

Pipeline Data-Driven para Previsão de Preços Agrícolas: Uma Abordagem de Engenharia de Dados para o Agronegócio Brasileiro

Karla C. Ferreira¹, Alessandra Martins Coelho¹, João Paulo Campolina Lamas¹

¹Departamento Acadêmico de Ciência da Computação
Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais -
Campus Rio Pomba –Bairro Lindo Vale – Rio Pomba – MG – Brazil

kferreira7581@gmail.com, {alessandra.coelho, joao.lamas}@ifsudestemg.edu.br

Abstract. *This research investigates the role of data engineering in predicting the price per arroba of beef cattle in Brazil, considering the integration of multiple sources, such as market data, macroeconomic indicators and climate variables. The approach involves constructing derived attributes and evaluating different machine learning algorithms under walk-forward temporal validation. The results show that performance varies depending on the forecast horizon, with errors of less than 2% in the short term and significant gains compared to traditional baselines. The findings indicate that the quality of the data pipeline has a decisive influence on predictive performance.*

Resumo. *Esta pesquisa investiga o papel da engenharia de dados na previsão do preço da arroba do boi gordo no Brasil, considerando a integração de múltiplas fontes, como dados de mercado, indicadores macroeconômicos e variáveis climáticas. A abordagem envolve a construção de atributos derivados e a avaliação de diferentes algoritmos de aprendizado de máquina sob validação temporal do tipo walk-forward. Os resultados mostram que o desempenho varia conforme o horizonte de previsão, com erros inferiores a 2% a curto prazo e ganhos expressivos em relação a baselines tradicionais. Os achados indicam que a qualidade do pipeline de dados exerce influência decisiva no desempenho preditivo.*

1. Introdução

O agronegócio brasileiro responde por aproximadamente 29,4% do Produto Interno Bruto (PIB) nacional [Ministério da Agricultura e Pecuária 2025], sendo a pecuária de corte um de seus principais segmentos. Nesse contexto, o preço da arroba do boi gordo constitui um indicador estratégico para produtores, frigoríficos e agentes financeiros, o qual é influenciado por fatores econômicos, climáticos e expectativas de mercado.

A dinâmica desse mercado apresenta uma elevada complexidade devido à interação entre variáveis heterogêneas e frequentemente não lineares, como a taxa de câmbio, os preços de contratos futuros, os indicadores macroeconômicos e as variáveis climáticas, dificultando a modelagem preditiva por abordagens convencionais. Além disso, a crescente disponibilidade de dados econômicos, financeiros e climáticos têm ampliado o potencial de utilização de técnicas *data-driven* na previsão de *commodities* agrícolas [Tran et al. 2023].

Nesse cenário, diferentes abordagens têm sido exploradas para previsão do preço da arroba do boi gordo, incluindo modelos de aprendizado de máquina e integração de múltiplas fontes de dados [Reis Filho et al. 2021].

Considerando esses aspectos, este artigo apresenta o impacto da engenharia de dados na previsão do preço da arroba do boi gordo, a partir de um *pipeline data-driven* capaz de integrar múltiplas fontes heterogêneas. Assim, a principal contribuição desta pesquisa está na valorização das etapas de construção e seleção de atributos, contemplando a geração de variáveis temporais e econômicas, e a análise sistemática da relevância de cada grupo de *features*, a fim de melhorar o desempenho preditivo dos modelos. Visando assegurar maior aderência a cenários reais, os experimentos foram conduzidos por meio de validação temporal do tipo *walk-forward*, considerando múltiplos horizontes de previsão.

2. Trabalhos relacionados

A previsão de preços de *commodities* agrícolas tem sido amplamente investigada devido à elevada volatilidade desses mercados e à sua relevância econômica. Abordagens estatísticas clássicas, como os modelos Autorregressivos Integrados de Médias Móveis (do inglês *Autoregressive Integrated Moving Average*, ARIMA), permanecem amplamente utilizados na modelagem de séries temporais por sua capacidade de capturar padrões autorregressivos e tendências [Kontopoulou et al. 2023]. Entretanto, essas abordagens apresentam limitações na representação de relações não lineares e na incorporação simultânea de múltiplas fontes heterogêneas de dados.

Nesse contexto, diferentes técnicas de aprendizado de máquina têm sido exploradas em tarefas de previsão agrícola e financeira, especialmente devido à capacidade de modelar relações complexas entre variáveis estruturadas [Tran et al. 2023]. Entre os métodos frequentemente utilizados destacam-se *Random Forest* [Breiman 2001], *Gradient Boosting* [Friedman 2001], *XGBoost* [Chen and Guestrin 2016], *Support Vector Regression (SVR)* [Drucker et al. 1997] e arquiteturas recorrentes como *LSTM* [Hochreiter and Schmidhuber 1997].

Mais recentemente, pesquisas como a de [Reis Filho et al. 2021] têm destacado a importância da integração de múltiplas fontes de dados, incluindo séries temporais econômicas e dados textuais como fator relevante para a melhoria do desempenho preditivo. Entretanto, apesar dos avanços nos modelos, ainda percebe-se trabalhos, como de [Tran et al. 2023], reforçando que os desafios relacionados à qualidade, integração e preparação dos dados permanecem centrais em aplicações preditivas baseadas em aprendizado de máquina.

Nesse contexto, este artigo diferencia-se ao enfatizar a construção de um *pipeline* de engenharia de dados, integrando fontes heterogêneas e avaliando sistematicamente o impacto das variáveis no desempenho preditivo.

3. Metodologia

Esta seção descreve a abordagem adotada para a construção do *pipeline* de dados, a integração das fontes utilizadas, a geração de atributos derivados e a avaliação experimental dos modelos de previsão.

3.1. Arquitetura do Pipeline de Dados

A abordagem proposta baseia-se em um *pipeline data-driven* que integra múltiplas fontes de dados heterogêneas, desde a fase de extração até a geração de *features*, seguindo o paradigma *ETL* (*Extract, Transform, Load*) [Kimball and Ross 2013] adaptado para previsão de séries temporais agrícolas, conforme ilustrado na Figura 1.

Na etapa de Transformação, a integração via merge e join unifica as bases pelo campo de data. Em seguida, a limpeza padroniza os tipos de dados e trata valores ausentes por interpolação linear, aplicando a normalização *Min-Max* sequencialmente. Por fim, a engenharia de atributos gera variáveis temporais (*lags* e médias móveis) restritas ao histórico de cada janela, mitigando riscos de *data leakage*.



Figura 1. Arquitetura do pipeline de dados proposto.

Fonte: Elaborado pela autora com auxílio de inteligência artificial (ChatGPT/DALL-E), 2026.

3.2. Fontes de Dados e Engenharia de Features

O *dataset* abrangeu o período de janeiro de 2015 a março de 2026, com granularidade diária (≈ 2.700 observações) e 137 *features* derivadas. Ele foi construído a partir da integração de: (1) preços históricos do boi gordo [CEPEA/ESALQ-USP 2024]; (2) contratos futuros de boi gordo [B3 - Bolsa Brasil Balcão 2024]; (3) indicadores macroeconômicos obtidos via *python-bcb* [Freitas 2024] e (4) dados climáticos históricos [Open-Meteo 2024]. Seu pré-processamento incluiu o tratamento de valores ausentes por interpolação linear e a normalização pelo método *Min-Max*.

A engenharia de *features*, por sua vez, foi realizada com o objetivo de capturar diferentes aspectos da dinâmica dos preços, incluindo variáveis temporais (defasagens e médias móveis), variáveis de mercado (contratos futuros) e indicadores macroeconômicos.

3.3. Modelagem, Validação e Avaliação Experimental

A previsão do preço da arroba do boi gordo foi abordada como um problema de regressão em séries temporais, considerando múltiplos horizontes de previsão: D+1, D+7, D+15, D+30, D+90 e D+180. Foram avaliados modelos de diferentes classes, incluindo *Ridge Regression*, métodos baseados em árvores (*Random Forest*, *Gradient Boosting*, *Extra*

Trees, *XGBoost* e *LightGBM*), além de *SVR* e *MLP*. A seleção desses algoritmos buscou contemplar abordagens lineares e não lineares amplamente utilizadas na literatura para previsão de séries temporais e regressão, especialmente em cenários com múltiplas variáveis e relações complexas entre os atributos. A persistência (*naïve forecast*) e o ARIMA foram utilizados como *baselines*.

Os dados foram divididos respeitando a ordem temporal, com 70% para treinamento e 30% para teste, utilizando validação com janelas de 30 dias. A avaliação foi conduzida por validação temporal *walk-forward* com janela expansiva, garantindo a ausência de *data leakage* e maior aderência ao uso real dos modelos [Bergmeir et al. 2016]. Foram utilizadas métricas amplamente adotadas na literatura, incluindo o *MAE* (*Mean Absolute Error*), que representa o erro médio absoluto em reais; o *RMSE* (*Root Mean Squared Error*), correspondente à raiz do erro quadrático médio; o *MAPE* (*Mean Absolute Percentage Error*), que mede o erro percentual médio; e o R^2 (*Coefficient of Determination*), indicador do coeficiente de determinação.

4. Resultados e Discussão

Nesta seção são apresentados os resultados obtidos nos diferentes horizontes de previsão, bem como as análises comparativas entre modelos e grupos de variáveis utilizados no *pipeline* proposto.

4.1. Desempenho por Horizonte

A Tabela 1 apresenta os melhores resultados obtidos para cada horizonte de previsão avaliado. Os experimentos evidenciam diferenças significativas de desempenho entre previsões de curto, médio e longo prazo. Observou-se elevada precisão nos horizontes de curto prazo (D+1 e D+7). Embora ocorra um recuo na assertividade nos horizontes intermediários devido ao aumento da incerteza, nota-se uma estabilização e leve redução nos erros absolutos em D+180 (como o MAE), impulsionada pela robustez do algoritmo *Random Forest* em mitigar o *overfitting* no longo prazo. Ressalta-se que a transição para valores negativos de R^2 nos horizontes mais longos reflete a dificuldade dos modelos em superar o comportamento da média histórica diante da alta volatilidade acumulada.

Tabela 1. Melhores resultados por horizonte de previsão

Horizonte	Melhor Modelo	MAE (R\$)	RMSE (R\$)	MAPE (%)	R^2
D+1	Ridge	3,78	4,96	1,38	0,9864
D+7	Extra Trees	12,10	15,97	4,60	0,8603
D+15	Gradient Boosting	14,02	19,03	5,43	0,8032
D+30	Extra Trees	22,06	28,80	11,84	0,5874
D+90	Extra Trees	40,87	52,23	13,29	-0,6362
D+180	Random Forest	37,05	42,96	12,83	-0,2871

Fonte: elaboração própria.

4.2. Análise dos Modelos e das Variáveis

A análise dos resultados revelou padrões consistentes entre os modelos avaliados. O modelo *Ridge* apresentou desempenho superior no horizonte D+1, sugerindo que a dinâmica

de curtíssimo prazo pode ser adequadamente capturada por relações aproximadamente lineares quando combinadas com *features* temporais.

Com o aumento do horizonte de previsão, observa-se uma mudança no comportamento dos modelos. Nos horizontes intermediários (D+7 a D+90), métodos baseados em árvores, em particular o *Extra Trees*, apresentam melhor desempenho, indicando que relações não lineares e interações entre atributos se tornam progressivamente mais relevantes. No longo prazo (D+180), o *Random Forest* demonstrou maior robustez, possivelmente devido à sua capacidade de reduzir o *overfitting* e lidar com maiores níveis de incerteza.

Em relação às *baselines*, os modelos propostos apresentaram desempenho superior à persistência e ao ARIMA, com reduções de erro variando entre 25% e 68%, evidenciando a capacidade do *pipeline* em capturar informações não exploradas por métodos tradicionais.

Para investigar o impacto de diferentes grupos de variáveis, foi conduzida uma análise de ablação, removendo sistematicamente grupos de *features*, conforme apresentado na Tabela 2.

Tabela 2. Análise de ablação — impacto de diferentes grupos de variáveis

Configuração	D+1 MAE	D+7 MAE	D+30 MAE
Completo (todas as features)	3,78	12,10	22,06
Sem lags	8,45 (+124%)	22,30 (+84%)	38,60 (+87%)
Sem variáveis macroeconômicas	4,12 (+9%)	13,20 (+9%)	23,40 (+14%)
Sem variáveis climáticas	3,82 (+1%)	12,50 (+3%)	21,20 (+3%)
Sem contratos futuros	4,35 (+15%)	14,80 (+22%)	25,70 (+25%)
Apenas lags	5,20 (+38%)	16,40 (+36%)	29,80 (+45%)

Valores entre parênteses indicam aumento percentual no MAE.

Fonte: elaboração própria.

A análise de ablação apresentada na Tabela 2 evidencia a contribuição das diferentes fontes de dados no desempenho dos modelos. A remoção de *features* temporais (*lags*) resulta em degradação significativa da performance, confirmando a forte autocorrelação presente na série. Adicionalmente, a exclusão de contratos futuros implica aumento relevante no erro, indicando que essas variáveis capturam expectativas de mercado não refletidas apenas nos dados históricos. Os indicadores macroeconômicos contribuem de forma moderada para o desempenho preditivo, enquanto as variáveis climáticas apresentam impacto limitado nos horizontes de curto prazo.

4.3. Importância das Variáveis

A análise de importância das variáveis, realizada via *permutation importance* em *Random Forest*, indica que os atributos mais relevantes estão associados à memória temporal da série e a fatores econômicos estruturais. Destacam-se o preço defasado, médias móveis de curto prazo e contratos futuros.

Esses resultados reforçam a predominância da autocorrelação e a relevância de fatores econômicos estruturais na formação do preço, evidenciando a importância da integração de múltiplas fontes no *pipeline* proposto.

4.4. Análise Visual

A Figura 2 apresenta a comparação entre valores reais e previstos no horizonte D+1, permitindo uma avaliação visual da aderência do modelo. A elevada correspondência entre as curvas real e prevista indica bom desempenho na captura da dinâmica de curtíssimo prazo, com $MAE = 3,78$ e $R^2 = 0,9864$. Os desvios residuais concentram-se em picos e vales específicos, sugerindo sensibilidade a choques exógenos de mercado, possivelmente relacionados a eventos como embargos sanitários e variações cambiais abruptas, que afetam temporariamente o padrão histórico da série.

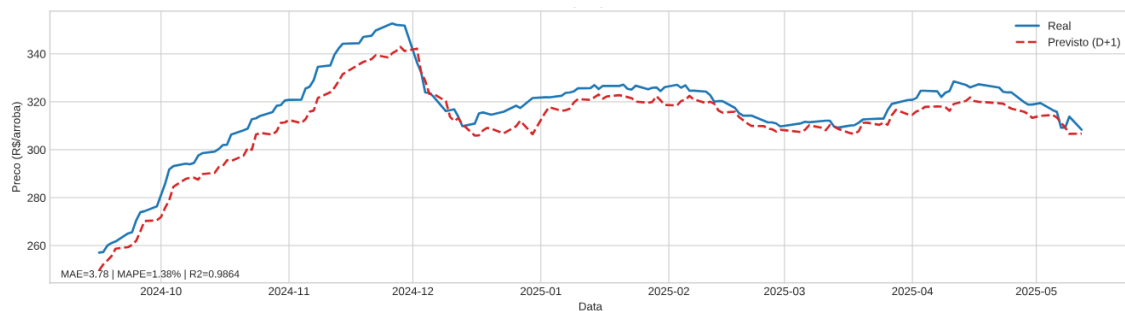


Figura 2. Comparação entre valores reais e previstos no horizonte D+1.

Fonte: elaboração própria.

4.5. Interpretação por Regime Temporal

Os resultados indicam três regimes distintos de previsibilidade: curto prazo com alta precisão ($MAPE$ inferior a 5%), médio prazo com desempenho moderado (entre 5% e 12%), e longo prazo, com menor precisão (superior a 12%). Essa segmentação sugere que a utilidade do modelo está diretamente associada ao horizonte de decisão, sendo mais adequada para suporte tático no curto prazo e para análise exploratória em horizontes mais longos.

5. Conclusões

Os resultados evidenciaram que a engenharia de *features*, especialmente por meio de variáveis temporais e da integração de múltiplas fontes de dados, como contratos futuros e indicadores macroeconômicos, exerceu impacto significativo no desempenho preditivo, com ganhos consistentes em relação às *baselines* tradicionais.

Observou-se, ainda, que o desempenho dos modelos depende do horizonte de previsão, não havendo um algoritmo universalmente superior. Modelos lineares apresentaram melhor desempenho em horizontes curtos, enquanto métodos baseados em árvores se destacaram em horizontes mais longos.

De forma geral, os resultados reforçam que, em cenários complexos e multivariados, a qualidade do *pipeline* de dados pode ser tão relevante quanto a escolha do algoritmo de aprendizado de máquina.

Como trabalhos futuros, pretende-se realizar testes estatísticos, como o teste de Diebold-Mariano, para validar se a diferença de desempenho entre os modelos é significativa. Além disso, planeja-se medir o tempo de execução e a eficiência do pipeline de dados, bem como testar modelos de aprendizado profundo para tentar melhorar os resultados nos horizontes de mais longo prazo.

6. Agradecimentos

Ao Programa de Educação Tutorial (PET) pelo apoio à realização desta pesquisa.

Referências

- B3 - Bolsa Brasil Balcão (2024). Contratos futuros de boi gordo. <http://www.b3.com.br/>.
- Bergmeir, C., Hyndman, R. J., and Benítez, J. M. (2016). Bagging exponential smoothing methods using stl decomposition and box–cox transformation. *International Journal of Forecasting*, 32(2):303–312.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.
- CEPEA/ESALQ-USP (2024). Centro de estudos avançados em economia aplicada. <https://www.cepea.esalq.usp.br/>.
- Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.
- Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., and Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*, 9.
- Freitas, W. (2024). python-bcb: Python wrapper for brazilian central bank web services. <https://github.com/wilsonfreitas/python-bcb>.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.
- Kimball, R. and Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. John Wiley & Sons.
- Kontopoulou, V. I., Panagopoulos, A. D., Kakkos, I., and Matsopoulos, G. K. (2023). A review of arima vs. machine learning approaches for time series forecasting in data driven networks. *Future Internet*, 15(8).
- Ministério da Agricultura e Pecuária (2025). Estatísticas de comércio exterior do agromercado brasileiro. <https://www.gov.br/agricultura/>.
- Open-Meteo (2024). Open-meteo historical weather api. <https://open-meteo.com/>.
- Reis Filho, I. J., Portari Junior, S. C., Oliveira, C. M., Barcelos, L. V., and Nunes Corrêa, G. (2021). A integração de séries temporais e dados de textos para a previsão de preços futuros de milho e soja. *Revista de Sistemas de Informação*, 1(1).
- Tran, N.-Q., Felipe, A., Nguyen Ngoc, T., Huynh, T., Tran, Q., Tang, A., and Nguyen, T. (2023). Predicting agricultural commodities prices with machine learning: A review of current research. *arXiv preprint arXiv:2310.18646*.