Relatório de Processamento e Rotulagem de Dados para Detecção de Pirataria de Produtos HP

Lancelot Chagas Rodrigues / 554707

Ana Carolina Martins da Silva / 555762

Kauan Alves Batista / 555082

---

1. Introdução

Este documento apresenta as etapas de tratamento de dados, o processo de feature engineering e a rotulagem heurística que conduzimos com o intuito de construir um dataset estruturado. Este dataset servirá como alicerce para o desenvolvimento futuro de uma solução de Machine Learning, visando a identificação de anúncios potencialmente fraudulentos ou de pirataria de produtos HP em plataformas de e-commerce. O foco aqui é demonstrar a transformação dos dados brutos coletados em um conjunto de informações enriquecido e preparado para modelagem, refletindo as decisões e aprendizados do grupo durante esta jornada.

2. Coleta de Dados (Web Scraping)

Os dados primários para este projeto foram obtidos através de um processo de Web Scraping, que se concentrou na varredura da plataforma de e-commerce Mercado Livre em busca de anúncios de produtos HP. Nesta etapa, foram coletadas informações como descrições de produtos, preços, nomes de vendedores, avaliações, entre outros atributos relevantes para a análise.

É importante destacar que a metodologia detalhada, as ferramentas empregadas e os desafios encontrados durante a fase de coleta de dados foram extensivamente documentados em um projeto anterior. Este trabalho de scraping foi desenvolvido durante o Challenge Sprint 1 de RPA. Para um aprofundamento completo nesta etapa, recomendamos a consulta ao repositório dedicado: https://github.com/lancelot-rodrigues/HP-Scraping.

3. Features (Variáveis) do Dataset

O dataset utilizado neste trabalho é composto por features originais (diretamente coletadas) e por novas features desenvolvidas através do processo de feature engineering. O objetivo do feature engineering foi enriquecer a análise e fornecer subsídios mais robustos para o futuro modelo de Machine Learning.

3.1. Features Originais (Coletadas do CSV de Entrada)

As seguintes features foram consideradas como dados de entrada, provenientes do arquivo ml_produtos_hp_processado_amostra.csv (cuja estrutura original se assemelha à lista abaixo, com nomes adaptados durante o processo):

- link_anuncio: URL do anúncio do produto.
- titulo: Título do anúncio conforme exibido no Mercado Livre.
- preco: Preço de venda do produto no anúncio.
- vendedor: Nome do vendedor ou da loja que publicou o anúncio.
- avaliacao_nota: Nota média de avaliação do produto ou vendedor, se disponível.
- avaliacao_numero: Quantidade de avaliações recebidas pelo produto ou vendedor.
- descrição: Texto descritivo do produto fornecido no anúncio.
- marca: Marca do produto (espera-se "HP" ou variações).
- modelo: Modelo específico do produto HP.
- cor: Cor do produto, se aplicável.
- tipo_cartucho: Tipo específico do cartucho (ex: "Tinta", "Toner"), utilizado como principal coluna de categorização.
- originalidade: Alegação do vendedor sobre a originalidade do produto (ex: "Original", "Compatível").
- rendimento_paginas: Estimativa de rendimento em páginas (comum para cartuchos/toners).
- impressoras_compativeis: Lista de modelos de impressoras compatíveis.
- volume_ml_ou_peso_g: Volume (para tintas) ou peso (para toners).
- estado_produto: Condição do produto (ex: "Novo", "Usado").
- nota_qualidade: Uma possível avaliação numérica da qualidade do anúncio ou produto (presente no dataset de entrada).
- justificativa_qualidade: Uma possível justificativa textual associada à nota_qualidade (presente no dataset de entrada).

3.2. Novas Features (Resultado de Feature Engineering)

Para aumentar o poder preditivo do dataset, o trabalho de feature engineering resultou na criação das seguintes features:

- **seller_trust_level (Nível de Confiança do Vendedor):**
    - **Descrição:** Um score numérico (1 para Suspeito, 2 para Neutro/Avaliar Internamente, 3 para Confiável Alto) atribuído a cada vendedor.
    - **Lógica:** Esta feature foi construída a partir de uma pesquisa manual extensiva realizada pelo grupo. Foram consultadas fontes de reputação online, como o site Reclame Aqui, e considerada a percepção geral sobre os vendedores listados. Criamos um mapeamento que associa nomes de vendedores a esses níveis de confiança. É crucial enfatizar que esta é uma classificação inicial, baseada no conhecimento e esforço de pesquisa do grupo até o momento. Em futuras versões do projeto, esta feature será significativamente enriquecida e validada com informações e evidências mais robustas, que poderão ser disponibilizadas pela HP.
- **preco_medio_categoria (Preço Médio da Categoria):**
    - **Descrição:** O preço médio dos produtos dentro de uma mesma categoria (definida pela coluna tipo_cartucho).
    - **Lógica:** Calculada agrupando-se os produtos pela coluna tipo_cartucho e computando a média aritmética dos seus respectivos preco (após limpeza). Para garantir a robustez deste cálculo, preços individuais que excediam um limite de plausibilidade (MAX_PLAUSIBLE_PRICE_INDIVIDUAL, ex: R$ 20.000) foram convertidos para NaN antes da média. Adicionalmente, as próprias médias de categoria calculadas foram submetidas a um "cap" (MAX_PLAUSIBLE_PRICE_AVERAGE, ex: R$ 5.000); médias que excedessem este valor ou fossem não positivas foram também convertidas para NaN. Este processo iterativo foi necessário para lidar com valores de preço atípicos ou mal formatados que poderiam distorcer as médias.
- **desvio_preco_media_categoria (Desvio do Preço em Relação à Média da Categoria):**
    - **Descrição:** O desvio percentual do preço de um produto específico em relação ao preco_medio_categoria correspondente.
    - **Lógica:** Calculado pela fórmula: (preco_do_produto - preco_medio_categoria) / preco_medio_categoria. Um valor

negativo indica que o produto está mais barato que a média da categoria, enquanto um positivo indica que está mais caro. Casos de divisão por zero ou resultantes de preco_medio_categoria NaN foram tratados para que o desvio se tornasse 0, evitando erros e mantendo a interpretabilidade.

4. Processo de Rotulagem Heurística (Criação da Feature Alvo label_heuristico_calculado)

O objetivo central desta fase foi criar uma feature alvo inicial, label_heuristico_calculado, para classificar os anúncios. Dada a ausência de um dataset previamente rotulado, optamos por um sistema de regras heurísticas.

Decisões e Lógica da Rotulagem:

A rotulagem foi implementada através de um conjunto de regras condicionais que consideram as features originais e aquelas desenvolvidas no processo de feature engineering. Acreditamos que uma abordagem multifatorial seria mais resiliente do que depender de um único indicador.

- **Consideração da Alegação de originalidade:** A informação fornecida pelo vendedor na coluna originalidade foi um ponto de partida importante.
    - Produtos explicitamente declarados como **"Compatível"** foram classificados com o novo rótulo compativel. Esta decisão foi tomada para distinguir claramente esses produtos, que não são falsificações diretas, mas são relevantes no contexto de alternativas aos originais.
    - Alegações como "Falso", "Não Original", "Alternativo" levaram ao rótulo nao_original_declarado.
- **Ceticismo com Alegações de "Original":** Reconhecendo que a simples alegação de "Original" por parte do vendedor não garante a autenticidade, cruzamos esta informação com outros fatores:
    - **Confiança no Vendedor (seller_trust_level):** Uma alegação de "Original" vinda de um vendedor com seller_trust_level = 1 (Suspeito) foi tratada com alto grau de ceticismo. Se acompanhada de um preço baixo (desvio_preco_media_categoria significativamente negativo), a tendência foi classificar como pirata_provavel_.... Mesmo com preço normal, a desconfiança no vendedor levou a rótulos como avaliar_manual_alegado_original_vendedor_suspeito.

    - **Desvio de Preço (desvio_preco_media_categoria):** Produtos alegadamente "Originais" mas com preços muito abaixo da média da sua categoria foram sinalizados. Se o vendedor também fosse suspeito, a classificação pendeu para pirata_provavel_.... Se o vendedor fosse confiável (trust == 3), mas o preço ainda assim baixo, classificamos como avaliar_manual_original_preco_baixo_vendedor_confiavel (podendo indicar uma promoção legítima, um erro, ou até mesmo que o vendedor confiável foi enganado).
    - **Nota de Qualidade (nota_qualidade):** Quando disponível, uma nota de qualidade muito baixa associada a um vendedor suspeito também serviu como um indicador negativo, influenciando a classificação para pirata_provavel_....
- **Estrutura das Regras:** As regras foram ordenadas para dar prioridade a indicadores mais fortes de não originalidade ou fraude. Por exemplo, um preço muito baixo de um vendedor de baixa confiança poderia levar a um rótulo de pirata_provavel mesmo que a originalidade alegada fosse "Original".
- **Rótulos de Avaliação Manual:** Para casos ambíguos onde as heurísticas não permitiam uma classificação definitiva com alta confiança, foram utilizados diversos rótulos como avaliar_manual_.... Estes indicam anúncios que necessitariam de uma análise humana ou seriam candidatos ideais para que o futuro modelo de Machine Learning aprendesse nuances mais sutis.

Este conjunto inicial de heurísticas é um ponto de partida. Entendemos que ele será iterativamente refinado à medida que mais dados forem analisados e, idealmente, com o feedback e informações adicionais da HP.

5. Desafios e Decisões Chave no Processamento

Durante o desenvolvimento, alguns desafios se destacaram, moldando nossas decisões:

- **Limpeza da Coluna preco:** Foi necessário um tratamento cuidadoso para converter diversos formatos de texto de preço em valores numéricos, incluindo a remoção de "R$", espaços e a correta interpretação de separadores decimais e de milhar. Um limite superior (MAX_PLAUSIBLE_PRICE_INDIVIDUAL) foi implementado para converter preços individuais absurdamente altos para NaN.

- **Cálculo e Interpretação de preco_medio_categoria:** Este foi um ponto particularmente desafiador. Inicialmente, observamos médias de categoria com valores astronômicos no arquivo CSV final. Após intensa depuração, que envolveu a análise dos print no console versus a exibição no Excel, identificamos que:
    1. Os cálculos no Python estavam, de fato, corretos para as categorias visíveis na depuração.
    2. A discrepância visual no Excel era causada pela interpretação do separador decimal: o Pandas salvava com . (padrão), mas o Excel, configurado para localidade brasileira, esperava ,. Isso foi **resolvido especificando decimal=',' no método to_csv**.
    3. Para garantir a robustez contra quaisquer outros valores extremos que pudessem surgir em categorias não inspecionadas ou devido a dados problemáticos, implementamos um "cap" também para a preco_medio_categoria calculada (MAX_PLAUSIBLE_PRICE_AVERAGE), convertendo para NaN qualquer média que excedesse esse valor ou fosse não positiva.
    4. A normalização da coluna de categoria (tipo_cartucho), removendo espaços e tratando valores nulos/inválidos, também foi crucial para a consistência do agrupamento.
- **Confiança do Vendedor (seller_trust_level):** Conforme mencionado, esta feature reflete o conhecimento atual do grupo. Foi uma decisão consciente utilizar esta pesquisa como uma heurística inicial forte, com o pleno entendimento de que ela é subjetiva e se beneficiará enormemente de validação e dados adicionais da HP no futuro.

6. Conclusão e Próximos Passos

O processo descrito resultou na criação de um dataset estruturado e rotulado heuristicamente, pronto para as próximas etapas do projeto, que incluem o treinamento de modelos de Machine Learning. As features originais, juntamente com aquelas desenvolvidas através do feature engineering, e a lógica de rotulagem que tenta discernir enganos, fornecem uma base sólida.

Os próximos passos recomendados incluem:

- Utilizar este dataset para treinar e avaliar diferentes algoritmos de Machine Learning.
- Refinar continuamente as heurísticas de rotulagem à medida que se ganha mais conhecimento sobre os dados e os padrões de pirataria.

- Incorporar futuras evidências e dados fornecidos pela HP para aprimorar, principalmente, a classificação de confiança dos vendedores e a validação dos rótulos.
- Expandir o conjunto de features, possivelmente explorando mais a fundo o conteúdo textual dos anúncios.

Acreditamos que o trabalho realizado representa um avanço significativo na direção de uma solução automatizada para auxiliar a HP na identificação de produtos suspeitos no e-commerce.