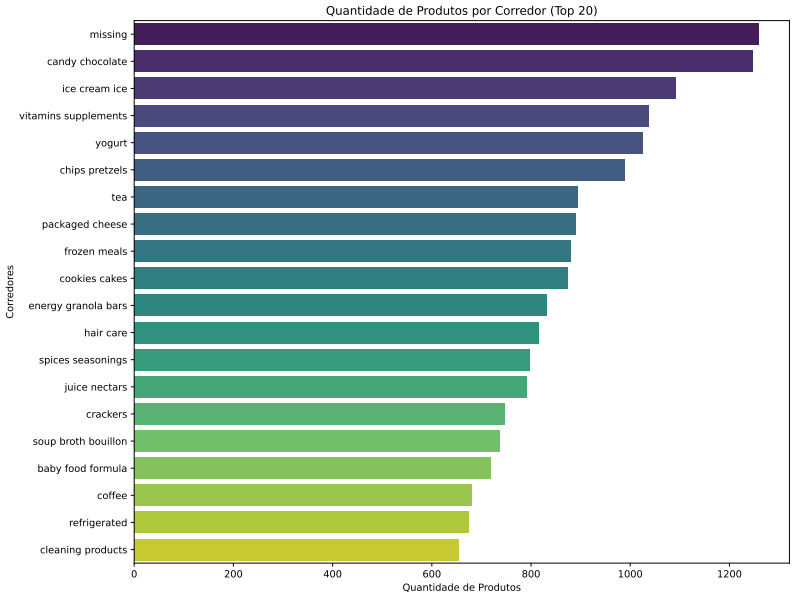

Figura 1: Quantidade de produtos por corredor

Corredores com mais Produtos:

- O corredor 'missing' tem a maior quantidade de produtos. Isso pode indicar produtos que não foram corretamente classificados.

Categorias Populares

- Corredores como 'candy chocolate' e 'ice cream ice' sugerem que produtos de sobremesa e doces têm uma alta presença no inventário.
- A presença de corredores como 'vitamins supplements' e 'hair care' indica uma boa variedade de produtos de saúde e cuidados pessoais.

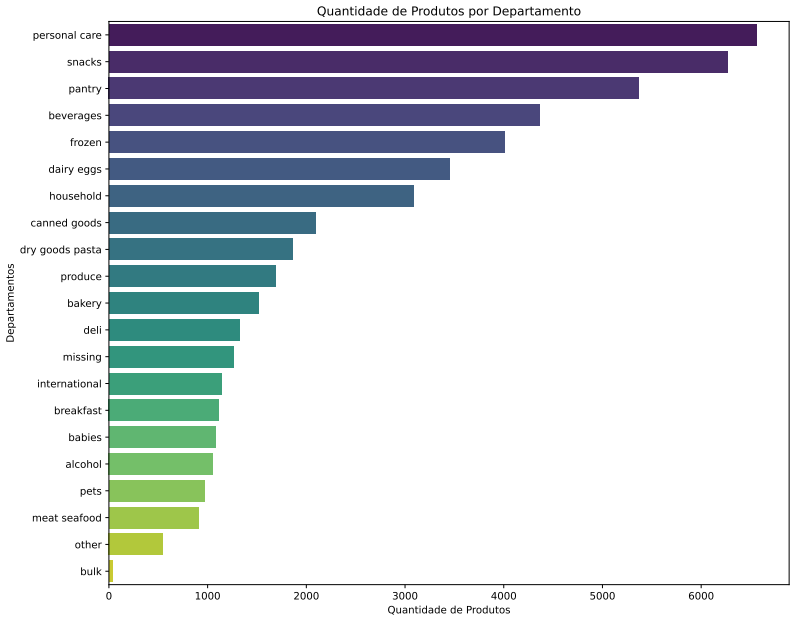

Figura 2: Quantidade de produtos por departamento

Departamentos com mais Produtos:

- **Personal Care**: O departamento de cuidados pessoais tem o maior número de produtos, indicando uma ampla variedade de itens como shampoos, sabonetes, produtos de higiene, entre outros.
- **Snacks**: O departamento de lanches também tem um grande número de produtos, sugerindo uma alta demanda por esses itens.
- **Pantry**: Produtos de despensa (como alimentos não perecíveis) estão em terceiro lugar, refletindo a necessidade contínua de itens básicos na alimentação diária.

Gerenciamento de Categorias Populares:

- **Alta Diversidade em Personal Care**: A ampla variedade de produtos de cuidados pessoais pode ser uma resposta à alta demanda e preferências variadas dos consumidores.
- **Snacks e Beverages**: Departamentos de lanches e bebidas são áreas chave para promoções sazonais, como durante eventos esportivos ou feriados.

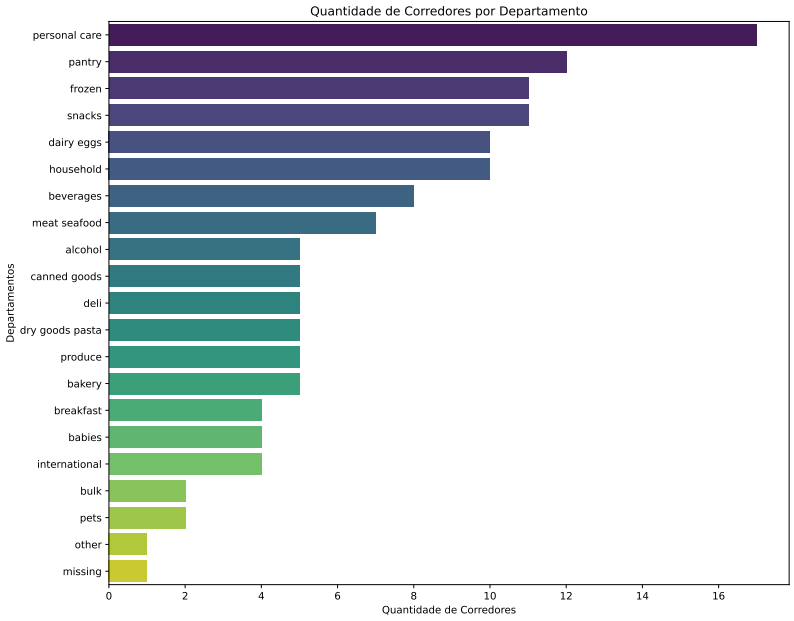

Figura 3: Quantidade de corredores por departamento

Departamentos com mais Corredores:

- **Personal Care**: Este departamento tem a maior quantidade de corredores, sugerindo uma alta variedade de produtos de cuidados pessoais distribuídos em múltiplos corredores.
- **Pantry e Frozen**: Esses departamentos também têm um número significativo de corredores, indicando uma grande diversidade de produtos não perecíveis e congelados.

Departamentos com menos Corredores:

- **Bulk e Pets**: Esses departamentos têm menos corredores, o que pode indicar uma menor variedade de produtos.

Estratégias de Marketing:

- **Exibição de Produtos**: A variedade de produtos em departamentos como 'Personal Care' pode ser usada para criar seções temáticas ou displays promocionais para atrair a atenção dos clientes.

Figura 4: Nuvem de palavras

Termos Frequentes:

- **Organic (Orgânico)**: Este termo é o mais proeminente na nuvem, indicando uma forte presença de produtos orgânicos no catálogo. Isso reflete uma tendência crescente entre os consumidores que preferem produtos naturais e sustentáveis.
- **Original, Natural e Gluten Free (Sem Glúten)**: Outros termos frequentes que indicam a popularidade de produtos em suas formas puras e naturais.

Variedade de Produtos:

- **Alimentos Diversos**: Termos como 'Ice Cream', 'Chicken', 'Peanut Butter', 'Chocolate', 'Yogurt' e 'Apple' destacam uma ampla gama de produtos alimentares disponíveis, desde sobremesas e lanches até alimentos básicos.
- **Produtos de Saúde e Cuidados**: Palavras como 'Cat Food', 'Dog Food', 'Shampoo' e 'Detergent' mostram que a oferta inclui não apenas alimentos, mas também produtos de cuidados pessoais e para animais de estimação.

2.2 Explorando banco de dados orders

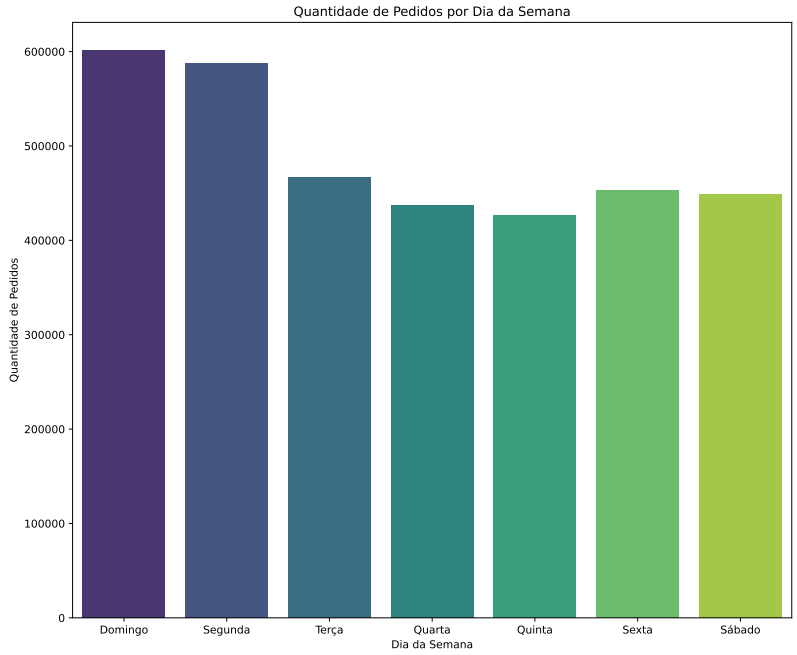

Figura 5: Quantidade de pedidos por dia da semana

Picos de Atividade:

- **Domingo:** O número de pedidos é mais alto aos domingos, sugerindo que este é o dia mais popular para os clientes fazerem pedidos. Isso pode ser devido à preparação para a semana seguinte.
- **Segunda-feira:** O segundo maior número de pedidos ocorre nas segundas-feiras, indicando que muitos clientes ainda estão completando suas compras no início da semana.

Dias de Menor Atividade:

- **Terça-feira a Sábado:** Há uma queda visível no número de pedidos durante esses dias.

Campanhas de Marketing:

- **Promoções Alvo:** Realizar promoções específicas durante os dias de menor atividade (quarta à quinta-feira) pode ajudar a distribuir a demanda de forma mais uniforme ao longo da semana.

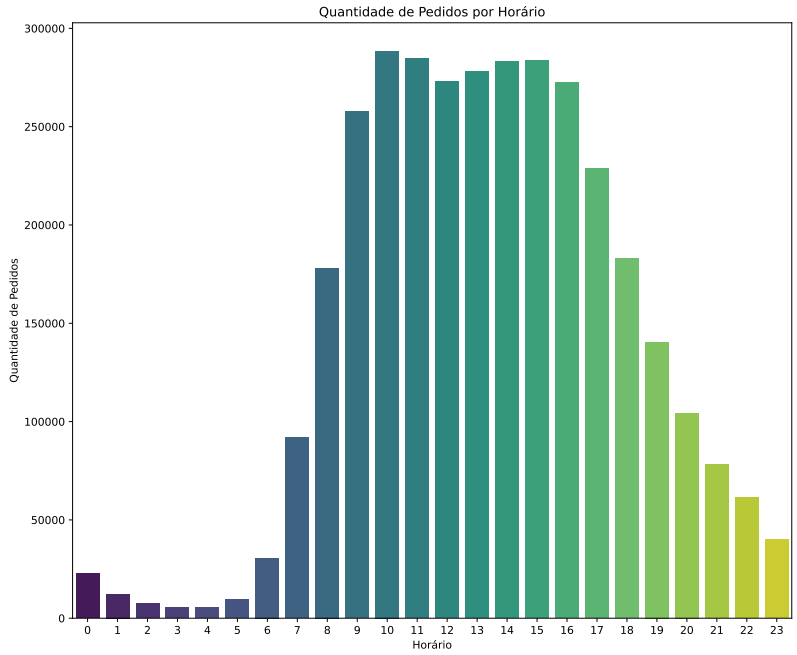

Figura 6: Quantidade de pedidos por hora

Picos de Atividade:

- **9h às 16h:** O maior número de pedidos ocorre entre as 9h e 16h. Isso sugere que a maioria dos clientes faz compras durante o meio do dia, provavelmente antes ou durante o horário de almoço.
- **10h e 11h:** Esses horários especificamente têm o maior volume de pedidos.

Baixa Atividade:

- **0h às 6h:** Há poucos pedidos durante a madrugada e as primeiras horas da manhã.

Atividade Moderada:

- **7h às 8h:** Há um aumento gradual no número de pedidos a partir das 7h, sugerindo que alguns clientes começam a fazer compras logo de manhã.

Campanhas de Marketing:

- **Promoções Matinais:** Oferecer promoções especiais nas primeiras horas da manhã (por exemplo, descontos até as 9h) pode incentivar mais compras durante esse período de baixa atividade.
- **Ofertas Noturnas:** Implementar promoções específicas após as 20h pode ajudar a aumentar as vendas durante as horas de menor movimento.

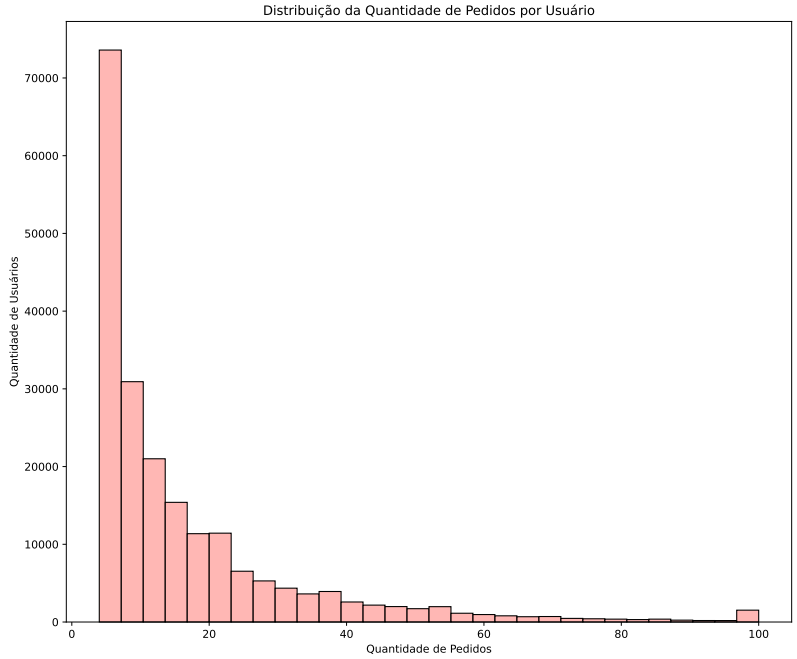

Figura 7: Distribuição quantidade de pedidos por usuário

Usuários com Poucos Pedidos:

- A maioria dos usuários fez um número pequeno de pedidos.

Decrescimento Exponencial:

- O gráfico mostra um decrescimento exponencial, onde o número de usuários diminui rapidamente à medida que o número de pedidos aumenta. Isso indica que há menos usuários que fizeram um grande número de pedidos.

Usuários Frequentes:

- Há uma pequena quantidade de usuários que fizeram muitos pedidos. Mesmo assim, existe uma diminuição acentuada no número de usuários conforme o número de pedidos aumenta para 30 ou mais.

Aquisição e Retenção de Clientes:

- **Alta Aquisição, Baixa Retenção**: O pico acentuado em 1-2 pedidos sugere que, embora muitos usuários sejam adquiridos, a retenção pode ser um problema, pois muitos não voltam para fazer compras adicionais.
- **Foco na Retenção**: Investir em estratégias de retenção, como programas de fidelidade, descontos para compras repetidas, e campanhas de reengajamento podem ajudar a aumentar o número de usuários que fazem pedidos adicionais.

Segmentação de Clientes:

- **Clientes Novos vs. Clientes Fiéis**: A segmentação dos clientes com base no número de pedidos pode ajudar a identificar novos clientes e clientes fiéis. Novos clientes podem precisar de incentivos para se tornarem clientes recorrentes.
- **Campanhas Personalizadas**: Desenvolver campanhas de marketing personalizadas para cada segmento pode aumentar a eficácia das campanhas. Por exemplo, oferecer um desconto para um cliente que fez apenas um pedido pode incentivá-lo a fazer um segundo pedido.

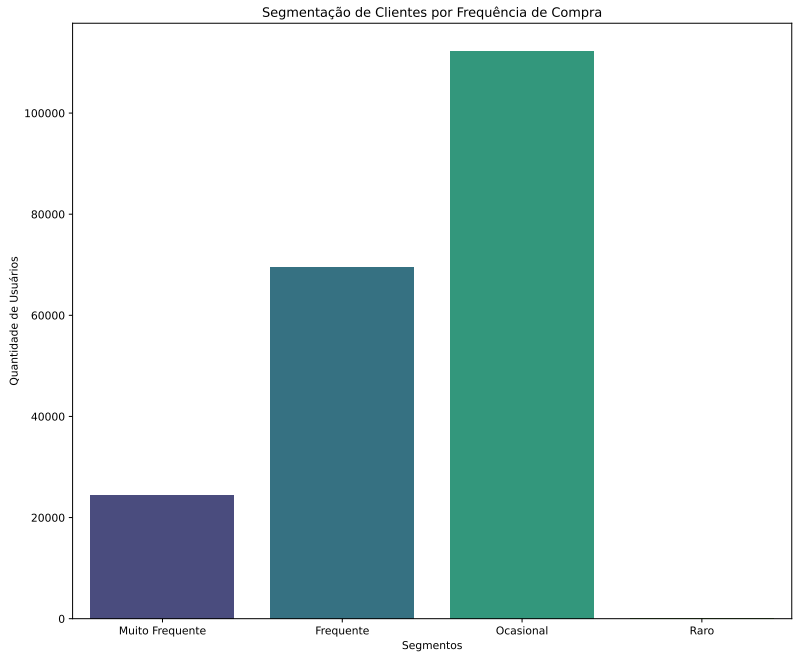

Figura 8: Segmentação de clientes por frequência de compra

Estratégias de Retenção de Clientes:

- **Foco em Clientes Frequentes e Muito Frequentes**: Clientes que fazem compras frequentes são valiosos e devem ser recompensados para mantê-los engajados. Programas de fidelidade, descontos exclusivos e ofertas personalizadas podem ajudar a aumentar a retenção desses clientes.
- **Reengajamento de Clientes Ocasional**: A maioria dos clientes está no segmento ocasional. Oferecer incentivos para aumentar a frequência de compras, como cupons de desconto para a próxima compra ou programas de recompensas baseados na frequência de pedidos, pode ajudar a mover esses clientes para segmentos mais frequentes.

Campanhas de Marketing Segmentadas:

- **Personalização de Ofertas**: As campanhas de marketing podem ser personalizadas com base no segmento de frequência de compra. Por exemplo, campanhas de reengajamento podem ser direcionadas aos clientes ocasionais, enquanto campanhas de fidelização podem ser direcionadas aos clientes muito frequentes.

Análise de Comportamento de Compra:

- **Identificação de Padrões**: Analisar os padrões de compra dentro de cada segmento pode ajudar a entender melhor as necessidades e preferências dos clientes. Isso pode incluir estudar *quais produtos são mais populares entre clientes frequentes versus ocasionais.*

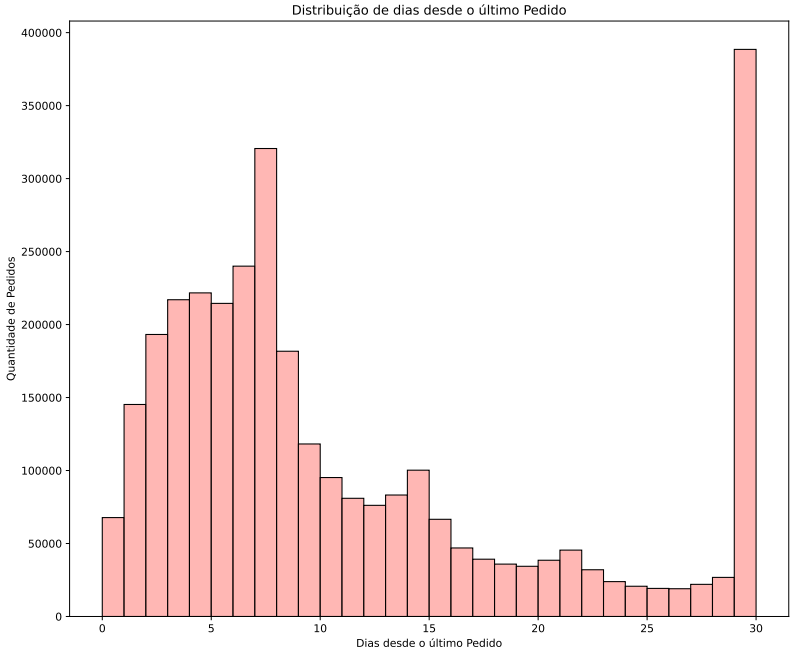

Figura 9: Segmentação de clientes por frequência de compra

Picos na Distribuição:

- **Dias 7, 14 e 30**: Existem picos significativos em intervalos específicos, especialmente aos 7, 14 e 30 dias. Isso sugere que muitos clientes têm ciclos de compra regulares, possivelmente semanais, quinzenais e mensais.

Tendências Gerais:

- **Dias 1 a 10**: Há uma alta frequência de pedidos logo após os primeiros dias, especialmente até o décimo dia, indicando que muitos clientes realizam compras dentro de uma a duas semanas após o pedido anterior.
- **Dias 11 a 20**: A frequência de pedidos diminui gradualmente, com um pequeno aumento em torno do dia 15.
- **Dia 30**: O número de pedidos aumenta drasticamente no dia 30, sugerindo que muitos clientes fazem compras mensais.

Segmentação de Clientes:

- **Campanhas de Reengajamento**: Clientes que tendem a fazer compras em intervalos regulares podem ser alvo de campanhas de reengajamento, oferecendo lembretes ou promoções próximas aos seus ciclos de compra típicos.

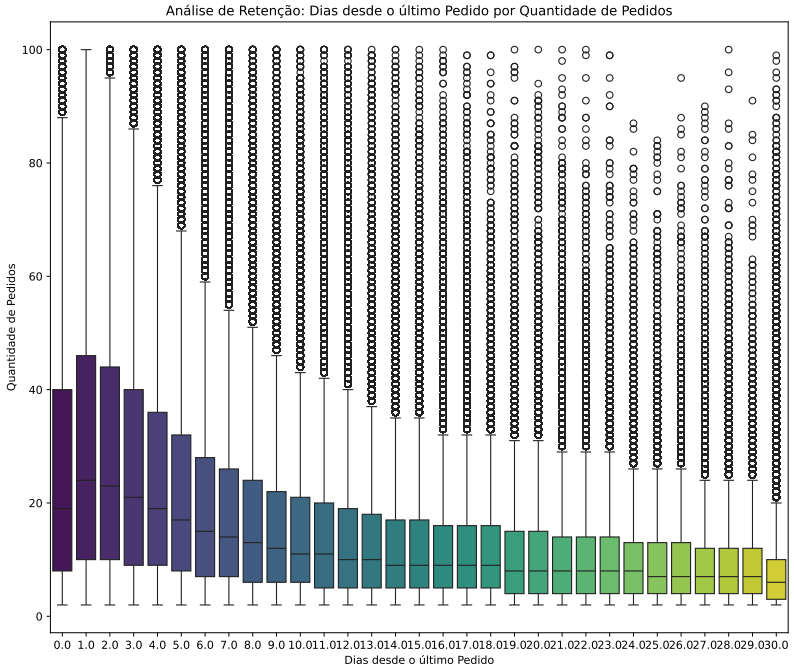

Figura 10: Analise de retenção

Tendência Geral:

- **Diminuição dos Dias Desde o Último Pedido**: À medida que o número de ordem aumenta, a mediana dos dias desde o último pedido tende a diminuir. Isso indica que, quanto mais pedidos um cliente faz, menor é o intervalo entre seus pedidos.

Estabilização com o Tempo:

- **Estabilização da Frequência de Pedidos**: Após aproximadamente 20 pedidos, a variação dos dias desde o último pedido diminui e a frequência de pedidos se estabiliza, com a maioria dos clientes fazendo pedidos em intervalos regulares.

Outliers:

- **Presença de Outliers**: Há muitos outliers presentes ao longo de todo o gráfico, indicando que há sempre alguns clientes que não seguem o padrão geral e podem ter intervalos de compra muito maiores ou muito menores.

Comportamento de Compra dos Clientes:

- **Fidelização Gradual**: À medida que os clientes fazem mais pedidos, eles tendem a se tornar mais frequentes e regulares, sugerindo que a fidelização aumenta com o tempo e com a familiaridade com o serviço.

Segmentação de Clientes:

- **Identificação de Clientes em Risco**: Clientes que, após muitos pedidos, ainda têm uma grande variação nos dias desde o último pedido podem estar em risco de churn (abandono) e podem ser alvo de campanhas de reengajamento.

6. Estratégias de Marketing e Retenção:

- **Incentivos para Novos Clientes**: Oferecer incentivos para reduzir os intervalos entre os primeiros pedidos pode ajudar a estabelecer um hábito de compra mais rápido.

- **Recompensas para Clientes Fiéis**: Desenvolver programas de fidelidade que recompensem a regularidade e frequência de pedidos pode ajudar a manter os clientes fiéis e incentivá-los a continuar comprando regularmente.

2.3 Analisando dados históricos

product_name	department	aisle
Organic Baby Spinach	produce	packaged vegetables fruits
Bag of Organic Bananas	produce	fresh fruits
Organic Raspberries	produce	packaged vegetables fruits
Organic Hass Avocado	produce	fresh fruits
Banana	produce	fresh fruits
Organic Avocado	produce	fresh fruits
Organic Strawberries	produce	fresh fruits
Organic Whole Milk	dairy eggs	milk
Large Lemon	produce	fresh fruits
Organic Yellow Onion	produce	fresh vegetables
Cucumber Kirby	produce	fresh vegetables
Organic Fuji Apple	produce	fresh fruits
Strawberries	produce	fresh fruits
Organic Blueberries	produce	packaged vegetables fruits
Organic Grape Tomatoes	produce	packaged vegetables fruits
Organic Lemon	produce	fresh fruits
Organic Garlic	produce	fresh vegetables
Limes	produce	fresh fruits
Apple Honeycrisp Organic	produce	fresh fruits
Organic Zucchini	produce	fresh vegetables

Tabela 8: Top 20 produtos mais comprados por departamento e corredores

product_name	department	aisle
Blueberry Blast Fruit and Chia Seed Bar	snacks	energy granola bars
Yellow Fish Breading	pantry	marinades meat preparation
Citronge Extra Fine Orange Liqueur	alcohol	spirits
1,000 Mg Vitamin C Tangerine Grapefruit Efferv...	beverages	energy sports drinks
String Of Pearl White Sprinkles	pantry	baking supplies decor
Pasta Shapes In Tomato Sauce	missing	missing
Multigrain Penne Rigate	dry goods pasta	dry pasta
All Natural Stevia Liquid Extract Sweetener	pantry	baking ingredients
Escapes Variety Pack	alcohol	beers coolers
Original Lager	alcohol	beers coolers
Drink Distinct All Natural Soda Pineapple Coco...	missing	missing
Berry Sprouted Blend Cereal	breakfast	cereal
Chocolate Peppermint Tart	bakery	bakery desserts
Orangemint Flavored Water	beverages	water seltzer sparkling water
Aged Parmesan Cheese Sticks	snacks	crackers
Rosa Mosqueta Rose Hip Seed Oil	personal care	skin care
Indulgent Cherry & Dark Chocolate Whole Milk G...	dairy eggs	yogurt
Orange Flavored Ice Cubes	frozen	ice cream ice
Lindor Peppermint White Chocolate Truffles	snacks	candy chocolate
Water With Electrolytes	missing	missing

Tabela 9: Top 20 produtos menos comprados por departamento e corredores

Diversidade de Departamentos:

- Na primeira tabela podemos notar uma predominância do departamento produce e dos corredores fresh fruits.
- Ao contrário dos produtos mais comprados, os produtos menos comprados são distribuídos por uma maior variedade de departamentos.

Identificação de Produtos de Baixa Rotatividade:

- Os produtos listados têm baixa rotatividade, o que pode indicar uma baixa demanda. Isso pode ser devido à falta de necessidade regular, produtos menos populares ou pouco conhecidos.

Product Name	**Department**	**Aisle**
Organic Baby Spinach	produce	packaged vegetables fruits
Bag of Organic Bananas	produce	fresh fruits
Organic Raspberries	produce	packaged vegetables fruits
Organic Hass Avocado	produce	fresh fruits
Banana	produce	fresh fruits
Organic Avocado	produce	fresh fruits
Organic Strawberries	produce	fresh fruits
Organic Half & Half	dairy eggs	cream
Organic Whole Milk	dairy eggs	milk
Large Lemon	produce	fresh fruits
Organic Yellow Onion	produce	fresh vegetables
Cucumber Kirby	produce	fresh vegetables
Organic Fuji Apple	produce	fresh fruits
Strawberries	produce	fresh fruits
Organic Blueberries	produce	packaged vegetables fruits
Organic Lemon	produce	fresh fruits
Organic Garlic	produce	fresh vegetables
Limes	produce	fresh fruits
Apple Honeycrisp Organic	produce	fresh fruits
Organic Zucchini	produce	fresh vegetables

Tabela 10: Top 20 produtos mais recomprados por departamentos e corredores

Importância de Frutas e Vegetais Frescos:

- A alta frequência de frutas e vegetais frescos nos produtos mais recomprados indica que esses itens são essenciais nas compras recorrentes dos consumidores.
- Corredores como “fresh fruits”e “fresh vegetables” são críticos para a loja.

Oportunidades de Expansão:

- Promover produtos de outros departamentos junto com produtos frescos e orgânicos pode aumentar a venda cruzada e a diversificação das compras dos consumidores.

Product Name	Department	Aisle
Witch Hazel For Face & Body	personal care	first aid
Lemon Dishwashing Detergent + Clorox Stain Fig...	household	dish detergents
Chili with Beans, Vegetarian	canned goods	canned meals beans
Cognac, Privilege V.S.O.P	alcohol	spirits
Coconut Hibiscus Conditioner	personal care	hair care
Ultra Antibacterial Dish Liquid	missing	missing
Cocktail Time Hummus	deli	fresh dips tapenades
Sun-Dried Tomato Garlic	canned goods	canned jarred vegetables
Decorating Sugar Raspberry Red	pantry	baking supplies decor
Arctic Cod Liver Oil Lemon	personal care	vitamins supplements
Premium Cheeses Parmesan & Romaso	dairy eggs	packaged cheese
Oreo Chocolate Candy Bar	missing	missing
Tuscan Peasant Soup	dry goods pasta	grains rice dried goods
Personal Lubricant	other	other
Citrus Power Liquid Cleaner	household	cleaning products
L-Tyrosine, 500 mg	personal care	vitamins supplements
Oven Crispy Clam Strips	frozen	frozen appetizers sides
Healing Lip Balm	personal care	facial care
Gluten Free Nut Bar Cranberry	snacks	energy granola bars
Cotes De Provence Rose	other	other

Tabela 11: Top 20 produtos menos recomprados por departamentos e corredores

Diversidade de Departamentos:

- Os produtos menos recomprados são distribuídos por uma variedade de departamentos, como 'household', 'personal care', 'meat seafood', 'beverages', 'pantry', 'canned goods', 'deli' e 'dry goods pasta'.

Presença de Produtos Específicos e de Uso Ocasional:

- Muitos desses produtos são itens específicos ou de uso ocasional. Produtos de nicho, como 'Decaf Sweet Coconut Thai Chai Tea' e 'Pomegranate Pizzazz Herbal Tea Bags Caffeine Free', também aparecem na lista.

Estratégias de Marketing Direcionadas:

- Para aumentar as vendas desses produtos, estratégias de marketing específicas podem ser implementadas. Isso pode incluir promoções ou descontos.

- Posicionamento estratégico na loja virtual, como destacar esses produtos em seções de 'Novidades' ou 'Ofertas Especiais', pode ajudar a aumentar a visibilidade.

Medidas de dispersão:

- Média do número de produtos por pedido: 10.11

- Moda do número de produtos por pedido: 5

- Mediana do número de produtos por pedido: 8.0

Segmentação de Clientes:

- Os valores da média, moda e mediana podem indicar diferentes comportamentos de compra entre os clientes. Por exemplo, alguns clientes podem fazer grandes compras esporadicamente (influenciando a média), enquanto outros fazem compras pequenas e frequentes (influenciando a moda).

Estratégias de Venda:

- Compreender essas estatísticas pode ajudar a empresa a ajustar suas estratégias de marketing e promoção. Por exemplo, incentivar os clientes que costumam fazer pedidos pequenos a adicionar mais produtos ao carrinho com ofertas de frete grátis ou descontos em compras maiores pode aumentar as compras.

2.4 Algumas perguntas

Durante a análise exploratória surgem algumas perguntas interessantes. A seguir, são compartilhadas duas perguntas bem como as respostas obtidas da análise.

2.4.1 Existe Influência dos Corredores e Departamentos na Ordem de Adição dos Produtos ao Carrinho?

Categoria	F-statistic	p-value
Aisles	57.63962113873165	0.0
Departments	231.41110720066246	0.0

Tabela 12: Anova

Categoria	**R-squared**	**Adj. R-squared**
Aisles	0.134	0.132
Departments	0.085	0.132

Tabela 13: Regressão OLS

Influência dos Corredores:

- **Significância Estatística**: A análise ANOVA e a regressão mostram que os corredores têm uma influência estatisticamente significativa na posição de adição ao carrinho.
- **Magnitude da Influência**: Com um R-squared de 13.4%, os corredores explicam uma parte substancial da variação na ordem de adição ao carrinho. Isso sugere que a forma como os produtos são categorizados em corredores na interface do usuário impacta significativamente o comportamento de adição ao carrinho.
- **Insights**: Os consumidores tendem a adicionar produtos de certos corredores em uma ordem específica, possivelmente influenciada pela navegação e apresentação dos produtos na loja virtual.

Influência dos Departamentos:

- **Significância Estatística**: Tanto a ANOVA quanto a regressão indicam que os departamentos também têm uma influência estatisticamente significativa.
- **Magnitude da Influência**: Com um R-squared de 8.5%, os departamentos explicam uma parte menor, mas ainda significativa, da variação na ordem de adição ao carrinho. Isso sugere que a categorização mais ampla dos produtos em departamentos também afeta o comportamento de compra.
- **Insights**: Produtos em certos departamentos são adicionados ao carrinho em momentos específicos, refletindo possivelmente preferências de compra ou a forma como esses produtos são agrupados na loja virtual.

Análise Comportamental:

- **Estudo de Padrões de Navegação**: Analisar como os usuários navegam pela loja virtual e quais corredores e departamentos visitam primeiro pode fornecer insights adicionais sobre o comportamento de compra.
- **Testes A/B**: Realizar testes A/B para diferentes organizações de produtos na interface pode ajudar a identificar a disposição que maximiza a eficiência de adição ao carrinho e a satisfação do usuário.

2.4.2 Como diferentes produtos são agrupados com base nas preferências de adição ao carrinho?

Cluster	Count	Mean	Std	Min	25%	50%	75%	Max
0	10666.0	4.107163	0.987712	1.0	3.5	4.0	5.0	5.0
1	2616.0	14.073012	3.261987	11.5	12.0	13.0	15.0	53.0
2	17995.0	8.844568	0.939505	8.0	8.0	9.0	9.0	11.0
3	18408.0	6.568856	0.551571	5.5	6.0	7.0	7.0	7.5

Tabela 14: Estatísticas Descritivas por Cluster

Product Name	**Median Add to Cart Order**	**Aisle**	**Department**	**Cluster**
#2 Mechanical Pencils	5.0	more household	household	0
'Swingtop' Premium Lager	5.0	beers coolers	alcohol	0
0% Fat Strawberry Greek Yogurt	5.0	yogurt	dairy eggs	0
0% Fat Vanilla Greek Yogurt	5.0	yogurt	dairy eggs	0
1% Lowfat Milk	5.0	milk	dairy eggs	0

Tabela 15: Estatísticas de Produtos por Cluster

Cluster 0 (Produtos adicionados cedo na compra):

- **Produtos de Necessidade Diária**: Itens como leite desnatado e iogurte são frequentemente comprados no início, sugerindo que os consumidores priorizam esses produtos básicos.
- **Cerveja Premium**: A presença de uma cerveja premium ('Swingtop' Premium Lager) pode indicar uma compra planejada ou regular de itens alcoólicos específicos.
- **Material Escolar**: Produtos como lapiseiras são adicionados cedo, possivelmente devido à sua necessidade imediata ou frequência de reposição.

Product Name	**Median Add to Cart Order**	**Aisle**	**Department**	**Cluster**
100% Natural Just Peaches	7.5	fruit vegetable snacks	snacks	3
100% Natural Chicken Broth	7.5	soup broth bouillon	canned goods	3
1,000 Mg Vitamin C Lemon Lime Effervescent Drink	7.5	vitamins supplements	personal care	3
100% Oil-Free Eye Makeup Remover	7.5	beauty	personal care	3
100% Pure Cotton Ovals	7.5	beauty	personal care	3

Tabela 16: Estatísticas de Produtos por Cluster

Cluster 3 (Produtos adicionados em posição intermediária):

- **Saúde e Beleza**: Produtos de beleza e suplementos vitamínicos aparecem com frequência neste cluster, indicando que os consumidores talvez adicionem esses itens após cobrir necessidades básicas.
- **Alimentos Naturais**: A presença de itens naturais e saudáveis, como caldo de galinha e pêssegos naturais, sugere um foco crescente em alimentos saudáveis.

Product Name	**Median Add to Cart Order**	**Aisle**	**Department**	**Cluster**
100% Apple Juice Concentrate	11.0	frozen juice	frozen	2
100% Extra Virgin Olive Oil	11.0	oils vinegars	pantry	2
100% Juice Peach Mango Juice	11.0	juice nectars	beverages	2
100% Mountain Spring Water	11.0	water seltzer sparkling water	beverages	2
100% Natural Skin & Hair Revitalizing Coconut Oil	11.0	hair care	personal care	2

Tabela 17: Estatísticas de Produtos por Cluster

Cluster 2 (Produtos adicionados em posições variadas):

- **Produtos de Conveniência e Saúde**: Itens como sucos, água mineral e azeite de oliva são comprados em diversas posições, indicando que são produtos de conveniência frequentemente adicionados ao carrinho.
- **Cuidados Pessoais**: Óleos naturais para pele e cabelo estão presentes, mostrando a importância dos cuidados pessoais na rotina de compra.

Product Name	**Median Add to Cart Order**	**Aisle**	**Department**	**Cluster**
Original Submarine Dressing	35.5	condiments	pantry	1
Eczema Control	41.0	first aid	personal care	1
Vanilla Bean Sheep Milk Ice Cream	46.0	ice cream ice	frozen	1
Strawberry Energy Gel	50.0	energy granola bars	snacks	1
Citronge Extra Fine Orange Liqueur	53.0	spirits	alcohol	1

Tabela 18: Estatísticas de Produtos por Cluster

Cluster 1 (Produtos adicionados mais tarde na compra):

- **Itens de Luxo e Especializados**: Produtos como licores finos e sorvetes de leite de ovelha indicam compras de indulgência ou especializadas feitas mais tarde na compra.
- **Cuidados com a Saúde**: Produtos para controle de eczema e géis energéticos aparecem, sugerindo compras mais ponderadas para necessidades específicas.
- **Condimentos e Adicionais**: Molhos e condimentos são adicionados mais tarde, possivelmente como complementos para refeições planejadas.

3 Feature engineering

Agora que já entendemos nossos dados, observamos melhor seus padrões e até já respondemos algumas perguntas, vamos passar para a fase de preparação para a modelagem. A ideia será a seguinte:

- Utilizar os dados históricos para criar novas features.
- Utilizar essas novas features juntamente com os dados de treino para a modelagem.
- Utilizar nosso modelo para prever os resultados dos dados de teste (out of time).

Iniciamos juntando os bancos de dados products, aisles, departments, orders e prior_orders.

Coluna	Valores Faltantes
order_id	0
user_id	0
eval_set	0
order_number	0
order_dow	0
order_hour_of_day	0
days_since_prior_order	2078068
product_id	0
add_to_cart_order	0
reordered	0
product_name	0
aisle_id	0
department_id	0
aisle	0
department	0

Tabela 19: Valores Faltantes por Coluna

Relembrando:

- days_since_prior_order: Número de dias desde o pedido anterior (NaN para o primeiro pedido).

Sendo assim, nossa abordagem para lidar com esses valores faltantes foi a seguinte:

- Criamos uma nova coluna indicadora para o primeiro pedido.
- Substituimos os NaN de days_since_prior_order por zero.

Observação: Usamos uma função para reduzir a quantidade de memória que os dados usavam e salvamos esses dados com a memória reduzida no formato .parquet.

3.1 Criação de novas features

A criação das novas features foi dividida em duas categorias:

- Features relacionadas ao usuário.
- Features relacionadas ao produto.

3.1.1 Features associadas ao usuário:

Banco de dados user_features.parquet:

- user_id: Identificador único do usuário.
- user_order_count: O número total de pedidos feitos por cada usuário.
- user_avg_products_per_order: A média de produtos comprados por pedido para cada usuário.
- user_unique_products: A diversidade de produtos (número de produtos únicos) comprados por cada usuário.
- user_reorder_ratio: A frequência com que os usuários recompram produtos, ou seja, a proporção de produtos comprados que foram pedidos novamente.

- user_avg_days_between_orders: A média de dias entre os pedidos feitos por cada usuário.

Banco de dados behavior_features.parquet:

- **user_id**: Identificador único do usuário.
- **user_order_count**: Número total de pedidos realizados por usuário.
- **user_avg_products_per_order**: Média de produtos por pedido, calculada para cada usuário.
- **user_unique_products**: Número de produtos únicos comprados por cada usuário.
- **user_reorder_ratio**: Frequência de recompra, medida pela média de vezes que o usuário reordenou produtos.
- **user_avg_days_between_orders**: Média de dias entre pedidos, calculada para cada usuário.

Banco de dados temporal_features.parquet:

- **user_id**: Identificador único do usuário.
- **order_dow_sin**: Representação cíclica do dia da semana (seno).
- **order_dow_cos**: Representação cíclica do dia da semana (cosseno).
- **order_hour_of_day_sin**: Representação cíclica da hora do dia (seno).
- **order_hour_of_day_cos**: Representação cíclica da hora do dia (cosseno).
- **user_hour_order_freq**: Frequência de pedidos por hora do dia para cada usuário.
- **user_dow_order_freq**: Frequência de pedidos por dia da semana para cada usuário.

Banco de dados user_session_features.parquet:

- **user_id**: Identificador único do usuário.
- **user_avg_days_since_prior**: Média de dias desde o pedido anterior por usuário.
- **user_days_since_last_order**: Número de dias desde o último pedido por usuário.

3.1.2 Features associadas ao produto

Banco de dados product_features.parquet:

- **product_id**: Identificador único do produto.
- **product_popularity**: Popularidade do produto, medida pelo total de pedidos para cada produto.
- **user_id**: Identificador único do usuário.
- **user_product_popularity**: Popularidade do produto por usuário, medida pelo número de pedidos por produto por usuário.

Banco de dados product_days_diff.parquet:

- **product_id**: Identificador único do produto.
- **product_avg_days_since_prior**: Diferença média de dias entre pedidos que contêm o mesmo produto.

Banco de dados purchase_features.parquet:

- **product_id**: Identificador único do produto.
- **product_order_freq**: Frequência de compra do produto.

- **product_reorder_ratio**: Proporção de reordem do produto.
- **avg_pos_incart**: Posição média do produto no carrinho.

Agora que as features estão criadas, vamos começar a criar de fato nosso banco de dados para a modelagem. Iniciamos fazendo um merge dos dados orders (filtrado pela coluna eval_set quando ela é igual a train) com o banco de dados orders_train. E, posteriormente, na mesma ordem em que os novos bancos de dados foram descritos, faremos a união. Ao finalizar, teremos o banco de dados que chamamos de train_orders.

Coluna	**Valores Nulos**
order_id	0
user_id	0
eval_set	0
order_number	0
order_dow	0
order_hour_of_day	0
days_since_prior_order	0
product_id	0
add_to_cart_order	0
reordered	0
user_order_count	0
user_avg_products_per_order	0
user_unique_products	0
user_reorder_ratio	0
user_avg_days_between_orders	0
user_reorder_rate	0
user_reorder_freq	0
user_reordered_products_ratio	0
order_dow_sin	0
order_dow_cos	0
order_hour_of_day_sin	0
order_hour_of_day_cos	0
user_hour_order_freq	0
user_dow_order_freq	0
user_avg_days_since_prior	0
user_days_since_last_order	0
product_popularity	9
user_product_popularity	9
product_avg_days_since_prior	9
product_order_freq	9
product_reorder_ratio	9
avg_pos_incart	9

Tabela 20: Quantidade de valores nulos por coluna no DataFrame

Para garantir que o modelo possa treinar adequadamente e não seja afetado por valores ausentes, utilizamos diferentes estratégias para preencher as colunas com valores faltantes no DataFrame `train_orders`. A seguir, descrevemos as estratégias aplicadas para cada coluna:

product_order_freq:

- Valores ausentes foram preenchidos com 0. Isso ocorre porque a frequência de pedido do produto pode ser considerada como zero quando não há dados disponíveis.

product_reorder_ratio:

- Valores ausentes foram preenchidos com 0. O raciocínio é que se não há informações sobre a proporção de reordem, assume-se que o produto não foi reordenado.

avg_pos_incart:

- Valores ausentes foram preenchidos com -1. Utilizar -1 ajuda a distinguir esses casos dos valores válidos de posição média no carrinho, que são sempre não-negativos.

product_avg_days_since_prior:

- Valores ausentes foram preenchidos com a média da coluna. Isso suaviza o impacto dos valores ausentes, substituindo-os por um valor médio representativo.

product_popularity:

- Valores ausentes foram preenchidos com 0. A popularidade do produto é considerada zero quando não há registros de pedidos para o produto.

user_product_popularity:

- Valores ausentes foram preenchidos com 0. Isso indica que o usuário específico não tem um histórico de pedidos para o produto, portanto, a popularidade do produto para o usuário é considerada zero.

Essas estratégias foram escolhidas para minimizar o impacto dos valores ausentes nos modelos de machine learning, garantindo que cada coluna seja tratada de forma que faça sentido para suas características específicas.

Para a criação dos dados de 'teste' ou 'oot' (out off time) ocorreu quase o mesmo processo, sofrendo apenas duas alterações:

- O banco test_orders é criado filtrando a coluna eval_test de orders por test.
- O merge de product_features com os dados test_orders é feito pelo user_id.

4 Métodos de Modelagem

4.1 Modelagem e Treinamento com LightGBM

Descrição: Implementa um pipeline completo de treinamento de um modelo LightGBM, desde o pré-processamento dos dados até a avaliação do modelo.
Processo:

- **Pré-processamento:** Divide os dados em conjuntos de treino e teste e remove características irrelevantes e altamente correlacionadas.
- **Construção do Modelo:** Define os parâmetros do modelo LightGBM.
- **Treinamento:** Treina o modelo com os dados de treino.
- **Avaliação:** Avalia o desempenho do modelo nos dados de teste.

4.1.1 Pré-processamento

- **Divisão dos Dados:** Divide os dados em conjuntos de treino e teste, utilizando 'train_test_split' com estratificação.
- **Remoção de Características Irrelevantes:** Remove características constantes que não contribuem para o modelo.
- **Remoção de Características Correlacionadas:** Remove características altamente correlacionadas para evitar redundância.

4.1.2 Construção do Modelo

- **Parâmetros do Modelo:** Define os parâmetros do LightGBM, como 'min_child_samples', 'max_depth', 'num_leaves' e 'n_estimators'.

4.1.3 Treinamento

- **Dados de Treino:** Treina o modelo nos dados de treino utilizando o método 'fit' do LightGBM.

4.2 Avaliação do Modelo

Descrição: Avalia o desempenho do modelo treinado em termos de várias métricas de classificação.
Processo:

- **Métricas de Treinamento:** Avalia o desempenho do modelo nos dados de treino utilizando AUC, KS, Log Loss, Accuracy, Precision, Recall, F1 Score, F1 Score Kaggle.
- **Métricas de Teste:** Avalia o desempenho do modelo nos dados de teste utilizando as mesmas métricas.

4.3 Plotagem da Curva ROC

Descrição: Gera e plota a Curva ROC para avaliar a capacidade do modelo de discriminar entre as classes.
Processo:

- **Previsão de Probabilidades:** Usa o modelo para prever as probabilidades nos dados de teste.
- **Cálculo da Curva ROC:** Calcula as taxas de verdadeiro positivo (TPR) e falso positivo (FPR) para diferentes limiares.
- **Plotagem da Curva:** Plota a curva ROC e calcula a AUC.

4.4 Plotagem da Importância das Características

Descrição: Gera e plota a importância das características para identificar quais características são mais relevantes para o modelo.
Processo:

- **Importância das Características:** Calcula a importância das características utilizando o modelo treinado.
- **Plotagem da Importância:** Plota a importância das características em um gráfico de barras.

4.5 Seleção de Características e Otimização de Hiperparâmetros

4.5.1 MRMR (Minimum Redundancy Maximum Relevance)

Descrição: Seleciona características que são altamente relevantes para o alvo e possuem baixa redundância entre si.
Processo:

- **Relevância:** Calcula a relevância de cada característica individual em relação ao alvo.
- **Redundância:** Avalia a redundância entre as características.
- **Seleção:** Escolhe características que maximizam a informação útil.

4.5.2 Boruta

Descrição: Utiliza o algoritmo Boruta para selecionar todas as características relevantes para o alvo.
Processo:

- **Modelo Base:** Usa um modelo de floresta aleatória como base.
- **Iteração:** Itera adicionando e removendo características até encontrar o conjunto de características relevantes.

4.5.3 Forward Feature Selection

Descrição: Seleciona características de forma incremental, adicionando uma característica por vez com base em sua contribuição para o desempenho do modelo.
Processo:

- **Inicialização:** Começa com um conjunto vazio de características.
- **Iteração:** Adiciona a característica que mais melhora a métrica de avaliação do modelo.
- **Parada:** Para quando não há mais melhoria significativa ou quando atinge o tempo máximo.

4.5.4 Full Optuna

Descrição: Utiliza a biblioteca Optuna para otimizar os hiperparâmetros do modelo LightGBM.
Processo:

- **Definição da Função Objetivo:** Define a função objetivo que será maximizada pela Optuna.
- **Espaço de Busca:** Especifica os hiperparâmetros a serem otimizados e seus respectivos intervalos.
- **Otimização:** Executa a busca de hiperparâmetros utilizando amostragem bayesiana.

4.6 Previsão

Descrição: Gera previsões binárias e probabilísticas para novos dados.
Processo:

- **Previsão:** Utiliza o modelo treinado para gerar previsões de probabilidade e aplica um limiar para obter previsões binárias.

4.7 Trecho do Código que faz a Previsão

```
def predict(self, X):
    if self.model is None:
        raise ValueError("O modelo não foi treinado ainda.")
    return self.model.predict(X[self.features])

def predict_proba(self, X):
    if self.model is None:
        raise ValueError("O modelo não foi treinado ainda.")
    return self.model.predict_proba(X[self.features])[:, 1]
```

4.8 Modelagem e Treinamento com Redes Neurais

Descrição: Implementa um pipeline completo de treinamento de um modelo de rede neural, desde o pré-processamento dos dados até a avaliação do modelo.
Processo:

- **Pré-processamento:** Normaliza ou padroniza os dados, divide os dados em conjuntos de treino e teste.
- **Construção do Modelo:** Define a arquitetura da rede neural com várias camadas densas e dropout para evitar overfitting.
- **Compilação:** Configura o modelo com a função de perda 'binary_crossentropy', otimizador 'Adam' e métrica 'accuracy'.
- **Treinamento:** Treina o modelo com os dados de treino, validando com os dados de teste.

4.8.1 Pré-processamento

- **Normalização:** Aplica a normalização (StandardScaler ou MinMaxScaler) aos dados de treino e teste.
- **Divisão dos Dados:** Divide os dados em conjuntos de treino e teste, utilizando 'train_test_split' com estratificação.

4.8.2 Construção do Modelo

- **Arquitetura:** Define uma rede neural sequencial com camadas densas e camadas de dropout.
- **Função de Ativação:** Utiliza a função de ativação 'relu' para as camadas ocultas e 'sigmoid' para a camada de saída.
- **Dropout:** Inclui camadas de dropout para prevenir overfitting.

4.8.3 Compilação

- **Função de Perda:** Utiliza a função de perda 'binary_crossentropy'.
- **Otimizador:** Utiliza o otimizador 'Adam' com uma taxa de aprendizado especificada.
- **Métrica:** Avalia o desempenho do modelo utilizando a métrica 'accuracy'.

4.8.4 Treinamento

- **Dados de Treino:** Treina o modelo nos dados de treino com um número especificado de épocas e tamanho de lote.
- **Validação:** Valida o modelo utilizando os dados de teste durante o treinamento.

4.9 Avaliação do Modelo

Descrição: Avalia o desempenho do modelo treinado em termos de várias métricas de classificação.
Processo:

- **Métricas de Treinamento:** Avalia o desempenho do modelo nos dados de treino utilizando AUC, KS, Log Loss, Accuracy, Precision, Recall, F1 Score, F1 Score Kaggle.
- **Métricas de Teste:** Avalia o desempenho do modelo nos dados de teste utilizando as mesmas métricas.

4.10 Plotagem da Curva ROC

Descrição: Gera e plota a Curva ROC para avaliar a capacidade do modelo de discriminar entre as classes.
Processo:

- **Previsão de Probabilidades:** Usa o modelo para prever as probabilidades nos dados de teste.
- **Cálculo da Curva ROC:** Calcula as taxas de verdadeiro positivo (TPR) e falso positivo (FPR) para diferentes limiares.
- **Plotagem da Curva:** Plota a curva ROC e calcula a AUC.

4.11 Plotagem da Curva de Perda

Descrição: Plota a curva de perda durante o treinamento e a validação.
Processo:

- **Histórico do Treinamento:** Obtém os valores de perda durante o treinamento e a validação.
- **Plotagem da Curva de Perda:** Plota a perda em função das épocas de treinamento.

4.12 Previsão

Descrição: Gera previsões binárias e probabilísticas para novos dados.
Processo:

- **Normalização dos Novos Dados:** Aplica o mesmo escalonamento usado no treinamento.
- **Previsão:** Gera previsões de probabilidades e aplica um limiar para obter previsões binárias.

4.13 Trecho de Código que faz a Previsão

```
def predict(self, new_data, threshold=0.5):
    if self.model is None:
        raise ValueError("O modelo não foi treinado ainda.")
    new_data_scaled = self.scaler.transform(new_data)
    probabilities = self.model.predict(new_data_scaled)
    binary_predictions = (probabilities > threshold).astype(int)
    return probabilities, binary_predictions
```

5 O modelo escolhido

Utilizando as métricas mencionadas para avaliar o desempenho dos modelos, notamos que todos os modelos tiveram um resultado muito próximo. A combinação LightGBM com MRMR demonstrou as maiores métricas.

5.1 Justificativa do Uso de Métricas no Desafio Kaggle Instacart Market Basket Analysis

5.1.1 AUC (Area Under the ROC Curve)

Descrição: A AUC mede a capacidade do modelo de distinguir entre classes. É a área sob a curva ROC, que plota a taxa de verdadeiros positivos (TPR) contra a taxa de falsos positivos (FPR).
Relevância: No contexto do desafio, é importante medir quão bem o modelo pode distinguir entre produtos que serão recomprados e os que não serão. A AUC fornece uma visão geral dessa habilidade sem depender de um limite específico de decisão.

5.1.2 KS (Kolmogorov-Smirnov Statistic)

Descrição: A estatística KS mede a diferença máxima entre as distribuições acumuladas de probabilidades preditas para as classes positiva e negativa.
Relevância: É útil para avaliar a separação entre as duas classes. Uma alta estatística KS indica uma boa distinção entre produtos recomprados e não recomprados.

5.1.3 Log Loss

Descrição: Log Loss mede a incerteza das previsões do modelo. Penaliza predições incorretas mais severamente quanto mais confiantes elas são.
Relevância: No desafio, onde a precisão das probabilidades preditas é crucial, o Log Loss garante que o modelo não apenas classifique corretamente, mas também atribua probabilidades que reflitam a incerteza real.

5.1.4 Accuracy

Descrição: Accuracy é a proporção de predições corretas sobre o total de instâncias.
Relevância: É uma métrica básica e fácil de interpretar que dá uma ideia geral de quão frequentemente o modelo está correto. No entanto, pode ser menos útil em datasets desbalanceados como o do desafio.

5.1.5 Precision

Descrição: Precision é a proporção de verdadeiros positivos sobre o total de positivos preditos (TP / (TP + FP)).
Relevância: Alta precision é crucial quando o custo de falsos positivos é alto. No contexto do desafio, significa que quando o modelo prevê uma recomprar, é provável que seja correto.

5.1.6 Recall

Descrição: Recall é a proporção de verdadeiros positivos sobre o total de positivos reais (TP / (TP + FN)).
Relevância: Alta recall é importante quando o custo de falsos negativos é alto. No desafio, significa que o modelo está capturando a maioria dos produtos que serão recomprados.

5.1.7 F1 Score

Descrição: O F1 Score é a média harmônica de precision e recall.
Relevância: Combina os benefícios de precision e recall, sendo especialmente útil quando há um trade-off entre as duas. No desafio, o F1 Score ajuda a balancear a precisão e a abrangência das previsões de recompras.

5.1.8 F1 Score Kaggle (usando user_id para agrupar)

Descrição: Calcula o F1 Score para conjuntos de itens por usuário. Considera a precisão e o recall para cada usuário individualmente e então calcula a média.
Relevância: Reflete a performance do modelo em nível de usuário, que é o foco do desafio. As previsões são feitas para cada usuário, e a métrica considera a agregação dessas previsões, proporcionando uma medida mais realista e prática da performance do modelo no contexto de uso real.

5.2 Resultados para o modelo escolhido

Neste relatório, avaliamos o desempenho de um modelo de aprendizado de máquina treinado para o desafio do Kaggle: Instacart Market Basket Analysis. Utilizamos diversas métricas de avaliação e interpretamos a importância das features utilizadas no modelo.

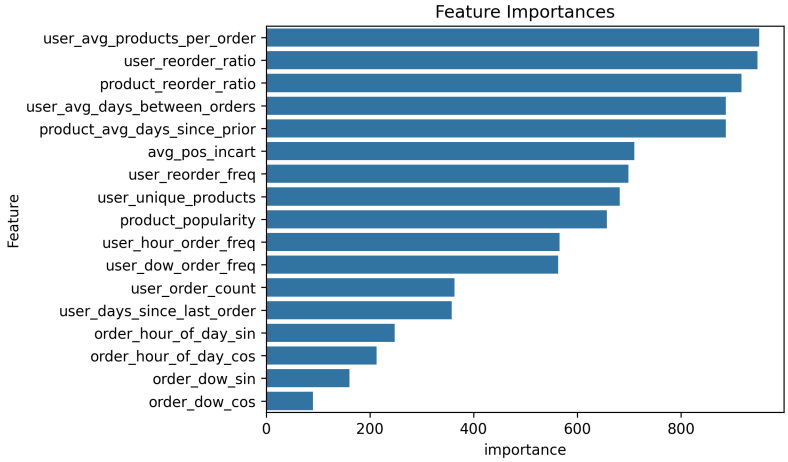

Figura 11: Feature importantes

Observamos que a média de produtos por pedido por usuário (*user_avg_products_per_order*) é a feature mais importante, seguida pela taxa de recompra do usuário (*user_reorder_ratio*) e a proporção de recompra do produto (*product_reorder_ratio*).

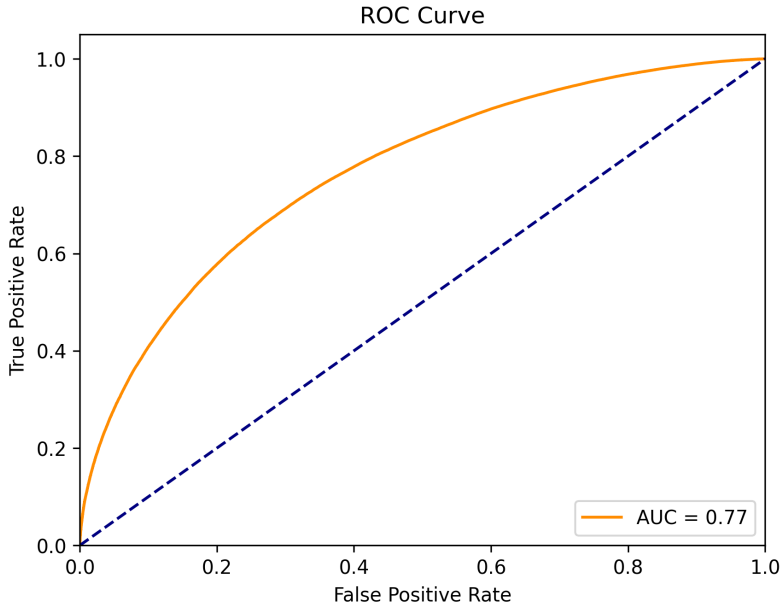

Figura 12: Curva ROC

A Curva ROC é uma representação gráfica que mostra o desempenho do modelo em termos de taxa de verdadeiros positivos versus taxa de falsos positivos.

Métrica	Valor
AUC	0.7668
KS	0.3924
Log Loss	0.5597
Accuracy	0.7076
Precision	0.7302
Recall	0.8120
F1 Score	0.7689
F1 Score Kaggle	0.7021

Tabela 21: Métricas de Avaliação do Modelo

A AUC (área sob a curva ROC) mede a capacidade do modelo em distinguir entre as classes positivas e negativas. Com um valor de 0.7668, nosso modelo apresenta uma boa capacidade discriminativa. A métrica KS avalia a diferença máxima entre as distribuições acumuladas de classes positivas e negativas. Um valor de 0.3924 indica uma boa separação entre as classes previstas. O Log Loss mede a incerteza das previsões probabilísticas do modelo. Com um valor de 0.5597, o modelo mostra uma incerteza moderada nas previsões. A acurácia indica a proporção de previsões corretas feitas pelo modelo. Com um valor de 0.7076, mais de 70% das previsões estão corretas. A precisão mede a proporção de previsões positivas corretas. Com um valor de 0.7302, o modelo tem uma boa taxa de precisão. O recall mede a capacidade do modelo em identificar corretamente as classes positivas. Com um valor de 0.8120, o modelo apresenta uma alta taxa de recall. O F1 Score é a média harmônica entre precisão e recall. Com um valor de 0.7689, o modelo apresenta um bom equilíbrio entre precisão e recall. O F1 Score Kaggle considera a avaliação por usuário, agrupando os resultados por *user_id*. Com um valor de 0.7021, o modelo apresenta um bom desempenho na identificação de produtos reordenados para cada usuário.

6 Conclusões

O modelo desenvolvido apresentou um desempenho robusto em várias métricas de avaliação. Através da análise da importância das features, identificamos quais características relacionadas ao comportamento de compra do usuário, como a média de produtos por pedido e a taxa de recompra, são cruciais para prever os pedidos futuros. As métricas como AUC, KS e F1 Score indicam que o modelo tem uma boa capacidade de discriminação e equilíbrio entre precisão e recall. O F1 Score específico do Kaggle também mostra que o modelo é eficaz em agrupar previsões por usuário, essencial para o desafio. Com base nesses resultados, o modelo está bem posicionado para uma boa performance no desafio do Kaggle.

7 Referências

Referências

[1] Saeed Asagar, *Instacart Market Basket Analysis - Part 1: Introduction & EDA*, Medium, 2020. `https://asagar60.medium.com/instacart-market-basket-analysis-part-1-introduction-eda-b08fd8250502`. Acesso em: 22 de julho de 2024.

[2] Colleen M. Farrelly, *Instacart Market Basket Analysis*, Medium, Kaggle Blog, 2017. `https://medium.com/kaggle-blog/instacart-market-basket-analysis-feda2700cded`. Acesso em: 22 de julho de 2024.

[3] Sultan Khan, *Kaggle Instacart Market Basket Analysis*, Medium, Geek Culture, 2022. `https://medium.com/geekculture/kaggle-instacart-market-basket-analysis-8bfbbf5f2efb`. Acesso em: 22 de julho de 2024.

[4] Susan Li, *Instacart Market Basket Analysis Part 3: Which sets of products should be recommended to shoppers?*, Medium, Towards Data Science, 2019. `https://towardsdatascience.com/`

instacart-market-basket-analysis-part-3-which-sets-of-products-should-be-recommended-to-shoppers-96517
Acesso em: 22 de julho de 2024.

[5] Kaggle, *Instacart Market Basket Analysis*, https://www.kaggle.com/competitions/instacart-market-basket-analysis. Acesso em: 22 de julho de 2024.

[6] Cookiecutter Data Science, https://cookiecutter-data-science.drivendata.org/. Acesso em: 22 de julho de 2024.

[7] OpenAI, *ChatGPT*, https://www.openai.com/chatgpt. Acesso em: 22 de julho de 2024.

[8] Google, *Google Colaboratory*, https://colab.research.google.com/. Acesso em: 22 de julho de 2024.

[9] Overleaf, *Overleaf: Online LaTeX Editor*, https://www.overleaf.com/. Acesso em: 22 de julho de 2024.