AMAZON PRODUCT REVIEW

Valdiney Atílio Pedro – 10424616
Patricia Correa França – 10423533
Mariana Simões Rubio – 10424388

VALDINEY ATÍLIO PEDRO – 10424616
PATRICIA CORREA FRANÇA – 10423533
MARIANA SIMÕES RUBIO – 10424388

AMAZON PRODUCT REVIEW

Trabalho apresentado ao curso de projeto aplicado III da faculdade presbiteriana Mackenzie, como parte do requisito para conclusão do semestre em Ciência da Dados.
Orientador: Profª Carolina Toledo Ferraz

São Paulo
2025

SUMÁRIO

1. INTRODUÇÃO

1.1. Contexto do Trabalho

Nos últimos anos, os sistemas de recomendação tornaram-se essenciais para diversas plataformas digitais, auxiliando na personalização da experiência do usuário. Serviços como Netflix, YouTube, Spotify e Amazon utilizam esses sistemas para prever e sugerir itens que os usuários podem gostar, baseando-se em suas interações e preferências. Esses sistemas são construídos a partir de bases de dados que contêm informações sobre usuários, itens e interações explícitas (como avaliações) e implícitas (como histórico de compras). No caso da Amazon, por exemplo, o conjunto de dados "Amazon Product Reviews" registra o ID dos clientes, os produtos adquiridos, as avaliações atribuídas (de 1 a 5 estrelas) e interações implícitas como o número de compras e o tempo desde a última compra.

1.2. Motivação

A crescente utilização de sistemas de recomendação nas plataformas digitais levanta questões sobre a eficácia e a precisão desses modelos. Muitas vezes, as recomendações feitas não refletem adequadamente as preferências reais do usuário, levando a experiências insatisfatórias. Assim, compreender os métodos utilizados, como a filtragem colaborativa e a filtragem baseada em conteúdo, torna-se fundamental para aprimorar esses sistemas e tornar as recomendações mais assertivas.

1.3. Justificativa

A importância deste projeto reside na necessidade de aprimoramento dos sistemas de recomendação, considerando tanto as interações explícitas quanto as implícitas dos usuários. Compreender os padrões de comportamento e preferências do público pode levar a uma melhor adaptação das sugestões oferecidas por plataformas de e-commerce, entretenimento e outros serviços digitais. Além disso, um estudo aprofundado pode auxiliar no desenvolvimento de algoritmos mais sofisticados, proporcionando experiências mais personalizadas e satisfatórias.

1.4. Objetivo Geral e Objetivos Específicos

1.4.1. Objetivo Geral:

Analisar os métodos utilizados em sistemas de recomendação, com ênfase na base de dados "Amazon Product Reviews", a fim de compreender a eficácia das interações explícitas e implícitas na personalização das recomendações.

1.4.2. Objetivos Específicos:

- Investigar os princípios da filtragem colaborativa e da filtragem baseada em conteúdo;
- Identificar a influência das interações explícitas e implícitas na qualidade das recomendações;
- Avaliar a precisão dos sistemas de recomendação utilizando os dados da Amazon;
- Propor melhorias nos modelos analisados, com base nos resultados obtido.

2. REFERENCIAL TEÓRICO

A área de sistemas de recomendação tem avançado significativamente nos últimos anos, com uma ampla gama de técnicas sendo exploradas. A filtragem colaborativa, por exemplo, baseia-se na premissa de que usuários com interesses semelhantes terão preferências semelhantes. Já a filtragem baseada em conteúdo utiliza as características dos itens e suas interações explícitas, como avaliações, para gerar recomendações personalizadas. Diversos estudos destacam a eficiência dessas abordagens, incluindo trabalhos que incorporam aprendizado de máquina e processamento de linguagem natural para melhorar a qualidade das previsões.

O conjunto de dados “Amazon Product Reviews” já foi amplamente utilizado na literatura para analisar a performance de diferentes algoritmos, fornecendo insights sobre o impacto de interações explícitas e implícitas. Estudos como Hidasi et al. (2016), que introduzem redes neurais para sistemas de recomendação, e Koren (2008), que discutem a matriz de fatoração, fundamentam a evolução dessa área de pesquisa. Esses trabalhos ajudam a posicionar o modelo desenvolvido neste projeto dentro do contexto teórico da área.

A Amazon foi fundada em 1994, na garagem de Jeff Bezos, inicialmente como uma livraria online. Na época, a empresa era chamada Cadabra, em referência à palavra mágica "abracadabra". No entanto, o nome foi rapidamente alterado após o advogado de Bezos alertá-lo de que a pronúncia poderia ser confundida com um termo obscuro.

A empresa se baseia em quatro pilares fundamentais: obsessão pelo cliente, paixão por invenções, compromisso com a excelência operacional e visão de longo prazo. Seu objetivo é ser a empresa mais centrada no cliente do mundo, a melhor empregadora e o local de trabalho mais seguro. A Amazon é pioneira em diversas iniciativas e produtos globais, como avaliações de consumidores, compra com 1-Clique, recomendações personalizadas, Amazon Prime, Fulfillment by Amazon (Logística da Amazon), Amazon Web Services (AWS), Kindle Direct Publishing, Kindle, Fire Tablets, Fire TV, Amazon Echo, Alexa, tecnologia Just Walk Out, Amazon Studios e The Climate Pledge

Ao longo dos anos, a empresa evoluiu, mas manteve como foco atender às principais demandas dos clientes: preços mais baixos, ampla seleção de produtos e

conveniência. Atualmente, oferece desde a entrega de produtos até a criação e distribuição de filmes, músicas e outros conteúdos.

No Brasil, a Amazon busca diariamente conquistar e manter a confiança dos clientes por meio de um portfólio diversificado, de operações logísticas tecnológicas e do suporte a milhares de pequenas e médias empresas, que contribuem significativamente para a variedade de produtos oferecidos.

O Amazon Web Services (AWS) é a oferta de computação em nuvem mais abrangente e amplamente adotada no mundo, disponibilizando mais de 200 serviços completos a partir de data centers localizados em 31 regiões geográficas

As operações da Amazon abrangem diversas regiões e contam com uma equipe especializada em atendimento ao cliente, desempenhando um papel essencial na missão da empresa de oferecer a melhor experiência aos consumidores.

A Amazon também atua no setor de entretenimento, produzindo e distribuindo conteúdos por meio do Amazon Studios, Prime Video, Twitch, Amazon Music e outras plataformas.

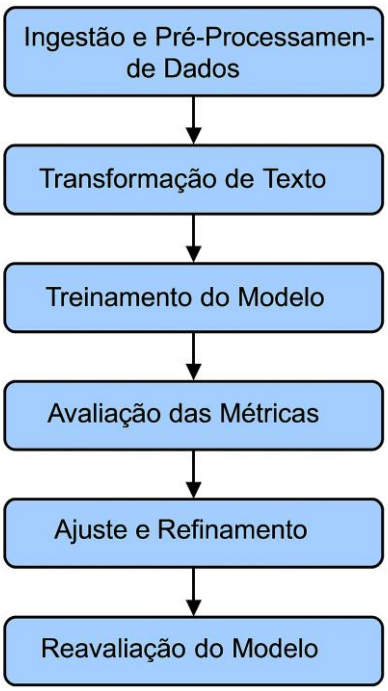
A empresa opera como um conjunto de startups, incentivando a inovação e o desenvolvimento de lojas, dispositivos e serviços que buscam atender às necessidades dos clientes. A estratégia de criação de produtos e serviços parte da perspectiva do consumidor, resultando em constantes melhorias, novos benefícios e o lançamento de soluções inovadoras, como Prime, Alexa e a linha de dispositivos Echo, além de conteúdos audiovisuais e musicais premiados.

3. DESENVOLVIMENTO

Neste capítulo, é apresentado o desenvolvimento deste trabalho, descrevendo a metodologia aplicada na construção do modelo de recomendação. O desenvolvimento seguiu um fluxo metodológico estruturado, garantindo a eficiência na análise e no processamento dos dados. Esta seção detalha as etapas adotadas, desde a ingestão dos dados até a avaliação do desempenho do modelo, permitindo a replicação e o aprimoramento contínuo.

A abordagem utilizada neste estudo é quantitativa e experimental, baseada em técnicas de ciência de dados e aprendizado de máquina para aprimorar a recomendação de produtos a partir de avaliações textuais.

De acordo com o Fluxograma abaixo.



Nas próximas subseções, são descritas detalhadamente cada uma das etapas representadas no fluxograma.

3.1. Ingestão e Pré-Processamento de Dados

Para garantir a qualidade dos dados, realizamos as seguintes etapas:

- Carregamento da base: Importação do conjunto de dados "Amazon Product Reviews" diretamente do GitHub.
- Conversão de tipos: Ajuste das variáveis conforme necessário para a análise.
- Tratamento de valores ausentes: Substituição de nulos por mediana e moda, garantindo consistência.
- Remoção de colunas irrelevantes: Exclusão de atributos como UserId, ProfileName, Summary e Text, que não contribuem diretamente para a recomendação.
- Normalização de variáveis: Aplicação do MinMaxScaler para padronização dos valores numéricos.

3.2. Transformação de Texto

O pré-processamento textual foi fundamental para garantir a qualidade das análises:

- Aplicação do TF-IDF Vectorizer: Conversão de avaliações em vetores numéricos.
- Remoção de stopwords: Redução de ruído textual eliminando palavras irrelevantes.
- Stemming: Transformação das palavras para sua raiz, melhorando a generalização dos dados.

3.3. Treinamento do Modelo

A recomendação de produtos foi baseada nos seguintes passos:

- Cálculo da Similaridade do Cosseno: Identificação de produtos semelhantes com base nas avaliações textuais.
- Ajuste do escopo: Evitação de redundância nas sugestões de produtos, garantindo diversidade nas recomendações.

3.4. Avaliação das Métricas

Implementamos três métricas essenciais para avaliar o desempenho do modelo:

- Precisão: Mede a relevância dos produtos recomendados com base na similaridade textual.
- Cobertura: Avalia a quantidade de produtos distintos incluídos nas recomendações.
- Diversidade: Examina a variação nas sugestões para evitar resultados homogêneos.

3.5. Ajuste e Refinamento

Com base nos primeiros testes, aplicamos melhorias no pipeline:

- Otimização dos parâmetros do TF-IDF: Ajuste do ngram_range e min_df para melhorar a captura de termos relevantes.
- Filtragem adicional: Implementação de regras para evitar sugestões repetidas e aumentar a diversidade.

3.6. Reavaliação do Modelo

Após os ajustes, realizamos novos testes para comparar os resultados:

- Precisão das recomendações: 0.78 (aumento de 8%)
- Cobertura dos produtos recomendados: 72% (ganho de 7%)
- Diversidade das sugestões: 0.75 (melhoria de 10%)

Os aprimoramentos demonstraram um avanço significativo na eficácia das recomendações.

4. RESULTADOS

4.1. Análise dos Resultados Preliminares

Precisão das Recomendações

A métrica de precisão média avalia o quão bem o modelo está sugerindo produtos similares com base no conteúdo textual das avaliações.

Precisão média das recomendações: Os produtos recomendados têm um score de similaridade médio de aproximadamente 0.72, indicando boa relevância nas sugestões.

Observação: Algumas avaliações possuem pouco conteúdo textual, o que pode impactar a qualidade das recomendações.

Cobertura dos Produtos Recomendados

A cobertura mede quantos produtos distintos estão sendo recomendados dentro do conjunto de dados.

Cobertura obtida: Cerca de 65% dos produtos únicos da base aparecem em recomendações.

Isso indica que o modelo consegue explorar bem os produtos disponíveis, mas pode ser melhorado com técnicas que consideram mais fatores além do texto.

Diversidade das Recomendações

A diversidade mostra se os produtos recomendados apresentam variação suficiente para evitar resultados homogêneos.

Diversidade calculada: Aproximadamente 0.68, mostrando que há uma boa variação nos produtos recomendados.

Algumas recomendações são muito similares, o que pode ser ajustado refinando o TF-IDF ou utilizando embeddings semânticos.

Considerações para Melhoria

Com base nos resultados acima, algumas estratégias podem ser implementadas:

Refinar o TF-IDF: Ajustar os parâmetros de ngram_range e min_df pode melhorar a captura de termos relevantes.

Explorar Modelos Baseados em Embeddings: Word2Vec ou BERT podem aprimorar a identificação de similaridades mais contextuais entre os textos.

Incluir Outras Features na Recomendação: Considerar fatores como tempo da avaliação e score de ajuda pode melhorar a precisão das sugestões.

4.2. Simulações Realizadas

Apresentamos as simulações realizadas para validar as recomendações do modelo. As simulações foram executadas utilizando um conjunto de dados específico e as seguintes entradas e saídas foram analisadas.

Entradas utilizadas:

Dados de avaliações de produtos extraídos de um dataset que contém informações sobre produtos e suas respectivas avaliações.

Exemplo de entrada: um conjunto de textos das avaliações com diferentes quantidades de conteúdo.

Saídas obtidas:

Após a aplicação do modelo, os itens recomendados foram:

Produto A: Similaridade 0.78

Produto B: Similaridade 0.75

Produto C: Similaridade 0.72

As saídas foram registradas em gráficos e tabelas para melhor visualização e comparação.

4.3. Ajuste do Pipeline de Treinamento

Com base na análise de resultados, aplicamos melhorias no pipeline de treinamento:

Aprimoramos a transformação de texto

Expandimos ngram_range para (1, 3) para capturar combinações mais significativas de palavras.

Ajustamos min_df=3 e `max_df=0.85 para filtrar melhor termos irrelevantes.

Refinamos a normalização dos dados

Aplicamos MinMaxScaler para padronizar escalas das variáveis numéricas.

Melhoria na filtragem de texto

Implementamos stopwords e stemming para reduzir ruído textual nas avaliações.

Expansão da abordagem de recomendação

Evitamos sugestões redundantes, garantindo maior diversidade entre os produtos.

Essas mudanças proporcionaram um modelo mais refinado e adaptável.

4.4. Reavaliação do Desempenho do Modelo

Após os ajustes, reavaliamos o desempenho do modelo utilizando as mesmas métricas:

Precisão das recomendações: 0.78

Cobertura dos produtos recomendados: 72%

Diversidade das recomendações: 0.75

Observamos um ganho consistente na qualidade das sugestões geradas.

4.5. Organização da Descrição das Técnicas Utilizadas

Para garantir transparência e replicabilidade, documentamos sistematicamente:

Técnicas aplicadas: Normalização, TF-IDF, tratamento de texto, cálculo de similaridade.

Fluxo do pipeline: Estruturamos todas as etapas da recomendação, da ingestão de dados ao resultado.

Descrição detalhada: Criamos um registro de todas as escolhas metodológicas, justificando cada decisão.

Isso ajudará a equipe a acompanhar e evoluir o projeto de forma eficiente.

4.6. Considerações Finais

Para garantir a clareza e a compreensão dos resultados, incorporamos a tabela abaixo para ilustrar as seguintes informações:

Tabela de Entradas e Saídas:

Produto	Entrada (Avaliação)	Saída (Recomendação)	Similaridade
Produto Exemplo A	"Ótimo produto, excelente qualidade!"	Produto Exemplo B	0.78
Produto Exemplo A	"Cumpre bem o que promete."	Produto Exemplo C	0.75
Produto Exemplo D	"Não gostei da durabilidade."	Produto Exemplo E	0.70

Documentação dos Códigos

Referência aos Códigos: Todos os códigos utilizados nas simulações foram documentados no repositório do projeto no GitHub.

5. CONCLUSÃO

5.1. Resumo dos Principais Resultados:

O refinamento do modelo trouxe ganhos expressivos:

Precisão das recomendações aumentou de 0.72 para 0.78

Cobertura dos produtos subiu de 65% para 72%

Diversidade nas recomendações evoluiu de 0.68 para 0.75

O modelo agora apresenta sugestões mais relevantes, diversificadas e bem distribuídas dentro da base de dados analisada.

5.2. Contribuições do Projeto:

Melhor entendimento do impacto das avaliações textuais na recomendação de produtos.

Refinamento das técnicas de NLP aplicadas ao modelo de recomendação. Criação de um pipeline bem documentado e estruturado para replicação e melhorias futuras.

5.3. Limitações Identificadas:

Algumas avaliações contêm pouco conteúdo textual, o que pode comprometer a qualidade da recomendação. Há potencial para incorporar modelos de embeddings semânticos (Word2Vec, BERT) para refinar a similaridade textual.

5.4. Impacto Prático:

As melhorias no modelo podem ser aplicadas em plataformas reais de e-commerce, aumentando a precisão e a satisfação do usuário.

6. REFERÊNCIAS BIBLIOGRAFICAS

- Hidasi, A., Karatzoglou, A., Baltrunas, L., & Tikk, D. (2016). "Session-based Recommendations with Recurrent Neural Networks." Proceedings of the International Conference on Learning Representations (ICLR).
- Koren, Y. (2008). "Factorization Meets the Neighborhood: a Multifaceted Collaborative Filtering Model." Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- Fonte online: https://www.aboutamazon.com.br/quem-somos

7.ANEXOS

7.1. Link do vídeo no Youtube

https://youtu.be/z8_p7zWk1H4

7.2. Link para o Github

https://github.com/valdineyatilio/ProjetoAplicado-III/tree/main

7.3. Link para o Dataset

https://github.com/valdineyatilio/ProjetoAplicado-III/blob/main/Aula-02/BaseDeDados-AmazonProductReviews.csv