# An E-Commerce Dataset Revealing Variations during Sales

Hélio Santos
Henrique Medeiros
João Dias

# Introdução

- Métodos existentes de Learning-To-Rank (LTR)
  - Dados independentes e identicamente distribuídos
  - Não considera evolução da distribuição
  - Não considera o ajuste fino
- Previsões imprecisas
- Problemas ao reajustar
- Solução: novo conjunto, práticas comuns
  - Dados com picos (spikes)

# Perguntas de Pesquisa

1. O paradigma padrão de LTR consegue superar os desafios de um dataset tempo-variante?
2. Abordagens mais avançadas consegue mitigar os desafios de um dataset tempo-variante?
3. Existem direções potencialmente efetivas para atacar o desafio de dados tempo-variantes com picos?

# Trabalhos Relacionados

- Diversos datasets para avaliar modelos de ranking
  - Pequenos conjuntos de dados
  - Características densas
  - Dados artificiais
- Cada dataset com uma especificidade
- Problema comum a todos os datasets: o foco do trabalho
- Por que o dataset proposto é bom?
  - Exploração de diferentes paradigmas de aprendizagem
    - Exemplos: continual learning, few-shot, sequence representation, etc.

# Trabalhos Relacionados

| | #Interactions | Time Range | E-Commerce | Implicty | Spikes |
|---|---|---|---|---|---|
| Amazon | 230 millions | 30 years | Yes | No | None |
| CIKMCUP | 1 million | 5 months | Yes | Yes | None |
| LETOR | < 1 million | Unknown | No | No | None |
| MovieLens 100K | 0.1 millions | 7 months | No | No | None |
| MovieLens 25M | 25 millions | 24 years | No | No | None |
| Taobao UVA | 100 millions | 9 days | Yes | Yes | None |
| Yelp | 7 millions | 8 years | Yes | Both | None |
| Ours | 16 millions | 80 days | Yes | Yes | Labeled |

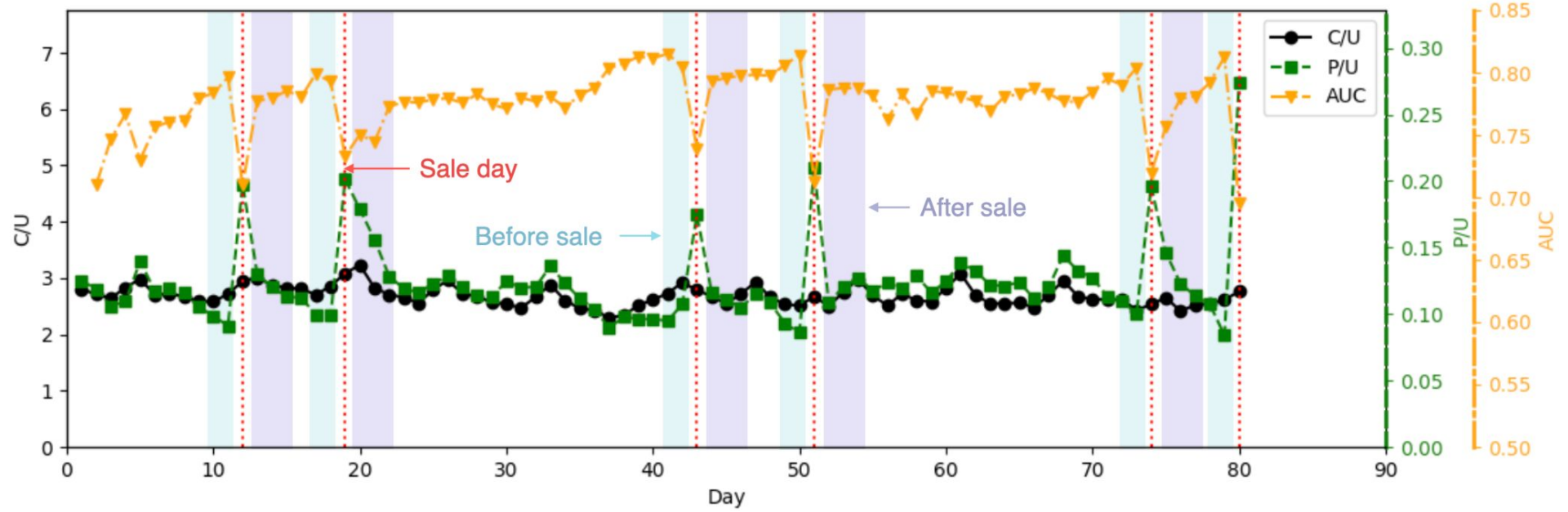
# O dataset

- Métricas
  - 80 dias de coleta
  - 19.713 queries
  - 109.940 usuários
  - 217.468 itens
- Interações Usuário-Item: exposição, click, compra
- Reduziram as exposições para próximo da quantidade de cliques: exposições pouco informativas, otimização de espaço e manutenção da utilidade.

# O dataset: estatísticas

- Foca nas análises de
  - Clicks per User (C/U)
  - Purchases per User (P/U)
  - **Métrica de avaliação:** AUC
- Separam os períodos em 5:
  - **Total:** estatísticas gerais
  - **Regular:** períodos não afetados pela liquidação; P/U reduzido; C/U estável
  - **Pré-liquidação:** os 2 dias anteriores à liquidação; spike em P/U
  - **Liquidação:** data da liquidação; spike em P/U
  - **Pós-liquidação:** os 3 dias posteriores à liquidação; C/U e P/U reduzem

# Distribuição das estatísticas no dataset

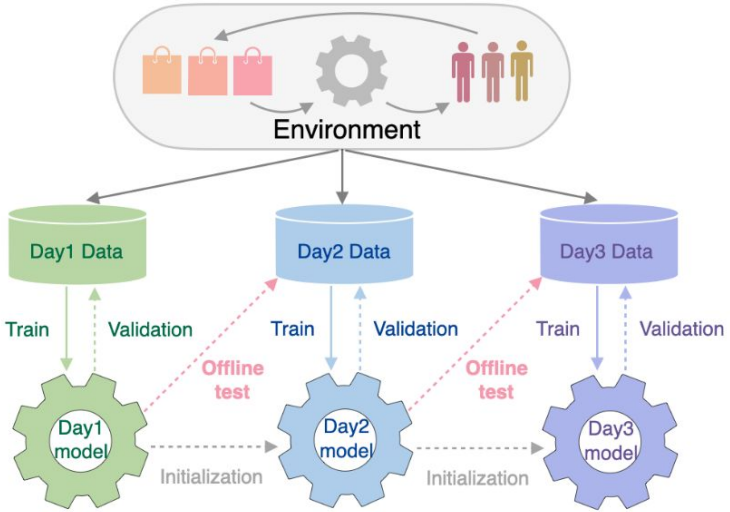


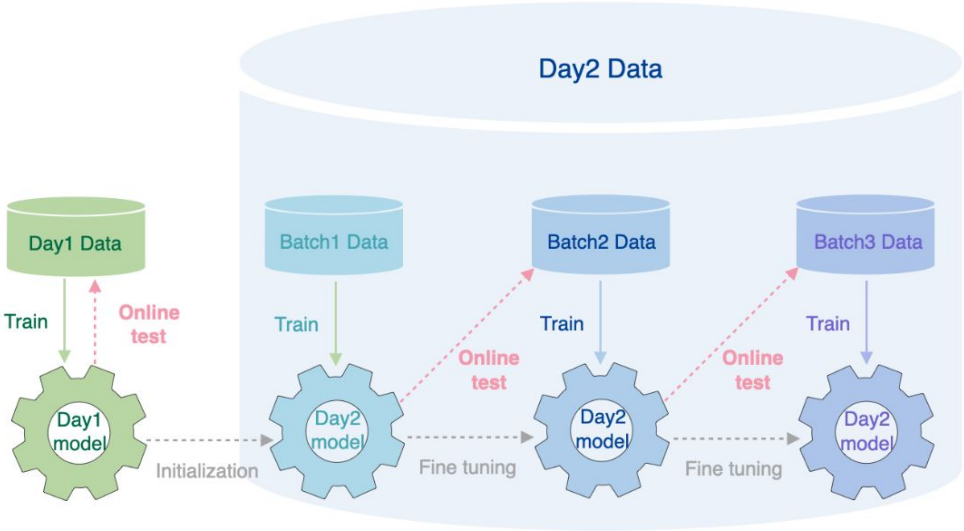
# Treinamento do modelo

- Para o dia *d*, o modelo é treinado com os dias [*1, d-1*]
- Os dados do dia *d* são particionados em lotes *D*
  - O ajuste fino é feito nos lotes [*D^1*, *D^k*]
- Após calculados os k lotes, é calculada a perda $\sum_{k=1}^{K} L(D^k, f_{\theta_k})$
  - f: o modelo de ranqueamento
  - L: alguma função de perda
  - θ_k: parâmetros do modelo após processados os k-1 lotes
- O modelo pode ser atualizado após novos dados serem revelados
  - Usa-se a retropropagação de uma passada
  - α: taxa de aprendizagem

$$\theta_k \leftarrow \theta_{k-1} + \alpha \partial L(D^{k-1}, f_{\theta_{k-1}}) / \partial \theta_{k-1}$$

# Treinamento do modelo: *offline vs online*



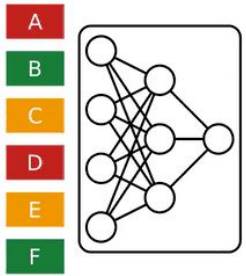



# Experimentos

- Focado em analisar as características distintas do dataset
- Técnicas que demonstraram performance distinta, foram testados na prática
- Modelos Avaliados
  - **MLP** (Multilayer Perceptron)
  - **WDL** (Wide & Deep Learning)
  - **MMoE** (Multi-gate Mixture-of-Experts)

# Contribuições

- Learning-to-Rank (LTR)
  - Dataset temporal com picos
  - Demonstração de Benchmarks
- Objetivos
  - Auxiliar no aprimoramento dos modelos de LTR
    - Especificamente em períodos de liquidação



# Conclusão

- Apresentação de um novo problema na área LTR + Benchmark dataset
- Focado em análise de dados reais, em específico: liquidações em e-commerce
- Pesquisas futuras
  - Metodologias que se adaptam ao longo do tempo
  - Técnicas refinadas para curadoria e remoção de dados

# Referências


- An E-Commerce Dataset Revealing Variations during Sales, ZHANG, J. et al.

# An E-Commerce Dataset Revealing Variations during Sales

Hélio Santos
Henrique Medeiros
João Dias