Evento: XV Seminário de Inovação e Tecnologia

DESENVOLVIMENTO DE UM MODELO PREDITIVO DE VENDAS EM ENGENHARIA DE SOFTWARE[1]

Mayara de Lourdes Schreiber Meotti[2], Edson Luiz Padoin[3]

[1] Trabalho desenvolvido no Componente Curricular Disciplinar de Estágio.
[2] Estudante do 9º semestre de Engenharia de Software.
[3] Professor orientador da disciplina de Estágio.

INTRODUÇÃO

A área do conhecimento abrange Engenharia de *Software*, Ciência de Dados e Inteligência Artificial (IA), com ênfase em aprendizado de máquina e técnicas de previsão baseadas em *gradient boosting*, destacando o algoritmo XGBoost. O trabalho integra técnicas de pré-processamento de dados, análise de séries temporais e desenvolvimento de soluções computacionais para problemas preditivos, promovendo uma abordagem interdisciplinar aplicada à gestão de vendas. Inicialmente, modelos de redes neurais recorrentes, como o LSTM (*Long Short-Term Memory*), foram considerados devido à sua capacidade de capturar dependências temporais de longo prazo. Contudo, após testes comparativos, o XGBoost foi selecionado como a abordagem principal por sua superioridade em precisão, eficiência computacional e adequação a dados tabulares.

METODOLOGIA

A metodologia adotada visa desenvolver um modelo preditivo de vendas mensais para 2025, com base em dados históricos de vendas diárias de 2022 a 2024, referentes a uma loja específica de uma rede de postos de combustíveis e itens numerados de 1 a 20. O conjunto de dados, contido em um csv autoral, inclui 32.661 registros com variáveis como data da venda (*DATE*), identificador da loja (*STORE*), item (*ITEM*) e quantidade vendida (*SALES*). O pré-processamento envolveu a conversão de datas para o formato datetime, agregação de vendas por data, item e loja para tratar duplicatas, preenchimento de valores ausentes com zero e limitação de outliers utilizando o intervalo interquartil (IQR) expandido para robustez.

A engenharia de *features* focou em capturar padrões temporais e sazonais, incluindo extrações como ano, mês, dia da semana, indicadores de fins de semana e dias desde o início, além de representações cíclicas (seno e cosseno) para periodicidade. Foram adicionadas variáveis binárias para eventos como feriados brasileiros, Páscoa, *Black Friday*, Natal, Ano Novo, Carnaval e dias de pagamento, com ajustes multiplicadores para fins de semana e períodos sazonais. Features de lag (1 a 28 dias), diferenças, médias móveis, desvios padrão e indicadores de tendência, momentum e volatilidade foram criadas, com preenchimento de ausentes pela média por item e loja. O modelo selecionado foi o XGBoost, otimizado via Optuna com validação cruzada de 5 dobras e métrica composta (0,6 MAE + 0,4 RMSE). O treinamento incluiu seleção de features com correlação absoluta >0,03, divisão temporal dos dados (85% treino, 15% teste), 150 experimentos de otimização de hiperparâmetros e treinamento final com parada antecipada.

As previsões para 2025 foram geradas iterativamente para até 365 dias, construindo um DataFrame futuro com *features* replicadas e atualizando lags e rollings dinamicamente com previsões anteriores. A avaliação no conjunto de teste utilizou métricas como RMSE, MAE e R^2, analisando a importância das *features* pelo ganho médio no XGBoost e apresentando estatísticas descritivas das previsões, incluindo média, volatilidade e tendências.

RESULTADOS E DISCUSSÃO

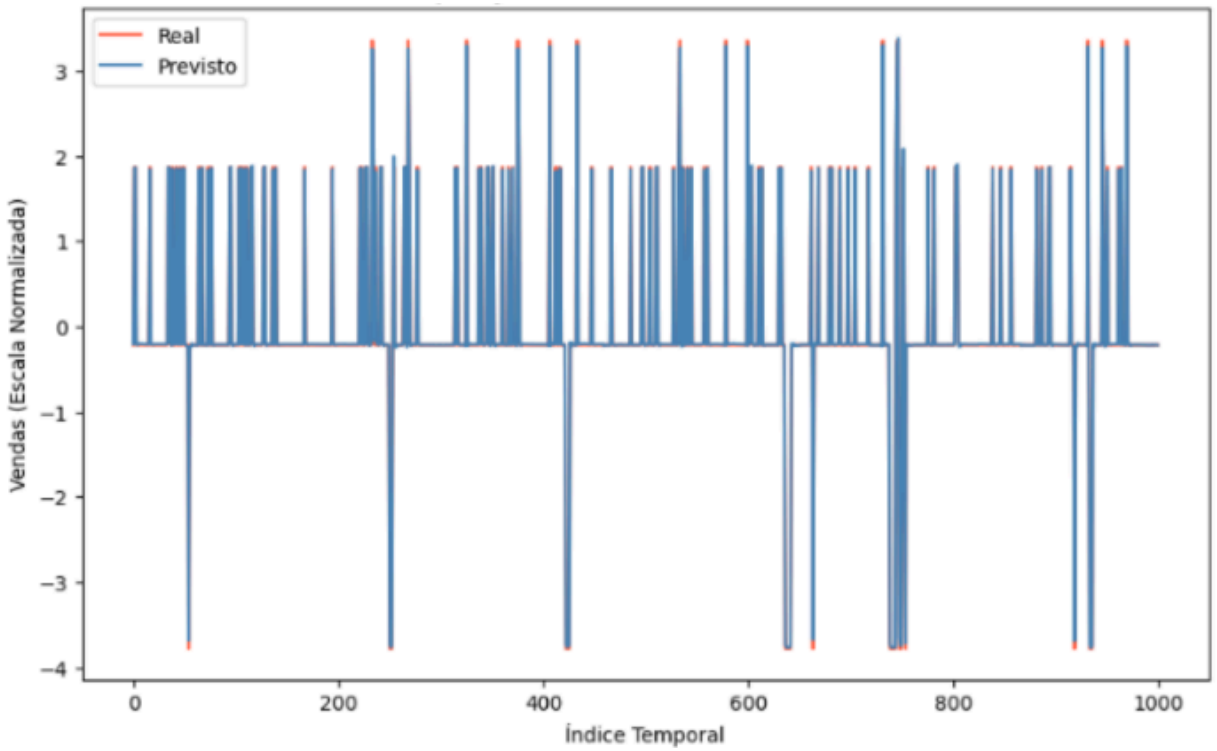
Os resultados do modelo preditivo, baseado no algoritmo XGBoost otimizado com Optuna, para estimar as vendas diárias de unidades em Janeiro de 2025 demonstram alta precisão e robustez. A média das previsões indica uma faixa de 3 a 4 unidades vendidas por dia (unidades totais divididas por 31 dias). Os dados reais registraram 106 unidades vendidas, ou seja, 3,41 unidades/dia. A proximidade entre as previsões e os valores reais reflete um baixo percentual de erro, evidenciando a confiabilidade do modelo.

O gráfico, figura 1, mostra uma ”Comparação entre Vendas Reais e Previstas”, com base no código fornecido, que utiliza um modelo LSTM para prever vendas (*SALES*). O eixo X representa os índices das observações no conjunto de dados ao longo do tempo, após a criação do *dataset* com *createdataset*. O código usa um *timestep* de 14, então cada ponto no eixo X corresponde a uma janela de 14 dias de dados, e os índices (0 a 2000) indicam a

posição das previsões no conjunto de teste ou treino. Como o *TimeSeriesSplit* divide os dados, esses índices refletem a sequência temporal das janelas de previsão.

Já o eixo Y representa os valores das vendas, tanto reais quanto previstas, após a transformação inversa. Assim, os valores no eixo Y (de -4 a 4) são as vendas reais (em vermelho) e previstas (em azul) nessa escala ajustada, com possível variação devido à normalização e ao recorte em 99% (salescap).

Figura 1: Gráfico comparativo.



Fonte: Autoral.

O modelo, avaliado por uma métrica composta (0,6 MAE + 0,4 RMSE), apresentou desempenho robusto. O MSE foi de 0,0004 (± 0,0005) no treinamento e 0,0013 (± 0,0018) no teste, indicando erros quadráticos reduzidos e alta precisão. A diferença entre os conjuntos sugere um leve *overfitting*, mas aceitável, pois o MSE de teste é apenas cerca de três vezes maior que o de treinamento.

O MAE registrou 0,0090 (± 0,0075) no treinamento e 0,0200 (± 0,0219) no teste, mostrando que as previsões desviam, em média, apenas 0,02 unidades de venda no teste. A variabilidade no MAE de teste (± 0,0219) pode indicar heterogeneidade nos dados ou sazonalidades, mas não compromete o desempenho geral. O R2 atingiu 0,9982 (± 0,0024) no

treinamento e 0,9933 (± 0,0095) no teste, valores próximos de 1 que demonstram que o modelo explica quase toda a variabilidade dos dados, com excelente generalização. A pequena diferença entre os R^2 reforça sua robustez.

CONSIDERAÇÕES FINAIS

O projeto não apenas atendeu aos objetivos propostos, como compreender fundamentos de modelos preditivos, realizar pré-processamento de dados e implementar um modelo otimizado, mas também destacou a relevância de soluções baseadas em aprendizado de máquina para a gestão empresarial. A experiência proporcionou o aprimoramento de competências técnicas, como manipulação de dados com Python, otimização de hiperparâmetros com Optuna e análise de séries temporais, além de habilidades profissionais, como trabalho em equipe e comunicação com stakeholders.

Como perspectivas futuras, sugere-se a expansão do modelo para outras unidades da rede, a incorporação de variáveis externas, como preços de mercado ou indicadores econômicos, e a exploração de arquiteturas híbridas, como a integração de redes neurais LSTM com XGBoost, para capturar padrões temporais ainda mais complexos. Essas iniciativas podem ampliar o impacto estratégico da solução, contribuindo para a competitividade da rede em um setor dinâmico.

Palavras-chave: Previsão de vendas. Séries temporais. Inteligência artificial. Engenharia de Software.

REFERÊNCIAS BIBLIOGRÁFICAS

ADADI, A.; BERRADA, M. Peeking inside the black-box: **A survey on explainable artificial intelligence (xai)**. IEEE Access, v. 6, p. 52138–52160, 2018. Disponível em: <https://doi.org/10.1109/ACCESS.2018.2870052>.

BELGIU, M.; DRăGUţ, L. **Random forests in remote sensing**. Remote Sensing of Environment, 2020. BIAMONTE, J.; WITTEK, P.; PANCOTTI, N.; REBENTROST, P.; WIEBE, N.; LLOYD, S.

BOJARSKI, M.; TESTA, D. D.; DWORAKOWSKI, D.; FIRNER, B.; FLEPP, B.; GOYAL, P. et al. **End to End Learning for Self-Driving Cars**. 2016. <https://arxiv.org/abs/1604.07316>. ArXiv preprint arXiv:1604.07316.

BOSTROM, N. **Superintelligence: Paths, Dangers, Strategies**. Oxford, UK: Oxford University Press, 2014.

OBERMEYER, Z.; POWERS, B.; VOGELI, C.; MULLAINATHAN, S. **Dissecting racial bias in an algorithm used to manage the health of populations**. Science, v. 366, n. 6464, p. 447–453, 2019.

RASCHKA, S.; MIRJALILI, V. **Python Machine Learning**. Second edition. Livery Place 35 Livery Street Birmingham B3 2PB, UK: Packt Publishing Ltd., 2022.

ZHANG, M.; FERNáNDEZ-TORRES, M. Ángel; CAMPS-VALLS, G. **Domain knowledge driven variational recurrent networks for drought monitoring**. Remote Sensing of Environment, v. 311, p. 114252, 2024.