Um Benchmark de Algoritmos de Recomendação Baseados em Sessão no Contexto do Mercado Imobiliário Brasileiro

Hygo Freire Tagarro[1], **Vinícius F. S. Mota**[1], **Giovanni Comarela**[1]

[1] Programa de Pós-Graduação em Informática (PPGI)
Universidade Federal do Espírito Santo (UFES)
Vitória – ES – Brazil

hygo.tagarro@edu.ufes.br, {vinicius.mota,gc}@inf.ufes.br

Abstract. *The real estate market accounts for a significant share of the Brazilian GDP, but recommending properties at scale imposes its own challenges: catalogs with high turnover, predominantly anonymous sessions, and short interactions. This work proposes a benchmark for session-based recommendation, using a proprietary Zap Imóveis dataset with 17.6 million interactions. Fifteen models are compared under full ranking and under negative sampling. Under full catalog, co-occurrence and factorization methods outperform neural models; under negative sampling, the order partially reverses. The divergence confirms, in a previously unstudied domain, that the evaluation protocol can alter the conclusion regarding which architecture is most suitable.*

Resumo. *O mercado imobiliário representa parcela significativa do PIB brasileiro, mas recomendar imóveis em larga escala impõe desafios próprios: catálogos com alta rotatividade, sessões predominantemente anônimas e interações curtas. Este trabalho propõe um benchmark para recomendação baseada em sessão, utilizando um dataset proprietário do Zap Imóveis com 17,6 milhões de interações. Quinze modelos são comparados sob ranking completo e sob amostragem negativa. Sob catálogo completo, métodos de coocorrência e fatoração superam modelos neurais; sob amostragem negativa, a ordem se inverte parcialmente. A divergência confirma, em um domínio até então não investigado, que o protocolo de avaliação pode alterar a conclusão sobre qual arquitetura é mais adequada.*

1. Introdução

O setor imobiliário tem papel central na economia brasileira e costuma refletir de forma bastante visível ciclos de expansão e retração econômica. Somente a construção civil terminou 2024 com PIB de R$ 359 bilhões, sendo o terceiro setor com maior crescimento no país [Instituto Brasileiro de Geografia e Estatística (IBGE) 2025]. Para uma parte considerável da população, o imóvel é o principal ativo de patrimônio: dados do Imposto de Renda indicam que 36,7% da riqueza declarada no Brasil correspondem a imóveis [Brasil. Ministério da Fazenda 2025]. Diferentemente do varejo convencional, comprar ou alugar um imóvel envolve compromisso financeiro de longo prazo, com custo elevado para reverter uma decisão ruim. Esse fato faz com que boas recomendações de imóveis sejam especialmente relevantes.

À medida que a oferta se expandiu, com o Anuário DataZAP 2025 registrando crescimento de 8,5% no número de lançamentos em 2024 (167 mil novas unidades)

e aumento de 7,4% no volume de transações de venda [DataZAP 2025], e os imóveis passaram a se dispersar por diferentes regiões metropolitanas, a busca baseada em visitas e anúncios locais tornou-se insuficiente. Os portais imobiliários preencheram essa lacuna: concentram o processo de descoberta e comparação, reduzem parte das incertezas da negociação e conectam compradores, inquilinos e corretores em escala que o modelo tradicional não alcançava.

Nesse contexto, a recomendação baseada em sessão se mostra adequada para o domínio imobiliário. A maioria dos usuários dos portais navega de forma anônima, sem autenticação, o que dificulta a construção de perfis persistentes de longo prazo. Na prática, o que conseguimos observar é apenas a sessão de navegação do usuário, ou seja, a sequência de anúncios acessados durante uma única visita ao portal. Modelos baseados em sessão lidam exatamente com esse tipo de problema: em vez de depender de um histórico acumulado entre diferentes visitas, eles extraem padrões a partir das transições entre itens dentro da própria sessão corrente [Hidasi et al. 2016].

Apesar disso, poucos estudos investigam de forma sistemática quais algoritmos de recomendação funcionam melhor nesse cenário. Embora modelos neurais tenham ganhado espaço, trabalhos de *benchmarking* recentes mostram que métodos heurísticos baseados em vizinhança podem superar arquiteturas profundas em tarefas baseadas em sessão, combinando bom desempenho com custo computacional menor [Ludewig et al. 2021]. Resultados semelhantes aparecem até em domínios com estrutura textual complexa, como o jurídico [Domingues et al. 2023], sugerindo que padrões locais de coocorrência e transição entre itens podem ser bastante robustos. Essa tendência, em que métodos conceitualmente mais simples superam abordagens mais complexas, também tem sido observada em outros paradigmas de recomendação: Lima et al. (2024) mostram que LLMs abertos com cerca de 8 bilhões de parâmetros podem superar modelos proprietários com mais de 175 bilhões de parâmetros na tarefa de recomendação de próximo item. No entanto, essa hipótese é pouco testada no contexto imobiliário brasileiro, em que a tomada de decisão é guiada por restrições rígidas (*hard filters*) como preço e número de dormitórios, fatores decisivos para 98% e 92% dos usuários, respectivamente [DataZAP 2025].

Diante disso, este trabalho propõe um *benchmark* de recomendação baseada em sessão em um *dataset* real do mercado imobiliário brasileiro, focando exclusivamente em padrões de interação sequencial, sem utilizar atributos explícitos dos imóveis. Essa escolha é deliberada: em portais imobiliários, as restrições rígidas tipicamente atuam na etapa de filtragem prévia (o usuário define faixa de preço, localização e tipologia), de modo que o sistema de recomendação opera sobre o conjunto já filtrado, onde o sinal sequencial tende a se tornar um dos principais diferenciadores. As contribuições são: (i) um protocolo experimental que combina deduplicação por *fingerprinting* (agrupamento de anúncios duplicados por atributos estruturais), particionamento temporal com retreinamento mensal e avaliação sob dois regimes distintos: ranking completo sobre todos os itens do catálogo e amostragem negativa com 100 candidatos; (ii) uma comparação empírica de 15 modelos cobrindo seis paradigmas arquiteturais, evidenciando que a escolha do regime de avaliação pode inverter a hierarquia entre modelos.

2. Trabalhos Relacionados

Este trabalho dialoga com três frentes de pesquisa: recomendação no mercado imobiliário, *benchmarks* de recomendação baseada em sessão e o efeito do protocolo de avaliação sobre a comparação entre modelos. As subseções a seguir discutem cada uma delas.

2.1. Recomendação no Domínio Imobiliário

A literatura sobre sistemas de recomendação para imóveis permanece escassa. Gharahighehi et al. (2021) identificaram que a maioria dos trabalhos compara modelos apenas contra *baselines* simples, usa métricas exclusivamente de acurácia e opera sobre *datasets* privados de pequena escala, limitações que dificultam a reprodutibilidade e a generalização dos resultados. Entre os desafios específicos do domínio, destacam-se: (i) alta rotatividade do catálogo, com anúncios entrando e saindo continuamente; (ii) esparsidade extrema, dado que cada usuário interage com uma fração mínima dos imóveis disponíveis; (iii) predominância de sessões anônimas, inviabilizando a construção de perfis de longo prazo; e (iv) papel central de restrições rígidas (preço, localização, tipologia) na navegação, que filtram o espaço de busca antes que qualquer modelo de recomendação atue [DataZAP 2025]. Polohakul et al. (2021) propõem *pipelines* híbridos para lidar com o *cold-start*, mas sem comparação sistemática entre múltiplas arquiteturas.

2.2. Benchmarks para Recomendação Baseada em Sessão

No contexto de sistemas de recomendação, um *benchmark* consiste em uma avaliação sistemática e comparável de múltiplos algoritmos sobre um conjunto de dados comum, com protocolo experimental controlado, permitindo conclusões sobre superioridade arquitetural que vão além de comparações pontuais [Ludewig and Jannach 2018].

A proposta de Hidasi et al. (2016) de aplicar redes recorrentes à recomendação baseada em sessão gerou dezenas de arquiteturas neurais, incluindo modelos baseados em atenção [Li et al. 2017, Liu et al. 2018, Kang and McAuley 2018, Sun et al. 2019], convoluções [Tang and Wang 2018, Yuan et al. 2019] e grafos.

Nesse contexto, os trabalhos de Ludewig e Jannach (2018, 2021) representam o esforço mais abrangente de *benchmarking* na área. Em 2018, os autores mostraram que heurísticas de vizinhança frequentemente igualam ou superam o GRU4Rec em múltiplos *datasets* de e-commerce e música. A versão ampliada de 2021 estendeu a comparação para doze algoritmos, incluindo NARM, STAMP e NextItNet, e confirmou que o progresso em acurácia dos métodos neurais permanece limitado em relação a *baselines* simples, especialmente quando o protocolo experimental é controlado de forma rigorosa. Um aspecto central desses *benchmarks* é a demonstração de que diferenças aparentes de desempenho podem decorrer de escolhas metodológicas (como a estratégia de divisão dos dados) e não de superioridade arquitetural. Domingues et al. (2023) estenderam essa análise ao domínio jurídico brasileiro, confirmando que características específicas do domínio influenciam significativamente o ranking entre algoritmos. O presente trabalho segue essa linha, adotando protocolo padronizado via RecBole [Xu et al. 2023] e estendendo a comparação ao domínio imobiliário.

2.3. Impacto do Protocolo de Avaliação

A avaliação de modelos de recomendação baseada em sessão pode seguir dois regimes distintos. No **ranking completo**, o modelo atribui uma pontuação a todos os itens do

catálogo e o item relevante é posicionado nesse ranking global, refletindo o cenário real de produção. Na **amostragem negativa**, o item relevante é posicionado contra um subconjunto pequeno de itens sorteados aleatoriamente (tipicamente 100 candidatos), o que reduz o custo computacional da avaliação mas simplifica artificialmente o problema.

Krichene e Rendle (2020) demonstraram formalmente que métricas calculadas sob amostragem negativa são inconsistentes com suas versões exatas: a conclusão de que A supera B sob amostragem não implica superioridade sob ranking completo. À medida que o número de amostras negativas diminui, todas as métricas convergem para a AUC, perdendo capacidade de discriminar o desempenho nas primeiras posições. Na prática, isso afeta diretamente a literatura de recomendação baseada em sessão, na qual a avaliação com amostragem negativa é amplamente adotada. O presente trabalho explora essa questão empiricamente, avaliando os mesmos *checkpoints* sob ambos os regimes.

3. Recomendação Baseada em Sessão

Em portais imobiliários, a maioria dos visitantes navega sem se autenticar, o que impede a construção de perfis de longo prazo. Na prática, a única informação disponível sobre o comportamento de um visitante é a sequência de anúncios que ele acessou durante uma visita ao portal, ou seja, a sua sessão de navegação. Esse cenário torna a recomendação baseada em sessão a abordagem natural para o domínio, pois ela opera exclusivamente sobre o comportamento de curto prazo, sem pressupor identidade persistente do usuário.

Neste trabalho, adota-se a formalização de recomendação baseada em sessão estabelecida na literatura [Hidasi et al. 2016, Ludewig et al. 2021]. O objetivo consiste em modelar o comportamento de curto prazo dos usuários a partir de sequências de interações observadas durante uma sessão. A seguir, são definidos os conceitos fundamentais, contextualizados no cenário de recomendação de anúncios de imóveis em plataformas digitais.

Denota-se por $\mathcal{I}_t$ o conjunto de todos os *itens* (anúncios) presentes no sistema e disponíveis para recomendação no instante de tempo t. Observe que, diferentemente de casos estáticos, o catálogo é dinâmico, ou seja, novos anúncios são continuamente adicionados, enquanto outros são removidos ou tornam-se inativos ao longo do tempo.

Uma sessão S é definida como uma sequência ordenada de interações realizadas por um usuário com um ou mais itens durante uma visita contínua à aplicação. Formalmente, $S = [v_1, v_2, \ldots, v_k]$, onde $v_i \in \mathcal{I}_t$ representa o i-ésimo item com o qual o usuário interagiu durante sua visita. Enfatiza-se que cada sessão está associada a um único usuário no contexto de uma visita, independentemente de seu estado de autenticação. As interações são tratadas como sinais de *feedback* implícito, sem distinção entre os diferentes tipos de ação registrados, refletindo apenas a ocorrência de interesse observável em um item.

Dessa forma, um *recomendador baseado em sessão* é uma função que recebe como entrada o histórico parcial de uma sessão, relativo à navegação de um usuário, e retorna os itens com os quais tal usuário terá maior chance de interagir. Para isso, o recomendador atribui a cada $v \in \mathcal{I}_t$ um *score*, que reflete o grau de interesse do usuário em interagir com v. Por fim, um sistema de recomendação pode apresentar ao usuário apenas os N itens com maiores *scores*. Nesse paradigma de recomendação, apenas o histórico parcial da sessão é considerado.

4. Dataset e Análise Exploratória

Para atingir os objetivos deste trabalho, foi utilizado um *dataset* proprietário extraído dos logs de interação do portal Zap Imóveis[1], cobrindo o período de 01/03/2024 a 01/07/2024, previamente anonimizado em conformidade com a LGPD. O *dataset* consolidado contém aproximadamente 17,6 milhões de interações distribuídas em 3,4 milhões de sessões. Em termos de volume, os dados brutos de eventos do portal totalizam aproximadamente 146 GB, cobrindo mais de 824 milhões de registros de interação no período analisado. Desse total, cerca de 461 milhões correspondem a anúncios de venda, dos quais 40 milhões são eventos de interação explícita (visualizações, favoritos, contatos e compartilhamentos).

A Figura 1 apresenta a distribuição acumulada (CDF) do número de eventos por sessão. A mediana é de apenas 3 eventos, enquanto a média atinge 5,14 (desvio padrão 4,9), refletindo uma distribuição fortemente assimétrica à direita: a maioria das sessões é curta, mas uma cauda longa de sessões de comparação puxa a média para cima. O percentil 90 corresponde a 10 eventos, e o máximo foi limitado a 50, seguindo protocolo de limpeza para reduzir ruído de *crawlers* [Hidasi et al. 2016]. Mesmo com sessões curtas, o engajamento médio é superior ao de *benchmarks* clássicos de e-commerce como RetailRocket (média 3,54) e CLEF News (média 3,38) [Ludewig and Jannach 2018].

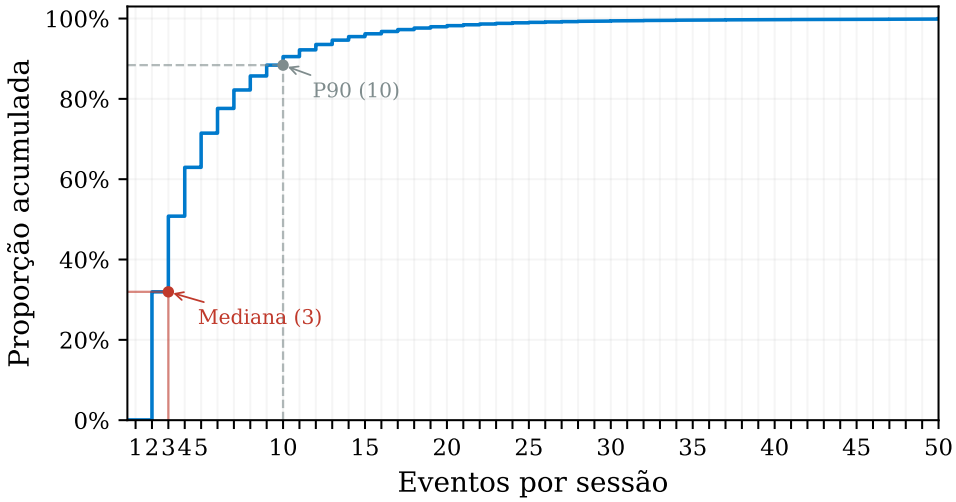

Figura 1. Distribuição acumulada (CDF) do número de eventos por sessão. A mediana de 3 eventos indica que metade das sessões contém no máximo três interações.

A análise por estado de autenticação reforça esse padrão. Usuários anônimos ($N \approx 2{,}9$M de sessões, 84% do total) apresentam sessões mais curtas (média 4,9, desvio padrão 4,37), compatíveis com navegação exploratória. Já os usuários identificados ($N \approx 464$k) exibem maior engajamento (média 6,7, desvio padrão 7,25), o que sugere uma fase de comparação mais direcionada.

No conjunto de dados utilizado, o mesmo imóvel pode ser anunciado por múltiplos corretores com IDs distintos e republicado com novo ID. Sem tratamento, um recomenda-

[1]Por questões de confidencialidade e acordo de não divulgação (NDA), os dados não podem ser disponibilizados. O código do protocolo e da avaliação está disponível em `https://github.com/hygo2025/Session-Based-Recommendation-Benchmark`, juntamente com instruções para aplicação em *datasets* públicos com estrutura similar.

dor trataria o mesmo imóvel como dezenas de itens diferentes, inflando o catálogo e aumentando a esparsidade. Para reduzir esse efeito, adotamos deduplicação por *fingerprinting*: cada imóvel recebe um identificador canônico calculado via MD5 sobre quatro atributos estruturais: geohash de precisão 7 (≈150 m), área útil discretizada em intervalos de $5\,m^2$, número de dormitórios e categoria do imóvel. Valores ausentes recebem sentinela padrão.

Essa forma de agrupar introduz perda de granularidade. Imóveis fisicamente distintos, mas com atributos idênticos (e.g., dois apartamentos de mesma planta no mesmo prédio) são mapeados para o mesmo identificador canônico. Esse é um compromisso deliberado: a redução de 63% no catálogo (de 1,1 milhão de anúncios brutos para 406 mil itens canônicos) atenua os problemas de *cold-start* e esparsidade dos dados.

Após o agrupamento, cada evento bruto é descrito por uma tupla $e = (u, v, t, tipo)$, onde u identifica o usuário, v o item canônico, t o *timestamp* e $tipo$ a ação realizada. O restante do pré-processamento segue duas etapas.

Resolução de identidade. Como um mesmo usuário pode alternar entre sessões anônimas e autenticadas, adotamos uma heurística de unificação: para cada identificador anônimo, selecionamos o identificador autenticado mais frequentemente associado a ele. Para evitar fusões indevidas em dispositivos compartilhados, identificadores anônimos associados a mais de δ usuários distintos não são unificados. Neste trabalho, adotamos $\delta = 7$; experimentos com $\delta \in [5, 10]$ produziram resultados equivalentes.

Construção de sessões. As interações são organizadas em sessões pelo identificador fornecido pela plataforma. Mantemos apenas eventos de interação explícita com anúncios de venda (visualização, favoritar, contato), evitando misturar dinâmicas comportamentais distintas (como aluguel versus compra). Dentro de cada sessão, as interações são ordenadas por *timestamp* e repetições consecutivas do mesmo item são removidas. Essa deduplicação é necessária porque diferentes ações sobre o mesmo anúncio (visualizar, favoritar, contatar) geram eventos distintos que, sem tratamento, produziriam sequências artificiais do tipo $v \to v \to v$. Padrões de retorno ($v \to w \to v$), comuns em cenários de comparação, são preservados.

Por fim, sessões com menos de 3 eventos são descartadas, garantindo ao menos um evento para treino, um para validação e um para teste no esquema *Leave-One-Out* (Seção 5), e sessões com mais de 50 eventos são truncadas nos 50 primeiros. Itens com frequência inferior a 5 são removidos para reduzir o impacto de itens extremamente raros.

5. Protocolo Experimental

Todos os experimentos foram conduzidos com a biblioteca *RecBole* [Xu et al. 2023].

5.1. Modelos Avaliados

A Tabela 1 resume os 15 modelos avaliados, organizados por paradigma. A seleção cobre desde *baselines* não personalizados até arquiteturas neurais profundas e baseadas em grafos, permitindo avaliar se o mecanismo de codificação influencia o desempenho em um domínio com sessões curtas.

Merece destaque a adaptação do LightGCN a este cenário. Por ser originalmente projetado para recomendação com perfis de usuário, sua aplicação a sessões exige um ajuste: cada sessão é tratada como um nó de "usuário" no grafo bipartido sessão–item,

Tabela 1. Modelos avaliados, organizados por categoria.

Categoria	Modelo	Mecanismo	Ref.
Não personalizado	Random	Aleatório (*baseline* inferior)	–
	Pop	Frequência global	–
Vizinhança	ItemKNN	Similaridade item-item (coocorrência)	[Sarwar et al. 2001, Deshpande and Karypis 2004]
Fatoração de Matrizes	BPR-MF	Fatoração otimizada via perda *pairwise*	[Rendle et al. 2009]
Latente / Transição	FPMC	Fatoração + cadeias de Markov	[Rendle et al. 2010]
	FOSSIL	Similaridade + transições markovianas	[He and McAuley 2016]
	TransRec	Translação no espaço latente	[He et al. 2017]
Neural sequencial	GRU4Rec	Recorrência (GRU)	[Hidasi et al. 2016]
	NARM	Recorrência + atenção dual	[Li et al. 2017]
	STAMP	Atenção pura (último item)	[Liu et al. 2018]
	Caser	Convolução 2D (horizontal + vertical)	[Tang and Wang 2018]
	NextItNet	Convolução dilatada causal	[Yuan et al. 2019]
	SASRec	Autoatenção unidirecional	[Kang and McAuley 2018]
	BERT4Rec	Mascaramento bidirecional	[Sun et al. 2019]
Grafo	LightGCN	Propagação em grafo bipartido	[He et al. 2020]

e o grafo é reconstruído a cada partição mensal. O LightGCN é um modelo transdutivo, ou seja, aprende representações (*embeddings*) apenas para os nós presentes no grafo de treino. Sessões do conjunto de teste, portanto, não possuem *embedding* próprio e são representadas na inferência pela média dos *embeddings* dos itens já visitados na sessão.

5.2. Particionamento e Avaliação

O método de avaliação dos diferentes recomendadores segue a estratégia ***Leave-One-Out*** [Ludewig and Jannach 2018, Kang and McAuley 2018]. Dada uma sessão $S = [v_1, \ldots, v_n]$, os itens $[v_1, \ldots, v_{n-2}]$ compõem o conjunto de treinamento, o item v_{n-1} é reservado para validação e v_n para teste. Essa divisão respeita a ordem cronológica, garantindo que nenhuma informação futura seja utilizada durante o treinamento. Antes do particionamento, itens com frequência inferior a 5 são removidos e o vocabulário resultante é fixado, eliminando *cold-start* de itens no teste. Além dessa divisão intra-sessão, o protocolo é aplicado de forma independente a cada mês do *dataset*, com todos os modelos treinados do zero em cada recorte.

A granularidade mensal foi adotada por duas razões complementares. Portais imobiliários apresentam renovação contínua do catálogo, com anúncios entrando e saindo em escala diária, de modo que janelas de 30 dias capturam variações relevantes na composição dos itens disponíveis entre recortes consecutivos. Além disso, partições mensais oferecem volume suficiente de sessões para treinamento estável dos modelos e permitem quatro avaliações independentes no período disponível, viabilizando a análise de estabilidade temporal apresentada na Seção 6.

Este trabalho adota dois regimes de avaliação, cuja diferença é central para a interpretação dos resultados. No **ranking completo**, o modelo atribui uma pontuação a cada item do catálogo (aproximadamente 406 mil itens canônicos, conforme a deduplicação descrita na Seção 4) e o item de teste é posicionado nesse ranking global. Esse regime reflete o cenário real de produção e fornece uma estimativa não enviesada do desempenho [Krichene and Rendle 2020]. Na **amostragem negativa com 100 candidatos**, o item de teste é posicionado contra apenas 99 itens sorteados aleatoriamente do catálogo. Esse regime é amplamente utilizado na literatura por ser computacionalmente mais ba-

rato [Kang and McAuley 2018, Sun et al. 2019], mas Krichene e Rendle demonstraram formalmente que as métricas resultantes são inconsistentes com suas versões exatas: a conclusão de que um modelo A supera B sob amostragem não implica a mesma relação sob ranking completo.

A qualidade dos modelos é avaliada por três métricas de ranking com $K \in \{5, 10, 20\}$ [Ludewig et al. 2021]. O **Recall@K** mede a proporção de sessões em que o item de teste aparece entre os K primeiros; no esquema *Leave-One-Out* com um único item relevante, coincide com o Hit Rate@K. O **MRR@K** computa a média do inverso da posição do item de teste: se ele aparece na posição $r \leq K$, a contribuição é $1/r$, caso contrário zero. O **NDCG@K** aplica desconto logarítmico à posição, com contribuição $1/\log_2(r+1)$, penalizando acertos em posições mais baixas.

5.3. Configuração de Treinamento

Para os modelos baseados em aprendizado (todos exceto Random, Pop e ItemKNN), o treinamento utiliza *early stopping* com paciência de 10 épocas sobre o MRR@10 de validação e limite máximo de 80 épocas (convergência média: 60 épocas), em uma GPU NVIDIA RTX 4090. A validação é feita sob amostragem negativa (100 candidatos), pois avaliar o catálogo completo a cada época seria impraticável; o estado com melhor MRR@10 de validação é então submetido aos dois regimes na avaliação final. Esse estado pode não ser ótimo sob ranking completo, mas a restrição é idêntica para todos os modelos, preservando a comparabilidade. Os hiperparâmetros seguem os artigos originais e os valores padrão do RecBole, sem busca em grade [Ludewig et al. 2021]; todos os modelos usam *embeddings* de dimensão 64, exceto Caser (dimensão 8). A configuração completa está disponível no repositório.

6. Resultados Experimentais

6.1. Resultados sob Ranking Completo e Amostragem Negativa

A Tabela 2 apresenta os resultados sob ambos os regimes de avaliação. Sob ranking completo, LightGCN alcança o melhor MRR@10, indicando maior precisão nas primeiras posições do ranking. Já FOSSIL apresenta os melhores valores de NDCG@10 e Recall@10, sugerindo maior capacidade de recuperar o item relevante dentro do Top-10. Cabe notar que a diferença entre os dois primeiros colocados em MRR@10 (0,0953 vs. 0,0951) está dentro da variabilidade mensal observada na Tabela 3, de modo que não é possível afirmar superioridade estatística de um sobre o outro nessa métrica. De maneira geral, modelos baseados em vizinhança e fatoração (ItemKNN, FOSSIL) permanecem altamente competitivos, superando diversas arquiteturas neurais profundas. Modelos sequenciais como SASRec e BERT4Rec apresentam desempenho razoável, porém sem superioridade consistente neste cenário mais rigoroso. Resultados para $K = 5$ e $K = 20$ apresentaram hierarquia consistente com $K = 10$ e estão disponíveis no repositório do projeto[2].

Sob amostragem negativa, valores de MRR@10 superiores a 0,8 tornam-se frequentes, e modelos neurais sequenciais (SASRec, GRU4Rec, NARM) tornam-se mais competitivos, enquanto métodos de vizinhança (ItemKNN, BPR) perdem vantagem relativa. Essa inversão decorre da simplificação do problema com conjunto reduzido de candidatos,

[2] `https://anonymous.4open.science/r/anonrecsys-E814`

Tabela 2. Resultados agregados (média sobre 4 meses) por modelo sob ranking completo e amostragem negativa. Modelos ordenados por MRR@10 sob ranking completo.

		Ranking Completo			Amostragem Negativa		
Modelo	**Família**	MRR@10	NDCG@10	Recall@10	MRR@10	NDCG@10	Recall@10
LightGCN	Grafo	**0,0953**	0,1218	0,2088	0,8161	0,8436	0,9277
FOSSIL	Lat./Trans.	0,0951	**0,1366**	**0,2721**	**0,8793**	**0,9022**	0,9719
ItemKNN	Vizinhança	0,0900	0,1165	0,2038	0,5452	0,5637	0,6249
BPR-MF	Fat. Mat.	0,0881	0,1016	0,1455	0,5301	0,5755	0,7204
SASRec	Neural seq.	0,0795	0,1223	0,2603	0,8700	0,8931	0,9634
BERT4Rec	Neural seq.	0,0775	0,1071	0,2046	0,8040	0,8353	0,9317
GRU4Rec	Neural seq.	0,0752	0,1024	0,1920	0,8639	0,8905	0,9709
NARM	Neural seq.	0,0637	0,0876	0,1666	0,8619	0,8898	**0,9739**
TransRec	Lat./Trans.	0,0561	0,0822	0,1678	0,8431	0,8751	0,9726
STAMP	Neural seq.	0,0543	0,0761	0,1486	0,8213	0,8543	0,9554
FPMC	Lat./Trans.	0,0492	0,0646	0,1154	0,6248	0,6708	0,8157
Caser	Neural seq.	0,0111	0,0161	0,0324	0,6141	0,6889	0,9231
Pop	Não pers.	0,0015	0,0021	0,0041	0,1668	0,2178	0,3865
NextItNet	Neural seq.	0,0015	0,0021	0,0042	0,1648	0,2156	0,3831
Random	Não pers.	0,0000	0,0000	0,0001	0,0290	0,0450	0,0990

confirmando que protocolos baseados em amostragem negativa tendem a superestimar o desempenho e podem alterar a conclusão sobre qual arquitetura é mais adequada.

6.2. Análise de Estabilidade Temporal

Como o catálogo de um portal imobiliário muda continuamente, um modelo que apresenta bom desempenho médio mas oscila muito entre meses pode ser inviável em produção. A Figura 2 ilustra essa dimensão para os quatro modelos mais competitivos sob ranking completo: o LightGCN parte abaixo do FOSSIL em março e só assume a liderança em maio, evidenciando que a vantagem média não se sustenta em todos os recortes. Modelos como ItemKNN e SASRec, em contrapartida, exibem trajetórias mais estáveis ao longo dos meses.

A Tabela 3 quantifica essa variação para todos os modelos. Entre os competitivos, SASRec apresenta a menor oscilação (desvio padrão de 0,0010), seguido por NARM e FOSSIL (ambos com desvio padrão inferior a 0,0015). O LightGCN, apesar de liderar em MRR@10 médio, exibe desvio padrão mais de quatro vezes superior, indicando maior sensibilidade às variações mensais do catálogo. Sob amostragem negativa, essa diferença praticamente desaparece, sugerindo que a instabilidade emerge apenas quando o ranking é calculado sobre o catálogo completo. STAMP também apresenta oscilação elevada, possivelmente por sua dependência do último item da sessão.

7. Discussões e Limitações

Entre os achados do trabalho, destaca-se a inversão de ranking entre os dois regimes de avaliação: os mesmos *checkpoints* que colocam LightGCN e FOSSIL no topo sob catálogo completo rebaixam esses modelos quando avaliados sob amostragem negativa, onde modelos neurais sequenciais passam a ocupar as primeiras posições. Esse padrão já foi documentado em *benchmarks* de e-commerce [Krichene and Rendle 2020], reforçando

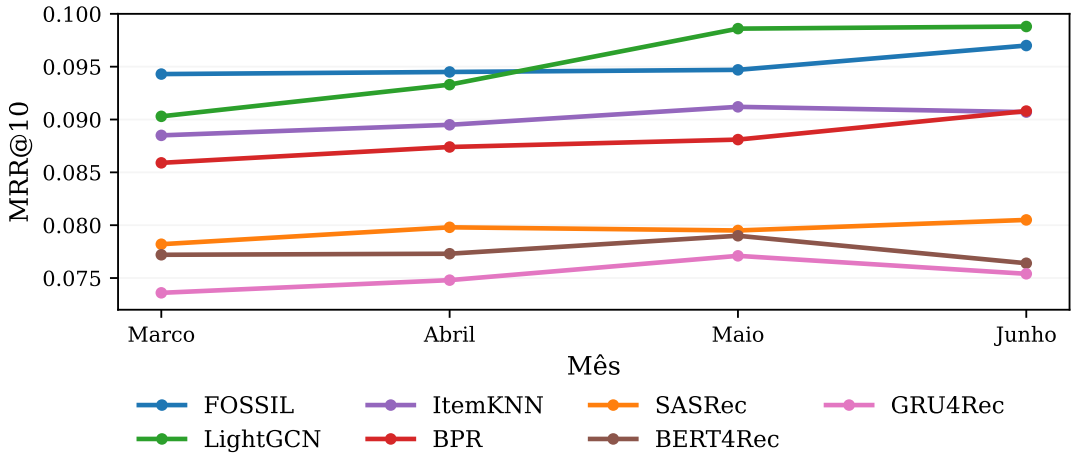

Figura 2. MRR@10 mensal dos sete modelos mais competitivos sob ranking completo. O LightGCN só assume a liderança a partir de maio, enquanto FOSSIL, ItemKNN e SASRec apresentam variação mensal mais contida.

que comparações baseadas exclusivamente em amostragem negativa podem conduzir a conclusões enviesadas sobre superioridade arquitetural.

Uma possível explicação para a competitividade de métodos baseados em vizinhança e fatoração sob catálogo completo reside na natureza das sessões imobiliárias. Com mediana de apenas três interações e alta proporção de sessões anônimas, o sinal disponível por sessão é limitado. Nesse regime, modelos que exploram padrões locais de coocorrência (como ItemKNN e FOSSIL) fazem menos suposições sobre a sequência e por isso sofrem menos com dados escassos, enquanto arquiteturas profundas, com maior capacidade paramétrica, podem tornar-se mais suscetíveis ao sobreajuste em sequências curtas e ruidosas. O caso extremo são Caser e NextItNet, cujo desempenho ficou próximo ao *baseline* aleatório. Ambos dependem de filtros convolucionais que foram projetados para operar sobre sequências substancialmente mais longas do que a mediana de 3 interações observada neste *dataset*. Na prática, a maior parte da entrada desses modelos consiste em vetores de *padding*, o que compromete a capacidade de capturar padrões úteis.

A análise temporal complementa a comparação por métricas médias. A diferença entre o primeiro e o segundo colocados em MRR@10 é de apenas 0,0002, enquanto o desvio padrão mensal do LightGCN (0,0042) é mais de vinte vezes superior a essa margem, o que significa que a liderança pode se inverter em qualquer mês individual. Ao reconstruir representações globais a cada retreinamento mensal, o modelo se torna sensível a variações na composição do catálogo, efeito esperado em portais onde anúncios entram e saem diariamente. Em contrapartida, SASRec, NARM e FOSSIL exibem desvio padrão abaixo de 0,0015, indicando predições mais estáveis. No portal Zap Imóveis, onde o retreinamento é periódico e a experiência do usuário precisa ser consistente entre ciclos, essa estabilidade importa. Um modelo que oscila muito entre meses pode gerar recomendações percebidas como erráticas, mesmo quando suas métricas agregadas são competitivas.

Do ponto de vista do mercado imobiliário, os resultados indicam que portais com sessões curtas, predominantemente anônimas e catálogo de grande escala podem obter desempenho competitivo com modelos de baixa complexidade. O FOSSIL, em particular,

Tabela 3. Estabilidade temporal (MRR@10) sobre 4 recortes mensais, sob ranking completo e amostragem negativa.

		Ranking Completo		Amostragem Negativa	
Modelo	**Família**	Média	Desvio Padrão	Média	Desvio Padrão
LightGCN	Grafo	0,0953	0,0042	0,8161	0,0059
FOSSIL	Lat./Trans.	0,0951	0,0013	0,8793	0,0045
ItemKNN	Vizinhança	0,0900	0,0012	0,5452	0,0064
BPR-MF	Fat. Mat.	0,0881	0,0021	0,5301	0,0040
SASRec	Neural seq.	0,0795	**0,0010**	0,8700	0,0060
BERT4Rec	Neural seq.	0,0775	0,0011	0,8040	0,0084
GRU4Rec	Neural seq.	0,0752	0,0015	0,8639	0,0056
NARM	Neural seq.	0,0637	0,0008	0,8619	**0,0045**
TransRec	Lat./Trans.	0,0561	0,0009	0,8431	0,0062
STAMP	Neural seq.	0,0543	0,0032	0,8213	0,0072
FPMC	Lat./Trans.	0,0492	0,0010	0,6248	0,0098
Caser	Neural seq.	0,0111	0,0058	0,6141	0,0797
Pop	Não pers.	0,0015	0,0008	0,1668	0,0072
NextItNet	Neural seq.	0,0015	0,0008	0,1648	0,0058
Random	Não pers.	0,0000	0,0000	0,0290	0,0002

apresenta o melhor Recall@10 sob ranking completo, estabilidade temporal elevada e custo computacional reduzido, o que o sugere como candidato promissor para cenários de produção em que retreinamentos frequentes são necessários. LightGCN é uma alternativa quando o foco recai sobre precisão nas primeiras posições do ranking, desde que se aceite maior oscilação entre ciclos de retreinamento.

Uma limitação inerente a este *benchmark* é a exclusão de atributos explícitos dos imóveis. Considerando que filtros rígidos (preço, localização) guiam a navegação neste domínio [DataZAP 2025], a incorporação dessas variáveis contextuais pode alterar a hierarquia de modelos observada e é o próximo passo natural deste trabalho. Além disso, a seleção do estado do modelo via MRR@10 sob amostragem negativa pode afetar modelos de forma assimétrica: é possível que, com recursos para otimizar diretamente sob ranking completo, a vantagem relativa dos modelos baseados em vizinhança e fatoração se acentue ou se reduza. Verificar essa hipótese exigiria avaliar o catálogo inteiro a cada época de validação, o que não foi viável com a infraestrutura disponível. Cabe mencionar também que cada configuração experimental (15 modelos $\times$ 4 meses) foi executada uma única vez. O custo total dos experimentos foi de aproximadamente 450 horas de GPU em uma NVIDIA RTX 4090, o que tornou impraticável múltiplas execuções. Embora isso impeça a construção de intervalos de confiança, o retreinamento mensal independente oferece uma forma parcial de avaliar a variabilidade dos resultados. Por fim, o *dataset* é proprietário, o que restringe a reprodutibilidade direta, embora o código do protocolo e da avaliação esteja disponível publicamente. Adicionalmente, a filtragem de sessões por segmento de imóvel (e.g., apartamentos em uma determinada região) pode viabilizar análises mais aprofundadas no domínio imobiliário, reduzindo o espaço de busca e expondo padrões de recomendação específicos a cada nicho. A validação em cenário online (testes A/B) e a extensão a outros portais ficam como trabalhos futuros.

8. Conclusão

Este trabalho apresentou um *benchmark* para recomendação baseada em sessão no mercado imobiliário brasileiro, comparando 15 modelos sobre um *dataset* de 17,6 milhões de interações do portal Zap Imóveis. O protocolo proposto combina deduplicação de anúncios por *fingerprinting*, particionamento *Leave-One-Out* com retreinamento mensal independente e avaliação sob ranking completo e amostragem negativa. Essa configuração permitiu isolar o efeito do regime de avaliação sobre a ordenação entre modelos.

Sob ranking completo, modelos baseados em fatoração e vizinhança (FOSSIL e LightGCN) superam arquiteturas neurais profundas; sob amostragem negativa, essa ordenação se inverte parcialmente. A divergência confirma, em um domínio até então não investigado, a inconsistência de métricas amostradas documentada por Krichene e Rendle (2020). Caser e NextItNet ficaram próximos do *baseline* aleatório, resultado coerente com sessões cuja mediana é de 3 interações. A análise de estabilidade temporal mostrou que o LightGCN, apesar de liderar em MRR@10 médio, apresenta variação mensal bem superior à dos modelos sequenciais competitivos, instabilidade que só se manifesta sob ranking completo.

Em síntese, este trabalho oferece duas contribuições não reportadas em estudos anteriores no domínio imobiliário: a demonstração de que o regime de avaliação pode inverter a hierarquia entre modelos no domínio imobiliário brasileiro; e a evidência de que métodos de baixa complexidade, como FOSSIL, combinam desempenho competitivo com estabilidade temporal, sugerindo-os como candidatos promissores para portais de larga escala. Esses achados indicam que a prática corrente de avaliar modelos sob amostragem negativa pode estar mascarando a vantagem de métodos mais simples em domínios com catálogos grandes. Permanece em aberto se a incorporação de atributos rígidos (preço, localização) como *features* de entrada inverteria novamente essa hierarquia. O protocolo e o código disponibilizados podem servir de base para futuros *benchmarks* em domínios com características similares.

Agradecimentos

Os autores agradecem ao Grupo OLX pela disponibilização dos dados utilizados neste trabalho, conforme Termo de Autorização de Uso de Dados para Pesquisa Acadêmica firmado entre as partes. Os autores agradecem o financiamento parcial das seguintes agências: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001; Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) - Códigos 442344/2023-6, 444602/2024-0 e 407466/2025-8; Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES) (2023-RWXSZ, 2024-9G1WP, 2025-1H3FP e 2026-B74MN); e MCTI/Fapesp/CGI.br (Porvir-5G, 2020/05182-3). Os autores também declaram o uso de ferramenta de inteligência artificial generativa exclusivamente para fins de revisão ortográfica e gramatical do texto, permanecendo os autores como únicos responsáveis por todo o conteúdo intelectual e científico gerado neste trabalho.

Referências

Brasil. Ministério da Fazenda (2025). Relatório de distribuição de renda 2025. Acesso em: 27 fev. 2026.

DataZAP (2025). Anuário do mercado imobiliário 2025. Grupo OLX.

Deshpande, M. and Karypis, G. (2004). Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems (TOIS)*, 22(1):143–177.

Domingues, M. A., de Moura, E. S., Marinho, L. B., and da Silva, A. (2023). A large scale benchmark for session-based recommendations in the legal domain. *Artificial Intelligence and Law*.

Gharahighehi, A., Pliakos, K., and Vens, C. (2021). Recommender systems in the real estate market—a survey. *Applied Sciences*, 11(16).

He, R., Kang, W.-C., and McAuley, J. (2017). Translation-based recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, RecSys '17, page 161–169, New York, NY, USA. Association for Computing Machinery.

He, R. and McAuley, J. (2016). Fusing similarity models with markov chains for sparse sequential recommendation. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 191–200. IEEE.

He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., and Wang, M. (2020). Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 639–648.

Hidasi, B., Karatzoglou, A., Baltrunas, L., and Tikk, D. (2016). Session-based recommendations with recurrent neural networks. In Bengio, Y. and LeCun, Y., editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Instituto Brasileiro de Geografia e Estatística (IBGE) (2025). Contas nacionais trimestrais: 4º trimestre de 2024. Technical report, IBGE, Rio de Janeiro. Divulgado em 07 de março de 2025.

Kang, W.-C. and McAuley, J. (2018). Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 197–206.

Krichene, W. and Rendle, S. (2020). On sampled metrics for item recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1748–1757, New York, NY, USA. Association for Computing Machinery.

Li, J., Ren, P., Chen, Z., Ren, Z., Lian, T., and Ma, J. (2017). Neural attentive session-based recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 1419–1428, New York, NY, USA. Association for Computing Machinery.

Lima, M., Silva, E., and da Silva, A. (2024). Um estudo sobre o uso de modelos de linguagem abertos na tarefa de recomendação de próximo item. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 510–522, Porto Alegre, RS, Brasil. SBC.

Liu, Q., Zeng, Y., Mokhosi, R., and Zhang, H. (2018). Stamp: short-term attention/memory priority model for session-based recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1831–1839.

Ludewig, M. and Jannach, D. (2018). Evaluation of session-based recommendation algorithms. *User Model. User-adapt Interact.*, 28(4-5):331–390.

Ludewig, M., Mauro, N., Latifi, S., and Jannach, D. (2021). Empirical analysis of session-based recommendation algorithms. *User Model. User-adapt Interact.*, 31(1):149–181.

Polohakul, J., Chuangsuwanich, E., Suchato, A., and Punyabukkana, P. (2021). Real estate recommendation approach for solving the item cold-start problem. *IEEE Access*, 9:68139–68150.

Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-Thieme, L. (2009). Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th conference on uncertainty in artificial intelligence*, pages 452–461.

Rendle, S., Freudenthaler, C., and Schmidt-Thieme, L. (2010). Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 811–820.

Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295.

Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., and Jiang, P. (2019). Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1441–1450, New York, NY, USA. Association for Computing Machinery.

Tang, J. and Wang, K. (2018). Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 565–573.

Xu, L., Tian, Z., Zhang, G., Zhang, J., Wang, L., Zheng, B., Li, Y., Tang, J., Zhang, Z., Hou, Y., Pan, X., Zhao, W. X., Chen, X., and Wen, J. (2023). Towards a more user-friendly and easy-to-use benchmark library for recommender systems. In *SIGIR*, pages 2837–2847. ACM.

Yuan, F., Karatzoglou, A., Arapakis, I., Jose, J. M., and He, X. (2019). A simple convolutional generative network for next item recommendation. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 582–590.