

Otimização da Distribuição de Leads via Aprendizado por Reforço: Uma Abordagem com Multi-Armed Bandits em um Ambiente de Vendas Simulado

Enzo Pinheiro Pierazolli

Departamento de Ciência da Computação

Universidade Federal de Minas Gerais

Belo Horizonte, Brasil

eppierazolli@gmail.com

Abstract—Este trabalho apresenta o desenvolvimento de um ambiente de vendas simulado para a otimização da distribuição de leads por meio de técnicas de *Multi-Armed Bandits* (MAB). Partindo da observação de que CRMs comerciais normalmente aplicam regras fixas (*Round-Robin*, *score-based*, etc.), propõe-se modelar um ambiente que permita comparar heurísticas tradicionais com algoritmos adaptativos de MAB (ϵ -Greedy, ϵ -Decaying, UCB1, UCB-Tuned, Gradient Bandit e Thompson Sampling). A contribuição principal é (i) um ambiente de simulação parametrizável que reproduz variações realistas de demanda e heterogeneidade entre vendedores, incorporando uma penalidade de carga que degrada a conversão em função da ocupação do vendedor; (ii) a implementação e integração de nove estratégias de alocação, incluindo heurísticas fixas e políticas adaptativas, que atualizam suas decisões a partir do desempenho observado e (iii) avaliação experimental que mensura taxa de conversão e concentração de carga em diferentes cenários. Resultados indicam que políticas adaptativas superam heurísticas estáticas especialmente quando há heterogeneidade significativa entre vendedores e o sistema opera sem saturação; em contrapartida, sob alta demanda ou vendedores homogêneos, estratégias de distribuição uniforme demonstram robustez.

I. INTRODUÇÃO

A. Caracterização do problema

Empresas geralmente lidam com um fluxo contínuo de contatos de potenciais clientes, denominados leads. A eficiência do processo comercial depende fortemente de como esses leads são distribuídos entre os vendedores: uma má distribuição resulta em perda de receita, baixa taxa de conversão e alocação injusta de trabalho. Sistemas de Customer Relationship Management (CRM), plataformas responsáveis pela gestão dos leads e pelo suporte às operações comerciais, normalmente implementam heurísticas simples de distribuição, como *round-robin* e *score-based*, que podem não se adaptar bem a cenários dinâmicos nos quais: (i) vendedores apresentam habilidades distintas e variáveis; (ii) leads têm perfis heterogêneos (valor, dificuldade, tempo de atendimento); (iii) a demanda oscila (sazonalidade, campanhas, picos) e (iv) existe capacidade limitada por vendedor (número simultâneo de atendimentos, tempo por lead).

B. Motivação

Métodos estáticos não capturam a necessidade de aprender com o histórico de desempenho nem de reagir a mudanças operacionais. Algoritmos de *Multi-Armed Bandits* (MAB) oferecem um trade-off explícito entre testar vendedores menos selecionados ou com maior incerteza sobre seu desempenho, e exploração do conhecimento acumulado, atribuindo leads aos vendedores de melhor performance conhecida. Adicionalmente, em ambientes nos quais a taxa de conversão de um vendedor é penalizada pela sua ocupação, concentrar leads indiscriminadamente nos vendedores mais eficientes pode ser prejudicial, tornando o problema de alocação não trivial e reforçando a necessidade de políticas que equilibrem desempenho e carga. Aplicar MABs na distribuição de leads pode, portanto, aumentar conversões, adaptar políticas automaticamente a mudanças no ambiente e ajustar a distribuição de carga conforme o desempenho observado, pontos de relevante impacto prático.

II. REFERENCIAL TEÓRICO E TRABALHOS RELACIONADOS

Os referenciais teóricos apresentados nesta seção conectam-se diretamente ao problema analisado, a distribuição adaptativa de leads em ambientes de CRM, servindo como base para a formulação do problema e para a escolha metodológica.

A. *Multi-Armed Bandits*

O problema de *Multi-Armed Bandits* (MAB) modela um cenário de decisão sequencial no qual um agente deve selecionar repetidamente entre múltiplas ações, cada uma associada a recompensas estocásticas e desconhecidas. O desafio central consiste no dilema entre *exploration*; testar ações pouco conhecidas para reduzir a incerteza sobre seu desempenho; e *exploitation*; selecionar a ação de melhor desempenho observado. Uma formulação rigorosa desse dilema, bem como algoritmos fundamentais para resolvê-lo, é apresentada em Sutton e Barto [1].

Esse arcabouço teórico se adequa a sistemas de distribuição de leads, nos quais cada vendedor pode ser interpretado como um braço do bandit, a atribuição de um lead como uma ação e a conversão resultante como a recompensa observada. No contexto do modelo proposto, o desempenho dos vendedores em diferentes cenários é tratado como desconhecido a priori e precisa ser inferido a partir das interações.

B. Algoritmos Clássicos de MAB

Diversos algoritmos foram propostos para resolver o problema de MAB. Entre os mais estudados e aplicados encontram-se:

- **ϵ -Greedy**: método simples que seleciona a melhor ação observada com probabilidade $1 - \epsilon$ e realiza exploração aleatória com probabilidade ϵ [1]. Uma variante relevante é o ϵ -Decaying, no qual ϵ decresce ao longo do tempo, priorizando exploração nas etapas iniciais e convergindo progressivamente para *exploitation* [6].
- **Upper Confidence Bound (UCB)**: algoritmos que combinam a média empírica de recompensa com um termo de incerteza para favorecer ações menos exploradas. [1]. O UCB-Tuned [6] estende essa formulação incorporando a variância amostral ao termo de confiança, tornando a política mais conservadora em braços com alta variabilidade de recompensa.
- **Gradient Bandits**: políticas que ajustam preferências sobre as ações com base em gradientes estimados das recompensas, atribuindo probabilidades via softmax e permitindo comportamento totalmente estocástico [1].
- **Thompson Sampling**: abordagem bayesiana que mantém uma distribuição de probabilidade sobre a recompensa esperada de cada braço e seleciona ações por amostragem dessas distribuições [8].

Essas abordagens formam a base dos algoritmos avaliados neste trabalho.

C. Estratégias de Referência (Baselines)

Para permitir a comparação com abordagens sem aprendizado, três estratégias de referência foram incluídas na avaliação:

- **Random**: seleciona um vendedor disponível de forma uniformemente aleatória, sem considerar histórico de desempenho. Por não incorporar nenhum mecanismo de aprendizado, serve como limite inferior de desempenho e referência para quantificar o ganho mínimo de qualquer política adaptativa.
- **Round-Robin**: distribui os leads de forma circular e determinística entre os vendedores disponíveis, garantindo equidade máxima na alocação de carga independentemente do desempenho individual. Sendo, portanto, uma abordagem que oferece simplicidade e previsibilidade.
- **Greedy**: após uma fase inicial de exploração forçada, na qual os primeiros leads são distribuídos aleatoriamente para que todos os vendedores sejam observados ao menos uma vez, passa a selecionar sempre o vendedor de maior taxa de conversão observada, sem exploração adicional.

Por escolher de forma determinística, é vulnerável à convergência prematura a vendedores subótimos e tende a concentrar a carga.

D. Bandits em Atendimento e Encaminhamento de Clientes

Pesquisas recentes mostram que técnicas de MAB são adequadas para problemas de encaminhamento e alocação de clientes, que apresentam dinamicidade, heterogeneidade entre agentes e restrições operacionais. Wen, Zeng e Liu [2] propõem um *contextual bandit* não paramétrico com controle de elegibilidade em nível de braço para direcionar contatos a especialistas de suporte. O método, denominado K-Boot, é avaliado em dados sintéticos e em um dataset da Amazon, demonstrando desempenho superior a políticas fixas tradicionais.

Os resultados reforçam a adequação de modelos de bandits para contextos similares ao presente trabalho, como a distribuição de leads comerciais, nos quais agentes possuem habilidades distintas e disponibilidade variável.

E. Bandits em Precificação e Decisões Comerciais

Outra linha relevante de pesquisa aplica MAB em precificação dinâmica. Trovò et al. [3] investigam variantes de UCB adaptadas para ambientes online de vendas, utilizando conhecimento prévio sobre ordem de grandeza das taxas de conversão para guiar a política. Os resultados mostram ganhos significativos frente a políticas genéricas de MAB e demonstram que incorporar características específicas do domínio pode melhorar substancialmente o desempenho dos algoritmos.

Esse tipo de investigação fornece diretrizes metodológicas úteis para modelagem e análise de algoritmos em ambientes corporativos simulados, como o proposto neste trabalho.

F. Síntese e Relação com o Presente Trabalho

A literatura demonstra que:

- algoritmos de MAB são adequados para ambientes dinâmicos e estocásticos;
- lidam naturalmente com heterogeneidade entre agentes e incerteza;
- permitem aprendizado incremental e adaptação contínua;
- podem incorporar restrições operacionais, como capacidade e disponibilidade.

Esses aspectos sustentam a escolha metodológica deste trabalho: o uso de algoritmos de MAB, tradicionalmente aplicados em sistemas online, para resolver o problema de distribuição adaptativa de leads em um ambiente de simulação projetado para capturar dinamicidade, heterogeneidade e restrições típicas de operações comerciais reais.

Entretanto, é válido destacar que o ambiente de simulação incorpora um fenômeno operacional relevante que não se

ajusta perfeitamente ao arcabouço tradicional de Multi-Armed Bandits: a variação da taxa de conversão em função da carga de trabalho do vendedor (isto é, da quantidade de leads simultâneos sob sua responsabilidade). Esse efeito introduz dependência temporal e não-estacionaridade estruturada, pois a recompensa esperada de um “braço” deixa de ser fixa e passa a depender do estado interno do vendedor. Embora tal característica seja realista em processos comerciais, ela ultrapassa a suposição clássica de recompensas i.i.d. do MAB e aproxima o problema de formulações mais gerais, como restless bandits, bandits com estado ou reinforcement learning contextual. Essa observação é relevante tanto para a interpretação dos resultados quanto para o direcionamento de extensões futuras do trabalho

III. MODELAGEM DO AMBIENTE DE SIMULAÇÃO

Esta seção apresenta os principais componentes que constituem o ambiente de simulação desenvolvido, incluindo a representação dos leads, dos vendedores e da dinâmica temporal responsável pela geração, alocação e processamento dos leads. A simulação opera em unidades discretas de tempo chamadas *steps*, que representam ciclos completos de operação nos quais todo o fluxo de chegada, distribuição, processamento e atualização de políticas ocorre.

A. Lead

Um *lead* corresponde a uma oportunidade de venda individual. Cada lead possui um fator opcional de dificuldade que pode ajustar sua probabilidade efetiva de conversão, bem como uma duração de atendimento sorteada aleatoriamente em uma janela discreta, que determina quantos steps são necessários até sua conclusão. A conversão só é avaliada ao final desse período, resultando em um desfecho binário: convertido ou não convertido. Nos experimentos realizados neste trabalho, o fator de dificuldade não é utilizado, mantendo a conversão dependente exclusivamente das características do vendedor.

B. Seller

Cada vendedor é modelado como um agente independente com capacidade limitada de atendimento simultâneo. Sua taxa de conversão base representa sua habilidade individual e, nessa modelagem, foi sorteada de uma distribuição normal $\mathcal{N}(\mu, \sigma^2)$, onde μ é a taxa média e σ controla a heterogeneidade entre vendedores. O número de leads que pode atender em paralelo é delimitado por uma capacidade máxima C . Um aspecto central da modelagem é o mecanismo de penalização por carga:

$$\text{conversão efetiva} = \text{conversão base} \cdot e^{-2 \cdot \text{load factor}}$$

onde o *load factor* é definido como a razão entre o número de leads ativos e a capacidade total do vendedor. Esse fator modela de forma simples, suave e monotônica

o efeito de saturação: conforme o vendedor se aproxima de sua capacidade, sua taxa efetiva de conversão decresce, reproduzindo fenômenos comuns em operações reais como fadiga, sobrecarga e aumento de tempo de resposta.

C. Ambiente de Distribuição de Leads

O ambiente coordena a dinâmica global da simulação. Cada step representa uma unidade de tempo na qual todo o ciclo operacional ocorre:

- 1) **Geração de leads:** novos leads são gerados em cada step em função da taxa média de chegada λ .
- 2) **Alocação:** cada lead é encaminhado a um vendedor pela estratégia ativa. Vendedores sem capacidade disponível ou que atingiram o limite de alocações por step L são excluídos temporariamente. Se todos os vendedores forem excluídos, o lead é classificado como perdido.
- 3) **Processamento:** os vendedores avançam o atendimento dos leads ativos, finalizando aqueles cujo tempo se esgotou e realizando a tentativa de conversão com base na taxa ajustada por carga.
- 4) **Atualização da política:** para cada lead concluído, é computada uma recompensa binária (1 para conversão, 0 caso contrário), utilizada pela estratégia para atualizar sua política de seleção.

Todas as estratégias seguem o mesmo contrato de interface: recebem a lista de vendedores disponíveis, selecionam um deles e atualizam sua política com a recompensa binária resultante, permitindo comparação direta entre abordagens.

D. Regimes de Demanda no Ambiente Simulado

Os diferentes regimes analisados correspondem a cenários operacionais que também podem ser observados na prática em equipes comerciais. No ambiente simulado, esses cenários são definidos pela relação entre o fluxo de chegada de leads e a capacidade total de atendimento dos vendedores, determinada diretamente pelos parâmetros configurados nos experimentos. Ao variar tais parâmetros, torna-se possível reproduzir situações de baixa, média ou alta demanda, permitindo investigar como cada estratégia se comporta frente a diferentes condições de carga e disponibilidade.

Baixa demanda ocorre quando o volume de leads é substancialmente inferior à capacidade total disponível. Nesse cenário, os vendedores operam com ampla folga, a probabilidade de saturação é baixa e o número de leads perdidos tende a ser nulo.

Demanda média corresponde ao regime em que o fluxo de leads se aproxima da capacidade de atendimento. Nessa condição, alguns vendedores podem atingir seus limites temporariamente, tornando a política de alocação mais relevante para o desempenho global.

Alta demanda caracteriza ambientes em que a chegada de leads supera sistematicamente a capacidade de atendi-

mento. O fluxo eleva a ocupação média dos Sellers, intensificando a penalidade de carga e reduzindo a conversão efetiva. Nesse contexto, a política de alocação torna-se crítica para mitigar gargalos e preservar o desempenho sob pressão operacional.

IV. CONFIGURAÇÃO EXPERIMENTAL

Os experimentos foram projetados para garantir reprodutibilidade e comparabilidade entre estratégias. Todos os cenários são definidos por parâmetros externos; número de vendedores, distribuição das taxas de conversão, capacidade, taxa de chegada de leads e número de steps; sem necessidade de alteração nos mecanismos internos de simulação.

A. Execução dos Experimentos

Cada experimento consiste na execução de simulações, utilizando diferentes sementes aleatórias, para uma combinação específica de parâmetros e estratégia. Em cada step, leads são gerados, alocados, processados e utilizados para atualizar a política da estratégia em avaliação. Ao final, são coletadas as seguintes métricas agregadas:

- **Taxa de conversão:** razão entre leads convertidos e leads concluídos, principal métrica de desempenho.
- **Leads perdidos:** leads não alocados por indisponibilidade de todos os vendedores, indicativo de saturação do sistema.
- **Distribuição de carga:** quantidade de leads recebidos por cada vendedor ao longo da simulação, utilizada para avaliar o grau de concentração da alocação.

Todas as estratégias são avaliadas sob as mesmas condições para cada cenário, garantindo que as comparações sejam diretas e controladas.

B. Cenários Experimentais

Três conjuntos de experimentos foram conduzidos, cada um investigando uma dimensão específica do problema (o experimento de fluxo fornece duas das quatro análises de resultados: a convergência por regime e a conversão agregada):

- 1) **Variação do fluxo de leads (λ):** mantendo fixos o número de Sellers, a capacidade e a heterogeneidade ($\mu = 0,12$, $\sigma = 0,05$), o fluxo de chegada é variado para avaliar como cada estratégia responde a diferentes regimes de demanda.
- 2) **Variação da heterogeneidade (σ):** mantendo fixo o fluxo de leads em regime de baixa demanda, o desvio padrão das taxas de conversão entre vendedores é variado para identificar em que condições estratégias adaptativas superam heurísticas fixas.
- 3) **Concentração de carga:** para um cenário fixo, é analisada a distribuição de leads entre os vendedores ao longo da simulação, permitindo comparar o grau de concentração produzido por cada estratégia.

V. RESULTADOS

Os experimentos foram organizados em quatro análises complementares: convergência por regime de demanda, impacto da heterogeneidade entre vendedores, concentração de carga e taxa de conversão agregada ao longo do fluxo. Em conjunto, essas análises permitem caracterizar o comportamento de cada estratégia sob diferentes condições operacionais.

A. Convergência por Regime de Demanda

As Figuras 1, 2 e 3 apresentam a evolução da taxa de conversão acumulada ao longo dos steps para os regimes de baixa, média e alta demanda, respectivamente.

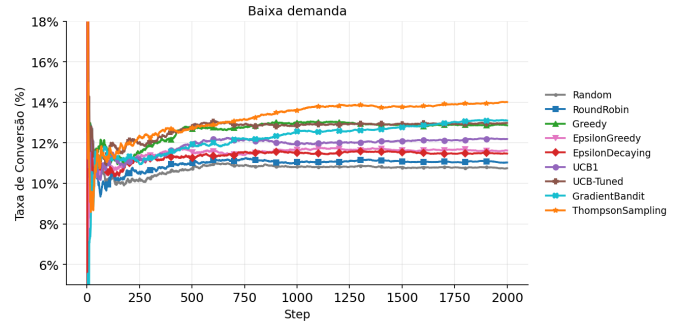


Fig. 1. Taxa de conversão acumulada: baixa demanda.

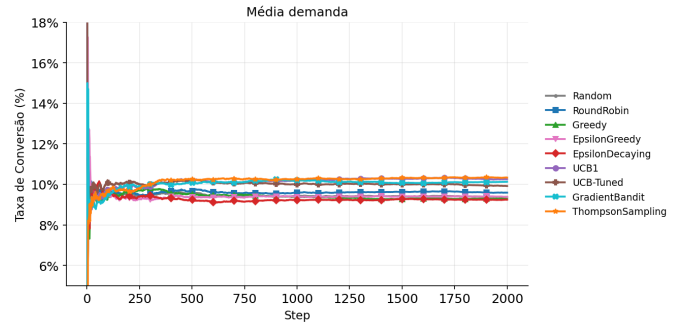


Fig. 2. Taxa de conversão acumulada: média demanda.

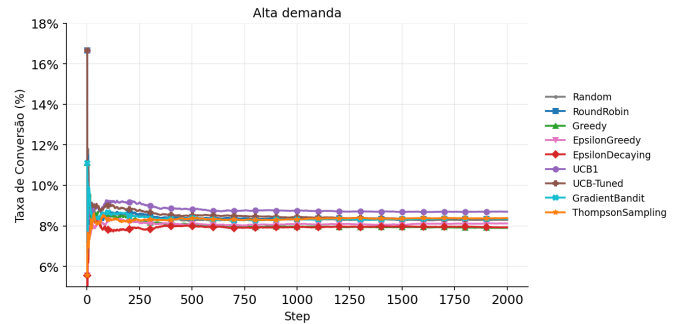


Fig. 3. Taxa de conversão acumulada: alta demanda.

No regime de baixa demanda, o sistema opera com ampla folga de capacidade e ausência de saturação.

Nesse contexto, estratégias que identificam rapidamente os vendedores mais eficazes, como *Thompson Sampling*, *Greedy*, *UCB-Tuned* e *Gradient Bandit*, obtêm as maiores taxas de conversão, beneficiando-se da heterogeneidade entre vendedores sem incorrer no custo de sobrecarga. Estratégias mais uniformes, como *Round-Robin* e *Random*, apresentam desempenho inferior, pois desperdiçam alocações em vendedores de baixa performance.

À medida que a demanda aumenta, o efeito da penalização por carga torna-se progressivamente mais relevante. No regime de média demanda, as diferenças entre estratégias se reduzem e os benefícios de concentrar leads nos melhores vendedores começam a ser prejudicados pelo aumento do fator de carga. No regime de alta demanda, observa-se uma inversão parcial: estratégias que distribuem a carga de forma mais equilibrada, como *Round-Robin* e *UCB1*, demonstram maior robustez, enquanto estratégias concentradoras, como *Thompson Sampling* e ϵ -Decaying, são mais penalizadas pela carga.

B. Impacto da Heterogeneidade entre Vendedores

A Figura 4 apresenta a taxa de conversão em função do desvio padrão (σ) das taxas de conversão entre vendedores, mantendo fluxo e capacidade fixos em regime de baixa demanda.

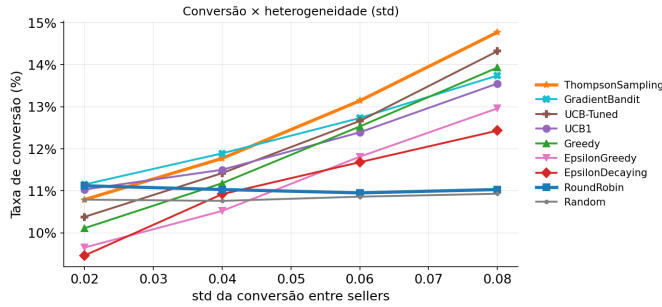


Fig. 4. Taxa de conversão em função da heterogeneidade entre vendedores (σ).

Quando os vendedores são mais homogêneos ($\sigma = 0,02$), há pouca informação a explorar e o *Round-Robin* fica entre as melhores, sem que nenhuma política adaptativa obtenha vantagem significativa. Conforme σ aumenta, estratégias adaptativas passam a identificar e priorizar os vendedores mais eficazes, obtendo ganhos crescentes. Os pontos de cruzamento ocorrem em torno de $\sigma \approx 0,04$, a partir do qual *Thompson Sampling* e *Gradient Bandit* superam consistentemente o *Round-Robin*. Em $\sigma = 0,08$, o *Thompson Sampling* atinge 14,77% de conversão (e o *UCB-Tuned*, 14,32%) contra 11,03% do *Round-Robin*, representando um ganho de até 34%.

C. Concentração de Carga

A Figura 5 apresenta a distribuição percentual de leads recebidos por cada vendedor ao longo da simulação, por meio de um box plot ordenado pelo grau de concentração.

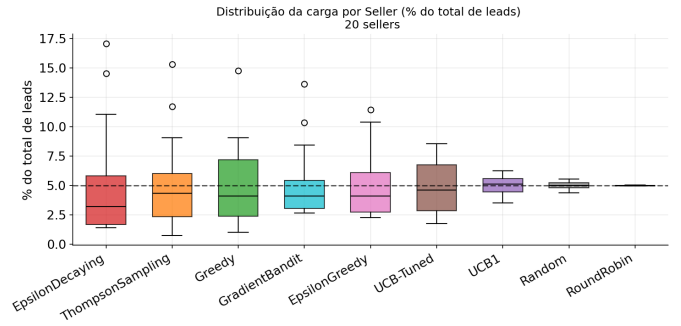


Fig. 5. Distribuição da carga por vendedor (% do total de leads): 20 sellers.

Os resultados revelam um espectro claro entre as estratégias. O *Round-Robin* distribui a carga de forma quase perfeitamente uniforme (Gini $\approx 0,003$), com todos os vendedores recebendo aproximadamente 5% dos leads, o equivalente à fração ideal de $1/N$ para os 20 vendedores deste experimento. No extremo oposto, o ϵ -Decaying apresenta Gini de 0,434, com os 20% de vendedores mais carregados absorvendo 47,8% dos leads. O *Thompson Sampling* apresenta concentração elevada (Gini = 0,369), reflexo de sua tendência em convergir rapidamente para os melhores vendedores. O *UCB1* destaca-se por manter concentração moderada (Gini = 0,093), o que explica sua robustez em cenários de alta demanda.

Esses resultados evidenciam que a concentração é o mecanismo pelo qual estratégias adaptativas obtêm ganhos de conversão e, ao mesmo tempo, o fator que as torna vulneráveis sob alta demanda.

D. Taxa de Conversão Agregada por Fluxo

A Figura 6 consolida a taxa de conversão média de cada estratégia ao longo de toda a varredura de fluxo, permitindo visualizar como o desempenho relativo entre estratégias se altera com o aumento da demanda.

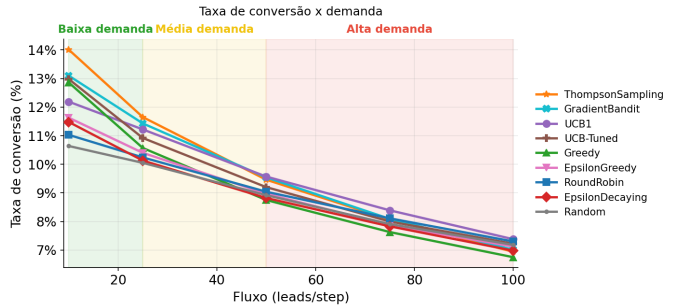


Fig. 6. Taxa de conversão média por estratégia em função do fluxo de leads.

Thompson lidera em fluxo baixo (14,01%); conforme o fluxo sobe, o UCB1 assume a liderança no regime de alta demanda.

VI. CONCLUSÃO

Os resultados evidenciam o potencial de estratégias baseadas em *Multi-Armed Bandits* para a distribuição adaptativa de leads em ambientes comerciais simulados. O *Thompson Sampling* destacou-se em cenários de baixa demanda e alta heterogeneidade entre vendedores, atingindo ganhos relativos de até 34% sobre o *Round-Robin* quando a variação entre vendedores é elevada. O *UCB1* demonstrou comportamento equilibrado entre os regimes, mantendo desempenho competitivo tanto em baixa quanto em alta demanda graças ao seu mecanismo de exploração orientado por incerteza, que naturalmente evita concentração excessiva de carga. O *Gradient Bandit* apresentou desempenho consistente em cenários intermediários, enquanto o *Round-Robin* também mostrou-se uma opção robusta sob alta demanda. Estratégias mais agressivas como o ϵ -*Decaying*, mostraram-se sensíveis ao regime operacional, obtendo os menores resultados nos extremos de demanda avaliados. Esses achados reforçam que a escolha da política de alocação deve considerar o contexto operacional: não há uma estratégia universalmente superior, mas sim um trade-off entre capacidade adaptativa e distribuição equilibrada de carga. O ambiente de simulação modular desenvolvido viabiliza a extensão natural deste trabalho para cenários mais complexos, incluindo a incorporação de algoritmos contextuais que considerem, por exemplo, características individuais dos leads na decisão de alocação, e até mesmo a validação dos achados em dados empíricos de operações comerciais reais.

REFERENCES

- [1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge: MIT Press, 2018.
- [2] Z. Wen, Y. Zeng, and Y. Liu, "A Nonparametric Contextual Bandit with Arm-level Eligibility Control for Customer Service Routing," 2022.
- [3] F. Trovò, L. Cella, A. Nedić, and M. Pontil, "Improving Multi-Armed Bandit Algorithms in Online Pricing Settings," 2018.
- [4] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge: Cambridge University Press, 2020.
- [5] D. M. Negoescu, P. Khosravi, S. Zhao, N. Chen, P. Ahmad, and H. Gonzalez, "Sales Channel Optimization via Simulations Based on Observational Data with Delayed Rewards: A Case Study at LinkedIn," 2022.
- [6] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," 2002.
- [7] A. Badanidiyuru, J. Langford, and A. Slivkins, "Resourceful Contextual Bandits," 2015.
- [8] O. Chapelle and L. Li, "An empirical evaluation of Thompson Sampling," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2011.