Predição de Margem EBITDA com Machine Learning Aplicada ao Fluxo de Caixa Descontado

Diogo Santos Pires Jandiroba[1], **Michel S. Soares**[1]

[1]Universidade Federal de Sergipe – São Cristóvão – SE – Brazil

diogo.jandiroba@dcomp.ufs.br, michel@dcomp.ufs.br

Abstract. *Company valuation using the Discounted Cash Flow (DCF) model suffers from subjectivity in volatile environments. In this context, this paper integrates Machine Learning techniques into the DCF framework to support financial projections. The methodology analyzed the most liquid stocks traded on the Brazilian stock exchange (B3) from 2010 to 2024, excluding the financial sector due to capital structure incompatibility. The Random Forest algorithm achieved the best performance, reaching a Mean Absolute Error (MAE) of 5.61 p.p. in the test set. In the validation set, errors stabilized below 2 p.p. after the disruption period caused by the COVID-19 pandemic. We conclude that the tool mitigates predictive instability.*
Keywords: *Valuation, Discounted Cash Flow, Machine Learning, Random Forest, Capital Markets.*

Resumo. *A avaliação de empresas pelo modelo de Fluxo de Caixa Descontado (FCD) sofre com a subjetividade em ambientes de volatilidade. Neste contexto, este artigo integra técnicas de Aprendizado de Máquina ao FCD para embasar as projeções financeiras. A metodologia analisou as ações de maior liquidez da bolsa brasileira (B3) no período de 2010 a 2024, excluindo o setor financeiro por incompatibilidade estrutural de capital. O algoritmo Random Forest obteve o melhor desempenho, alcançando um Erro Absoluto Médio (MAE) de 5,61 p.p. no conjunto de teste. No conjunto de validação, os erros estabilizaram abaixo de 2 p.p. após o período de instabilidade provocado pela pandemia de COVID-19. Conclui-se que a ferramenta mitiga a instabilidade preditiva.*
Palavras-chave: *Valuation, Fluxo de Caixa Descontado, Aprendizado de Máquina, Random Forest, Mercado de Capitais.*

1. Introdução

A avaliação de uma empresa (*valuation*) consiste no cálculo do seu valor intrínseco, isto é, o valor presente de todos os benefícios econômicos que se espera que ela gere ao longo do tempo. A relevância dessa determinação reside na distinção entre o valor intrínseco e o preço de mercado do ativo, pois esta diferença protege o investidor do risco de pagar excessivamente por um investimento [Penman 2011]. Em essência, o valor de uma empresa reflete a sua capacidade de geração de proventos e fluxos de caixa no presente e no futuro [da Cunha et al. 2014].

Os modelos de *valuation* dividem-se em diversas categorias, destacando-se o Fluxo de Caixa Descontado (FCD) como a técnica conceitualmente mais recomendada na literatura de finanças corporativas. A premissa central desse método é que o valor de

um ativo equivale ao valor presente de todos os seus fluxos de caixa futuros esperados, descontados por uma taxa que reflita o risco inerente ao investimento [Koller et al. 2020].

A principal lacuna na aplicação tradicional do FCD é a subjetividade atrelada às suas premissas, que torna o modelo sensível ao julgamento humano tanto na determinação da taxa de desconto quanto na previsão de crescimento de longo prazo. Em países emergentes como o Brasil, essa lacuna é agravada pelas frequentes instabilidades macroeconômicas, pela volatilidade cambial e pela interferência governamental em setores específicos, fatores que comprometem a capacidade analítica de projeção [da Cunha et al. 2014].

Para superar esse desafio, este trabalho integra técnicas de Aprendizado de Máquina à projeção de um direcionador do modelo FCD: a Margem EBITDA (Lucro antes de Juros, Impostos, Depreciação e Amortização, do inglês *Earnings Before Interest, Taxes, Depreciation and Amortization*, dividido pela Receita Líquida), que alimenta o cálculo do Fluxo de Caixa Livre. O objetivo é reduzir a dependência do julgamento subjetivo do analista nessa projeção, utilizando dados financeiros, macroeconômicos e de mercado de empresas brasileiras. Para isso, a metodologia treina e compara algoritmos de *Ensemble* baseados em árvores de decisão (*Random Forest* e *XGBoost*) sobre as trinta ações de maior liquidez na B3 no período de 2010 a 2024, excluindo entidades do setor financeiro por incompatibilidade estrutural de capital. Esse intervalo temporal foi escolhido a partir de 2010 devido à grande instabilidade geopolítica e econômica, nacional e internacional, provocada pela crise do *sub-prime* em 2008 e 2009. Essa escolha permite avaliar a eficácia dos algoritmos preditivos em isolar anomalias e sustentar a avaliação em cenários de alta volatilidade.

2. Referencial Teórico

2.1. O Modelo de Fluxo de Caixa Descontado e o Desafio da Volatilidade

Na modelagem clássica do FCD, o valor da empresa corresponde ao somatório dos fluxos de caixa livres projetados para um período explícito, acrescido do valor residual na perpetuidade, sendo o Custo Médio Ponderado de Capital (WACC) o balizador da taxa de desconto [Koller et al. 2020].

Apesar de sua solidez teórica, o FCD é criticado por sua sensibilidade a premissas subjetivas: a projeção de variáveis de longo prazo carrega alto grau de incerteza, abrindo margem para vieses cognitivos e especulação por parte do avaliador [Penman 2011]. Em economias emergentes como o Brasil, essa sensibilidade é agravada pela instabilidade macroeconômica e pelo risco-país elevado, tornando as projeções baseadas apenas em médias históricas e julgamentos analíticos [da Cunha et al. 2014]. A integração de grandes volumes de dados ao processo de *valuation*, contudo, mitiga essas distorções ao permitir a extração de cenários preditivos baseados em evidências empíricas [Ajayi et al. 2025].

2.2. Aprendizado de Máquina em Finanças Corporativas

As séries temporais financeiras caracterizam-se por ruídos, alta volatilidade, e dinâmicas não lineares [Tripathi et al. 2024]. Historicamente, as tentativas de previsão dessas séries apoiavam-se em modelos econômicos e estatísticos tradicionais, como o ARIMA (*Auto-regressive Integrated Moving Average*). Contudo, tais métodos dependem fortemente de premissas de linearidade e estacionariedade, falhando ao integrar dezenas de variáveis macroeconômicas e operacionais complexas [Wang and Zain 2025, Tripathi et al. 2024].

Para superar essas limitações, o uso de algoritmos de Aprendizado de Máquina vem se consolidando no campo financeiro [Santos et al. 2025]. Esses algoritmos aprendem iterativamente, mapeando interações multidimensionais ocultas nos dados históricos [Géron 2019]. Pesquisas recentes [Ajayi et al. 2025] [Tripathi et al. 2024] demonstram que modelos preditivos impulsionados por Inteligência Artificial superam abordagens estáticas convencionais, alcançando melhorias de até 25% na acurácia das projeções. Esses algoritmos conseguem modelar o impacto transversal de variáveis como juros, PIB e câmbio diretamente nas margens operacionais das companhias, reduzindo consideravelmente as margens de erro humano [Tripathi et al. 2024].

2.3. Arquiteturas de Aprendizado de Máquina

O desempenho de diferentes arquiteturas de Aprendizado de Máquina no mercado financeiro está ligado à frequência e ao formato estrutural dos dados [Géron 2019]. Modelos de Aprendizado Profundo (*Deep Learning*), especialmente as Redes Neurais Recorrentes do tipo LSTM (*Long Short-Term Memory*), possuem portas de memória capazes de modelar padrões sequenciais de longo prazo [Wang and Zain 2025]. Apesar de seu bom desempenho em negociações de alta frequência (*High-Frequency Trading*), as LSTMs demandam volumes massivos de dados para a correta convergência de sua função de perda [Wang and Zain 2025].

Quando aplicadas a dados fundamentalistas corporativos, como balanços anuais, as LSTMs sofrem com a escassez amostral. Em uma janela histórica de dez a quinze anos, o baixo número de instâncias temporais prejudica o ajuste dos pesos da rede neural, tornando-a suscetível à dissipação ou explosão do gradiente e ao sobreajuste [Géron 2019].

Em contrapartida, para demonstrações contábeis e dados tabulares de baixa frequência, a literatura aponta a superioridade dos métodos de *Ensemble* baseados em árvores de decisão. Algoritmos como o *Random Forest* e o *XGBoost* constroem múltiplas árvores simultâneas, conferindo resiliência ao ruído e capacidade de lidar com amostras restritas [Géron 2019]. Esses modelos extraem padrões seccionais, identificam a importância relativa dos atributos financeiros e isolam valores atípicos, entregando predições sem a necessidade de centenas de milhares de ciclos de treinamento [Tripathi et al. 2024].

3. Metodologia

Para avaliar o impacto do Aprendizado de Máquina nas projeções de fluxo de caixa corporativo, o mapeamento metodológico desta pesquisa fundamenta-se em um modelo preditivo centrado na Margem EBITDA como variável-alvo. Diferentemente do EBITDA em valores absolutos (R$), que escala diretamente com o porte de faturamento de cada companhia e inviabiliza a comparação direta entre empresas de capitalizações distintas, a Margem EBITDA é definida como a razão percentual entre o EBITDA e a Receita Líquida do período:

$$\text{Margem EBITDA}_t = \frac{\text{EBITDA}_t}{\text{Receita Líquida}_t} \tag{1}$$

Essa conversão para uma escala percentual relativa permite o nivelamento estatístico entre empresas de diferentes portes de capitalização, viabilizando o treinamento de um único modelo para todo o conjunto de firmas.

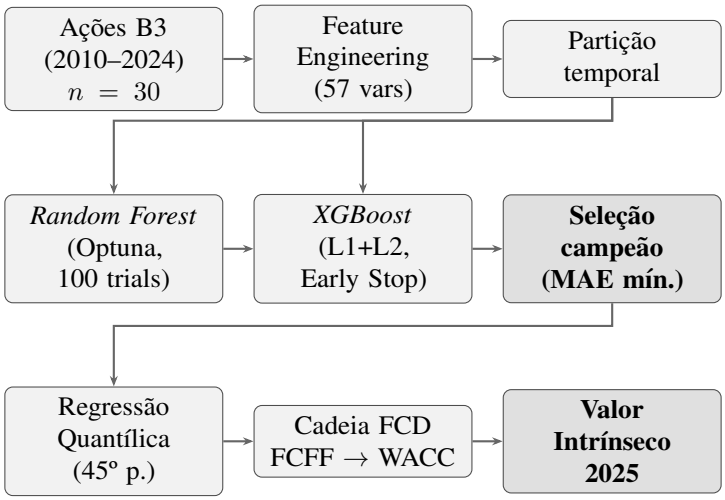


Figura 1. Pipeline metodológico da ferramenta proposta: da coleta de dados à geração do valor intrínseco via FCD.

3.1. Amostra e Particionamento Temporal

A base de dados inicial abrangeu informações financeiras, macroeconômicas e de mercado das trinta empresas de maior liquidez da B3 no período de 2010 a 2024, obtidas a partir do portal de Dados Abertos da B3. Para garantir a comparabilidade estrutural da EBITDA, realizou-se a eliminação de seis ativos do setor financeiro e de seguros (ITUB4, BBDC4, BBAS3, BPAC11, ITSA4 e B3SA3). Essa exclusão é necessária porque a dinâmica de intermediação financeira dessas instituições, em que as receitas e despesas com juros fazem parte da atividade-fim, descaracteriza o uso do EBITDA como um indicador real de eficiência operacional. A empresa Localiza (RENT3) foi removida do conjunto de modelagem para atuar como ativo de *backtesting*, totalizando 23 empresas nos conjuntos de treinamento, validação e teste.

Para evitar o vazamento de informações futuras para o passado, o particionamento temporal obedeceu a uma cronologia: (i) **Treinamento**: 2010–2021; (ii) **Validação**: 2022–2023; (iii) **Teste**: 2024; (iv) **Inferência**: projeções para 2025.

3.2. Engenharia de Variáveis (*Feature Engineering*)

A etapa de processamento originou 57 variáveis preditivas, tratadas com preenchimento de valores ausentes pela mediana calculada exclusivamente sobre o conjunto de treinamento, impedindo assim o contágio dos dados de teste. Entre as *features* criadas, destacam-se: (i) defasagens (*lags*) do período anterior ($t - 1$) para capturar a persistência operacional; (ii) métricas de risco da estrutura de capital, como o Grau de Alavancagem Operacional (GAO) e um score de distância de falência; e (iii) variáveis macroeconômicas defasadas, incluindo a taxa SELIC e a variação cambial.

3.3. Treinamento e Configuração das Arquiteturas

O núcleo preditivo testou abordagens de Aprendizado de Máquina para dados tabulares, com foco em métodos de *Ensemble*. A arquitetura *Random Forest* foi otimizada via inferência Bayesiana com o *framework* Optuna, em 100 iterações para ajuste fino de profundidade máxima, folhas mínimas e fração de atributos. O algoritmo *XGBoost* foi

implementado sob regime de regularização estrita (L1 e L2) combinado a *Early Stopping*. Para mensurar a eficácia real de ambos, seu desempenho foi comparado ao dos modelos de referência *Naive* e Média Histórica.

3.4. Integração com o Fluxo de Caixa Descontado

Após a predição da Margem EBITDA, adotou-se a abordagem do 45º percentil das árvores de decisão do *Random Forest* para estabelecer uma margem de segurança analítica. Ao operar abaixo da mediana da distribuição de previsões, o percentil injeta conservadorismo sistemático nas estimativas sem necessidade de parametrização adicional por parte do analista. A cadeia de cálculo projetou o Fluxo de Caixa Livre para a Firma (FCFF) estimando os impostos pela alíquota efetiva de 34% e deduzindo o CAPEX, sendo o montante descontado pelo WACC.

A cadeia de cálculo segue as equações a seguir. O Fluxo de Caixa Livre para a Firma é obtido por:

$$\text{FCFF}_t = \text{EBITDA}_t \times (1 - t_e) - \text{CAPEX}_t \tag{2}$$

onde $t_e = 0{,}34$ é a alíquota efetiva de IR/CSLL (Lei 9.430/1996).

O custo do capital próprio é estimado pelo CAPM adaptado a mercados emergentes [Koller et al. 2020]:

$$k_e = r_f + \beta_i \times \text{ERP} + \text{CRP} \tag{3}$$

$$\text{WACC} = k_e \cdot \frac{E}{D+E} \; + \; k_d \cdot (1 - t_e) \cdot \frac{D}{D+E} \tag{4}$$

onde r_f é a SELIC projetada para 2025 de 15%, β_i o beta anualizado do ativo, ERP = 5,5% o prêmio de risco histórico e CRP o CDS Brasil de 5 anos com o valor de 2,1% em 2025, como proxy do risco-país. O valor terminal na perpetuidade com crescimento nominal $g = 6{,}5\%$ a.a. , IPCA esperado de 4,26% em 2025 acrescido de crescimento real do PIB projetado em 2,3% para 2025:

$$TV = \frac{\text{FCFF}_{n+1}}{\text{WACC} - g} \tag{5}$$

O valor intrínseco de cada firma corresponde ao somatório dos FCFFs descontados no horizonte explícito de cinco anos acrescido do valor presente de TV.

4. Resultados e Discussão

A avaliação inicial concentrou-se na comparação direta entre os algoritmos testados.

4.1. Seleção de Modelos e Desempenho no Teste (2024)

O modelo *Random Forest* otimizado via Optuna superou o *XGBoost*, atingindo um MAE de 5,61 p.p. no conjunto de teste de 2024. Frente aos modelos de referência convencionais, o *Random Forest* evidenciou uma redução de aproximadamente 55% no MAE em relação à Média Histórica (12,39 p.p.), conforme ilustrado na Figura 2.

A Tabela 1 consolida as métricas de desempenho no conjunto de teste (2024) e no conjunto de validação (2022–2023). O *Random Forest* otimizado via Optuna é o modelo

campeão em MAE no teste, com 5,61 p.p., superando o *XGBoost* (6,15 p.p.) e reduzindo em aproximadamente 55% o erro da Média Histórica (12,39 p.p.). O modelo Naive (*Random Walk*) apresenta MAE de 5,65 p.p. praticamente idêntico ao RF. Essa paridade é esperada e documentada na literatura de previsão de séries financeiras [Wang and Zain 2025]: a inércia operacional da Margem EBITDA faz do valor corrente (y_t) um preditor naturalmente forte para o período seguinte.

O teste de Diebold-Mariano retorna $p = 0{,}48$, indicando que a diferença de MAE entre RF e Naive não é estatisticamente significativa dado o tamanho amostral disponível de $n = 23$ empresa-ano no conjunto de teste. Essa limitação é inerente a estudos com dados fundamentalistas anuais. A adoção do RF em detrimento do modelo de inércia justifica-se, portanto, por duas capacidades ausentes no Naive: (i) a regressão quantílica sobre a distribuição de árvores, que permite derivar estimativas conservadoras para o FCD sem parametrização adicional pelo analista; e (ii) a extração de importância de variáveis, que fornece fundamentação econômica interpretável para cada projeção, alinhando o modelo aos pilares da teoria corporativa.

Tabela 1. Desempenho dos modelos — Margem EBITDA (p.p.).

Modelo	MAE_{val}	MAE_{test}	$\text{RMSE}_{\text{test}}$	IC 95% $_{\text{test}}$
Média Histórica	–	12,39	–	–
Naive (RW)	–	5,65	–	[3,20; 8,44]
XGBoost	7,40	6,15	7,32	–
RF (Optuna)	**6,18**	**5,61**	**7,45**	**[3,77; 7,78]**

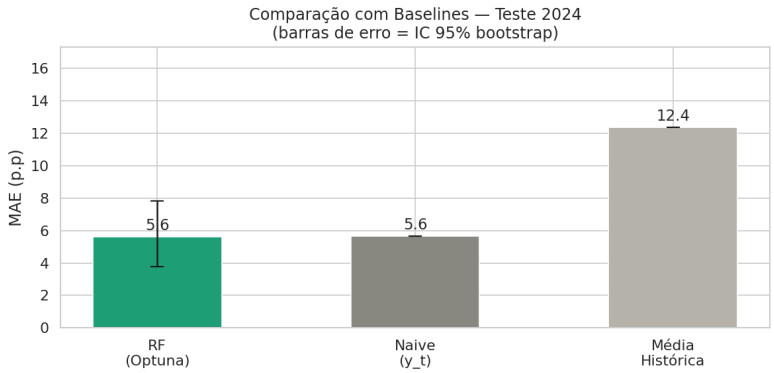


Figura 2. Comparação de desempenho (MAE) do Random Forest contra os modelos de referência (*baselines*) no ano de teste (2024).

4.2. Análise de Importância de Atributos e Comportamento dos Resíduos

A análise de importância de atributos (*Feature Importance*) confirmou que a variável de maior peso preditivo foi a própria margem EBITDA defasada da empresa, evidenciando a autocorrelação temporal da eficiência operacional. Indicadores de endividamento ocuparam as posições seguintes em relevância, conforme exibido na Figura 4.

A distribuição e a dispersão dos resíduos, apresentadas na Figura 5, confirmam a ausência de viés sistemático profundo: os erros concentram-se de forma simétrica em

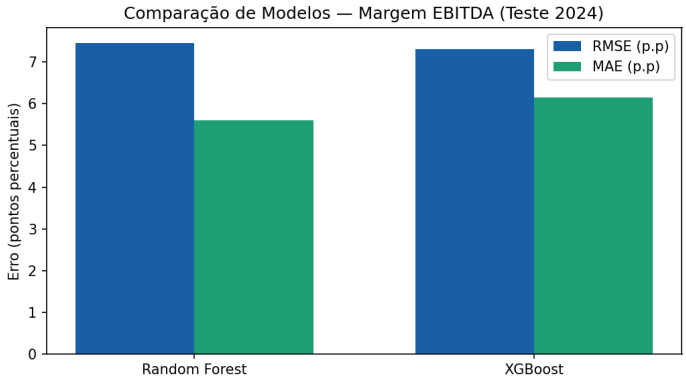


Figura 3. Comparação de RMSE e MAE entre *Random Forest* (Optuna) e *XGBoost* no conjunto de teste (2024). O RF vence em MAE (5,61 vs. 6,15 p.p.); o XGBoost apresenta RMSE marginalmente inferior (7,32 vs. 7,45 p.p.).

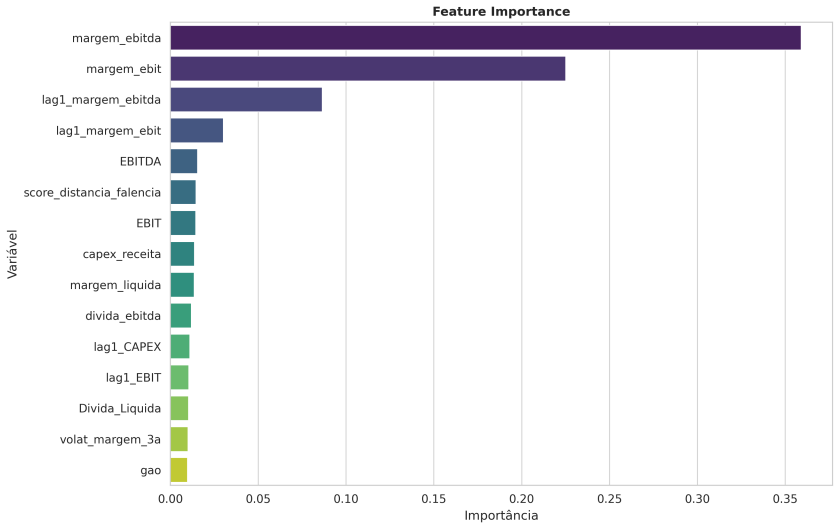


Figura 4. Importância relativa das variáveis (Top 15) no modelo preditivo campeão (*Random Forest* otimizado).

torno de zero, indicando que o modelo não superestima nem subestima sistematicamente as margens projetadas.

4.3. Validação: Estudo de Caso RENT3

A capacidade de generalização e a limitação do modelo frente a rupturas sistêmicas foram avaliadas utilizando a empresa Localiza (RENT3), mantida fora do conjunto de treinamento. Como demonstra a Figura 6, os anos de disrupção provocada pela pandemia de COVID-19 (2020–2022) geraram um deslocamento na distribuição das variáveis (*covariate shift*), elevando temporariamente o erro de projeção. Superada essa anomalia exógena, o modelo convergiu para um nível de excelência, atingindo erros inferiores a 2 p.p. em 2023 e 2024.

A Tabela 2 detalha os erros anuais. Os anos de 2020, 2021 e 2022 refletem o *covariate shift* induzido pela pandemia. Isolados esses três anos, o MAE médio para 2023–2024 é de 1,05 p.p., desempenho dentro da faixa de precisão reportada por analistas *sell-side* para previsões de margem no horizonte de um ano [Ajayi et al. 2025].

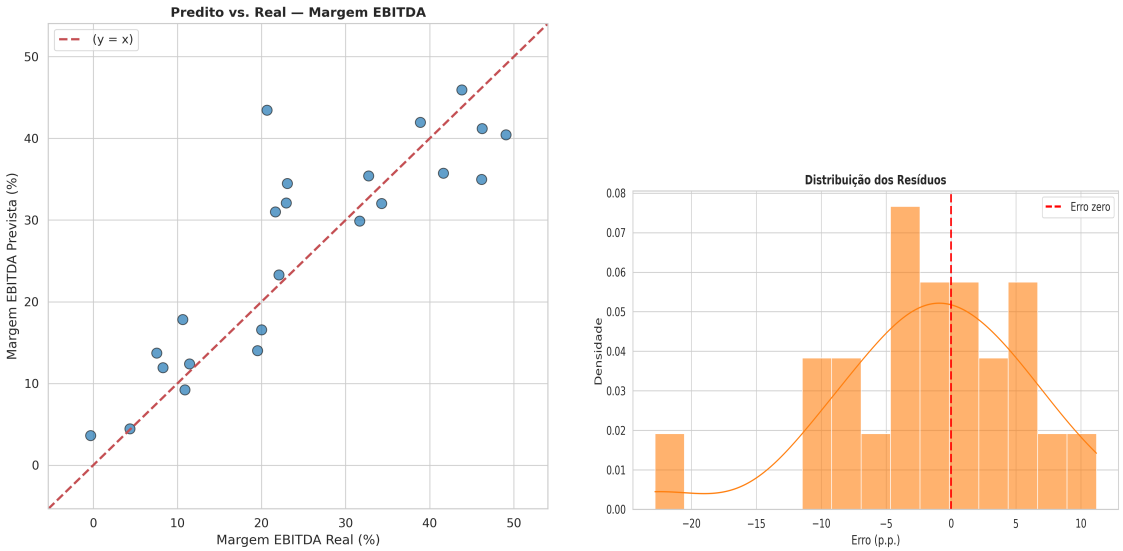


Figura 5. À esquerda, dispersão de aderência entre valores reais e previstos. À direita, distribuição dos resíduos do modelo.

Tabela 2. Erros anuais do *backtesting out-of-sample* — RENT3 (modelo treinado sem nenhum dado da empresa).

Ano	Real (%)	Previsto (%)	Erro Abs. (p.p.)
2018	21,70	21,80	0,10
2019	23,94	21,33	2,61
2020	29,58	20,89	8,69
2021	37,07	27,30	9,78
2022	36,41	28,17	8,24
2023	31,97	32,75	0,78
2024	32,92	31,61	1,31
MAE global (2018–2024)			**4,50**
MAE pós-COVID (2023–2024)			1,05

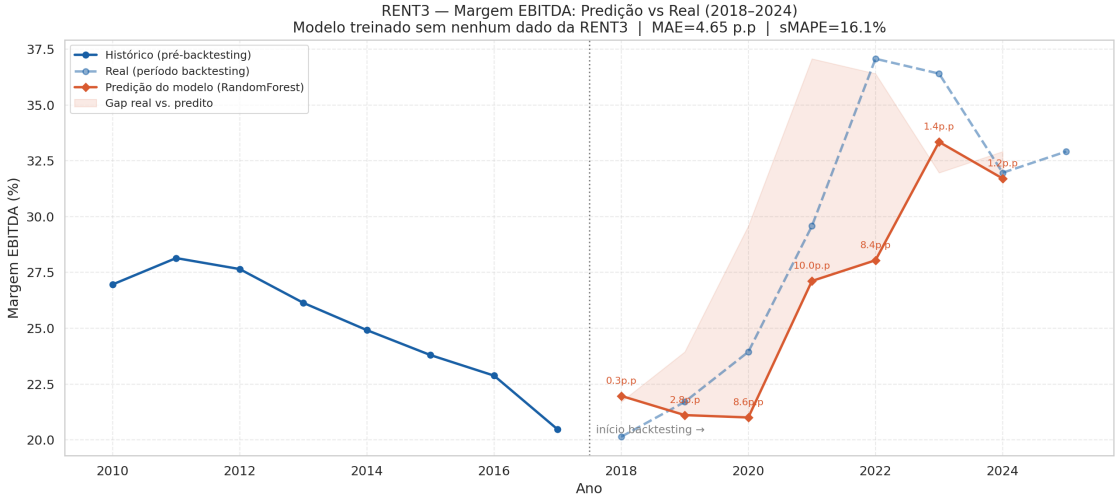


Figura 6. Desempenho preditivo *out-of-sample* e estabilização dos erros após o choque sistêmico da pandemia (RENT3).

4.4. Projeção de Caixa e Margem de Segurança (2025)

Com base na regressão quantílica no 45º percentil, que injeta uma margem de segurança conservadora nas estimativas, o modelo atuou de forma autônoma na inferência da geração de EBITDA das companhias para o exercício de 2025. As predições respeitaram a inércia microeconômica das firmas operacionais e detectaram precocemente regimes de normalização em empresas em processo de desestatização, provendo a âncora quantitativa necessária para estabilizar os fluxos contínuos no FCD, conforme ilustrado na Figura 7.

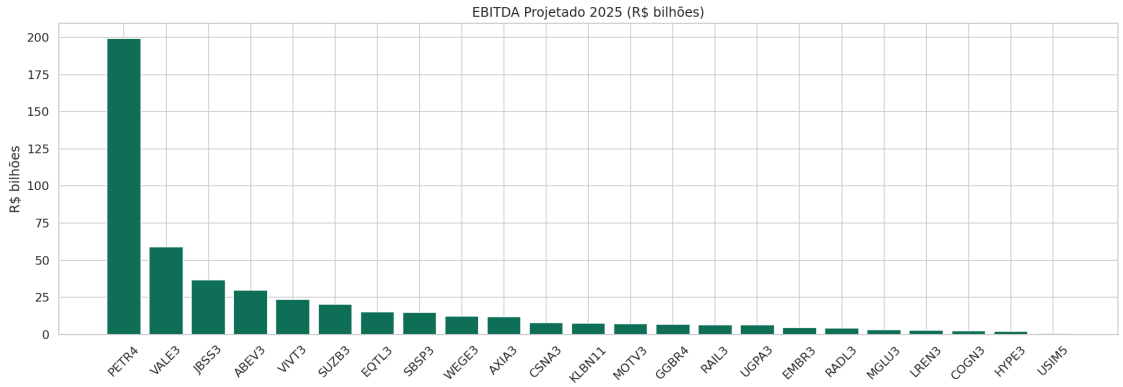


Figura 7. Projeção do EBITDA para o exercício financeiro de 2025 utilizada como insumo do FCD.

5. Conclusão e Trabalhos Futuros

Este trabalho demonstrou a viabilidade da integração de técnicas de Aprendizado de Máquina ao modelo de Fluxo de Caixa Descontado (FCD). Os resultados evidenciaram que algoritmos de *Ensemble*, destacando-se o *Random Forest* otimizado via Optuna, são adequados para dados financeiros tabulares de baixa frequência, reduzindo o erro médio absoluto (MAE) em aproximadamente 55% em relação à Média Histórica. Na validação com a RENT3, o modelo convergiu para erros inferiores a 2 p.p. após o período de disrupção da pandemia, desempenho competitivo com as estimativas do mercado financeiro.

A contribuição central da abordagem reside em três dimensões complementares: (i) a redução expressiva do erro preditivo em relação à extrapolação de médias históricas, que é o procedimento padrão em avaliações manuais; (ii) a capacidade de regressão quantílica, que permite derivar margens de segurança estruturadas e auditáveis para o FCD sem parametrização adicional por parte do analista; e (iii) a extração de importância de variáveis, que alinha as projeções à teoria das finanças corporativas e fornece justificativa econômica interpretável para cada estimativa.

A adoção de métodos baseados em árvores em detrimento de arquiteturas de Aprendizado Profundo justificou-se pelo volume amostral restrito da frequência anual, que inviabiliza a convergência de redes LSTM e aumenta o risco de *overfitting*. O estudo apresenta limitações importantes: a diferença de MAE entre o RF e o modelo Naive não é estatisticamente significativa ($p = 0{,}48$, $n = 23$), o que impede afirmar superioridade preditiva pontual com o tamanho amostral disponível; a análise foi restrita a ativos de alta liquidez da B3; e o setor financeiro foi excluído por incompatibilidade contábil. Essas

limitações delimitam o escopo de generalização dos resultados e motivam as extensões propostas a seguir.

Como trabalhos futuros, sugere-se a migração para bases de dados de frequência trimestral (ITR), o que ampliará o volume amostral e permitirá a aplicação de modelos de *Deep Learning* para capturar dependências temporais complexas. Recomenda-se também a expansão da amostra para uma maior quantidade de empresas, bem como a implementação de janelas de treinamento expansíveis para melhorar a adaptação dos algoritmos a choques econômicos. Por fim, a integração de dados multimodais via Processamento de Linguagem Natural (PLN) apresenta-se como uma fronteira promissora para extrair o sentimento de mercado a partir de relatórios e notícias financeiras.

Referências

Ajayi, A., Obi, I. G., Boboye, C. O., Egyam, E., and Dike, C. M. (2025). Advancing Valuation Accuracy in Mergers and Acquisitions through Artificial Intelligence and Financial Data Analytics in Investment Banking. *GSC Advanced Research and Reviews*, 24(3):186–200.

da Cunha, M. F., Martins, E., and Neto, A. A. (2014). Avaliação de Empresas no Brasil pelo Fluxo de Caixa Descontado: Evidências Empíricas sob o Ponto de Vista dos Direcionadores de Valor nas Ofertas Públicas de Aquisição de Ações. *Revista de Administração da Universidade de São Paulo (RAUSP)*, 49(2):251–266.

Géron, A. (2019). *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow*. Alta Books.

Koller, T., Goedhart, M., and Wessels, D. (2020). *Valuation: Measuring and Managing the Value of Companies*. John Wiley & Sons, 7th edition.

Penman, S. H. (2011). *Accounting for Value*. Columbia Business School Publishing.

Santos, J., Siqueira, L., Faria, E., Bispo, K., Silva, G., and Soares, M. S. (2025). Methodological Advances in Expert Systems with Artificial Intelligence. In *Anais do XXI Simpósio Brasileiro de Sistemas de Informação*, pages 703–712. SBC.

Tripathi, A., Tripathi, A., Trenz, O., and Mishra, P. K. (2024). Comparative Analysis of Selected Time Series Forecasting Approaches for Indian Markets. In *26th International Scientific Conference Economic Competitiveness and Sustainability*.

Wang, B. and Zain, A. B. M. (2025). A Hybrid XGBoost-LSTM Framework for Supply Chain Demand Forecasting: Empirical Evidence from Retail Multi-Store Data. *Journal of Cultural Analysis and Social Change*, 10(4):4056–4073.