Projeto de Data Science: Detecção de Fraudes em Entregas do Walmart

Objetivo do Projeto

O Walmart é a maior rede de varejo dos Estados Unidos gerando em média de US$ 1,6 bilhão em receita por dia. Em termos de vendas na loja, o Walmart faz uma média de: US$ 17.000 por segundo, US$ 1,1 milhão por minuto e US$ 68 milhões por hora.

A plataforma online e as vendas no varejo do Walmart contribuem com uma grande parte de seus lucros. O Walmart é o maior varejista de alimentos dos Estados Unidos, com mais de US$ 264 bilhões em vendas de alimentos nos EUA no ano passado.

De acordo com uma pesquisa recente da Lending Tree, o roubo em caixas de autoatendimento é um problema real, e os indivíduos que se envolvem nisso provavelmente repetirão o comportamento. O Walmart enfrentou perdas com roubos no varejo, com estimativas de 3 bilhões em 2021, 6,1 bilhões em 2022 e US$ 6,5 bilhões em 2023, o que mostra um crescimento de $400 milhões de dólares em perdas por furto no último ano.

Você é o cientista de dados do Walmart trabalhando exclusivamente para a área de e-commerce. Foi indentificado que o maior aumento proporcional em furtos aconteceu nas compras efetuadas via e-commerce em que usuários relatam não receberem todos os pedidos feitos nas suas compras. Do crescimento de roubos de 2022 para 2023, 53% do aumento veio das compras online.

Você foi designado a cuidar deste projeto para diminuir o número de fraudes e roubos através dos sites de e-commerce do Walmart. Foi identificado que estes roubos acontecem na entrega.

Neste projeto, o objetivo principal é identificar possíveis fraudes em entregas realizadas pelo Walmart na região Central da Flórida para servir de modelo (caso o resultado do projeto seja positivo) para as demais regiões dos EUA.

O foco é analisar os dados de entrega para detectar padrões e anomalias que possam indicar que itens declarados como entregues pelo motorista não foram efetivamente recebidos pelo cliente. O Walmart tem enfrentado reclamações de consumidores sobre entregas incompletas e, com a análise dos dados, deseja entender se a responsabilidade pela fraude pode ser atribuída aos consumidores, aos entregadores ou a outra causa.

Descrição do Problema

Nos EUA o Walmart possui um sistema similar ao Uber em que entregadores se cadastram para entregar pedidos feitos através do site do Walmart. Estes entregadores (motoristas) não são funcionários do Walmart, mas trabalham independentemente aceitando pedidos de entrega e fazendo o recebimento destes pedidos pela equipe do Walmart e entrega destes pedidos ao endereço do consumidor. Muitos consumidores relataram que certos itens de seus pedidos não foram entregues, apesar de o sistema marcar a entrega como concluída. Isso levanta algumas questões críticas:

1. Fraude do Entregador (Motorista): Há evidências de que motoristas possam estar reportando a entrega de itens que, na realidade, não chegaram até o cliente. Eles podem estar omitindo ou desviando itens do pedido, registrando, no entanto, a entrega total.
2. Erro do Sistema ou Processo: Pode ser que o problema esteja em falhas no sistema de registro ou no processo de entrega, não se limitando a fraudes intencionais.
3. Fraude do Consumidor: Em alguns casos, o consumidor pode estar declarando como não ter recebido um produto que foi entregue para assim pedir o reembolso do produto.

Fontes de Dados

Os dados fornecidos pelo Walmart incluem informações sobre as ordens de entrega, com detalhes relevantes para a análise, tais como:

Tabelas Disponibilizadas

1. Orders (pedidos)

Esta tabela possui uma amostra de pedidos realizados no Walmart e-commerce site no ano de 2023 na região de Central Florida, na Florida, EUA.

Amostra dos dados

date	order_id	order_amount	region	items_delivered	items_missing	delivery_hour	driver_id	customer_id
2023-11-01	c7a343f7	$634.57	Clermont	14	3	13:50:54	WDID09873	WCID5170
2023-07-16	2069829	$418.46	Apopka	16	3	5:12:52	WDID09874	WCID5901
2023-06-15	d7f690a0	$314.43	Sanford	12	3	10:49:04	WDID09875	WCID5652

- Date - Data do Pedido
- Order_id - Número identificador do pedido. Este valor é único por pedido
- Order_amount - Valor do Pedido
- Region - Região de entrega do pedido
- Items_delivered - Número total Itens entregues
- Items_missing - Número total Itens perdidos (Não entregues)
- Delivery_hour - Hora da entrega
- Driver_id - ID do entregador (motorista responsável por entregar o pedido)
- Customer_id - ID do consumidor (cliente que fez o pedido)

2. Missing Items Data (Dados de Produtos declarados como não recebidos pelo cliente)

Esta tabela possui uma amostra de produtos adquiridos na compra no Walmart e-commerce e que foram declarados como não recebidos pelo cliente.

Amostra dos dados

order_id	product_id_1	product_id_2	product_id_3
c7a343f7-3f1d-497c-8004	PWPX0982761090982	PWPX0982761090982	PWPX0982761090982
20698293-8399-4fda-af1	PWPX0982761090983	PWPX0982761090983	PWPX0982761090983
d7f690a0-c1c2-4b36-b05f	PWPX0982761090984	PWPX0982761090984	PWPX0982761090984
15cba1bc-6a92-4c97-b37	PWPX0982761091109	PWPX0982761091088	
304f3d20-4780-475a-aca	PWPX0982761091110	PWPX0982761091089	
d8b4a4b3-b35e-427c-a0c	PWPX0982761091111	PWPX0982761091090	
b0a31709-0fc4-46cf-b488	PWPX0982761091112		
16195ca2-121d-42e0-ba9	PWPX0982761091113		
c1ed403b-da93-47e8-894	PWPX0982761091114		
3077925f-6a4d-4026-936	PWPX0982761091115		
f2c09b49-874e-4f3b-a46b	PWPX0982761091116		

Algumas compras tiveram mais de um produto declarado pelo consumidor como não entregue. Esta tabela mostra o codigo do pedido e o código do produto que não foi recebido pelo cliente.Algumas compras tiveram apenas um produto não entregue enquanto outras tiveram 2 e outras 3.

- **Order_id** - Número identificador do pedido. Este valor é único por pedido
- **Product_id_1** - Primeiro ou único produto não entregue
- **Product_id_2** - Segundo produto não entregue
- **Product_id_3** - Terceiro produto não entregue

3. Driver’s Data (Dados dos Entregadores (motoristas))

Esta tabela possui os dados dos motoristas que coletaram as compras no Walmart e realizaram a entrega na casa do consumidor.

- **Driver_id** - Número único (ID) de identificação do motorista (entregador)
- **Driver_name** - Nome do motorista (entregador)
- **Age** - Idade do motorista (entregador)
- **Trips** - Quantas entregas este motorista (entregador) realizou no ano de 2023

driver_id	driver_name	age	Trips
WDID09873	Pamela Moore	18	64
WDID09874	Billy Lawson	18	37
WDID09875	Stephen Randolph	18	64
WDID09876	Jordan Daniel	18	53

4. Products Data (Dados dos Produtos)

Esta tabela possui os dados dos produtos que foram comprados. Esta tabela pode ajudar a identificar se existe um padrão no tipo de produto que mais é reportado como não entregue.

- **Product_id** - Código de identificação do Produto
- **Product_name** - Nome do Produto
- **Category** - Categoria do Produto
- **Price** - Preço do Produto

5. Customer’s Data (Dados dos clientes)

Esta tabela possui os dados dos clientes que realizaram as compras

- **Customer_id** = ID do consumidor
- **Customer_name** - Nome do Consumidor
- **Customer_age** - Idade do Consumidor

customer_id	customer_name	customer_age
WCID5170	Elijah Taylor	30
WCID5901	Alexis Ross	58
WCID5652	Carla Knox	23

Tarefas do Projeto

Os estudantes devem realizar as seguintes tarefas para entender a origem das falhas de entrega e propor soluções eficazes:

1. Análise Exploratória dos Dados (EDA):

- Entender as características principais do conjunto de dados.
- Identificar e lidar com dados ausentes, se houver.
- Analisar a distribuição dos dados de entrega para detectar padrões de comportamento incomum, como uma alta frequência de itens faltando em pedidos específicos ou áreas geográficas.
- Combinar os Dados usando SQL para entender os padrões baseado na combinação de tabelas diferentes através dos IDs em comum entre as tabelas.

2. Detecção de Padrões de Fraude:

- Identificar padrões ou fatores que possam sugerir fraudes, como:
- Entregadores com taxas de reclamação acima da média que apresentam uma alta frequência de “entregas completas” em que clientes relatam itens faltando.
- Avaliar a variação entre os registros de itens entregues e os itens efetivamente recebidos pelos clientes, buscando correlações entre motoristas específicos e ordens com problemas.
- Usar técnicas de modelagem para detectar comportamentos anômalos (ex.: clusterização, análise de outliers).

3. Avaliação de Causas e Responsabilidades:

- Investigar se a fraude é atribuível a motoristas específicos, a entregadores, ou se existem problemas sistêmicos.
- Analisar se há problemas recorrentes em certas regiões geográficas, horários ou em tipos específicos de itens.
- Identificar se o problema aumentou ou diminuiu depois em determinados períodos
- Existem produtos específicos que são mais declarados como não recebidos do que outros? Existe um padrão nestes tipos de produtos?

5. Recomendações e Medidas Preventivas:

- Propor medidas para reduzir as ocorrências de fraudes (ex: um sistema de verificação de entregas com foto dos produtos entregues é viável, assinatura digital do cliente?)

- Sugerir melhorias no processo de entrega, como auditorias periódicas dos motoristas e entregadores com maior frequência de reclamações.
- Se no atual cenário de 2023 aumentou em 200 milhões (32%) a porcentagem de fraudes nos pedidos online, qual sua expectativa de redução caso suas medidas sejam estipuladas?

6. Propor Melhorias nos Dados

Existem dados que não estão disponíveis para a análise que ajudariam a melhorar a identificação do problema e as iniciativas de mitigação? Que dados ou informações seriam relevantes?

Entrega Final do Projeto

Ao concluir o projeto, os estudantes deverão apresentar:

1. Relatório Completo (Report): Um relatório que descreva as etapas da análise,insights, os resultados obtidos, as conclusões e a viabilidade de implementação.
2. Recomendações de Medidas Preventivas: Sugestões de ações que o Walmart pode tomar para minimizar fraudes futuras.
3. Dashboard: Um modelo de dashboard em Looker, Excel, Google Sheets ou PowerBI de monitoramento das entregas baseado nos dados disponíveis que ajudaria a identificar em tempo real problemas nas entregas.
4. Propostas de aprimoramento: Além da análise o que você propõe em termos de análise para melhor a acuracidade da detecção e prevenção deste tipo de fraude? Testes A/B? Quais? Melhoria nos dados? (Quais dados seriam necessários?), Pesquisas com o consumidor? Pesquisas qualitativas com os motoristas?

Pontos de Avaliação

O projeto será avaliado com base nos seguintes critérios:

- Compreensão e uso adequado das técnicas de EDA e análise estatística.
- Eficácia dos modelos de detecção de fraudes implementados.
- Clareza e detalhamento das recomendações finais.
- Qualidade e organização do relatório e dashboard entregue
- Coerência e profissionalismo na apresentação dos resultados.

Conclusão

Esse projeto permitirá que vocês apliquem habilidades analíticas e de ciência de dados a um problema real, onde a detecção de padrões e a modelagem preditiva podem ter impacto direto nos processos e na experiência do cliente. Além disso, a tarefa incentiva o desenvolvimento de uma mentalidade orientada à resolução de problemas e o pensamento analítico, ao abordar fraudes e medidas preventivas em um cenário prático.

Link para acesso aos Dados

Os dados estão disponíveis para download nesta aula abaixo, ou podem ser acessados via Data World pelo Link:

Data World

https://data.world/jerrys/introduction-to-data-analytics/workspace

Tabelas

- orders
- products_data
- missing_items_data
- customers_data
- drivers_data