Estimating Forecast Difficulty in N-BEATS using Latent Representations and Meta-Learning

Lauryane Santos Siqueira[1], **Leonardo Nogueira Matos**[1],
Paulo Salgado Gomes de Mattos Neto[2], **Filipe Coelho de Lima Duarte**[3]

[1] Departamento de Computação – Universidade Federal de Sergipe (UFS)
São Cristóvão – SE – Brasil

[2]Centro de Informática – Universidade Federal de Pernambuco (UFPE)
Recife – PE – Brasil

[3]Departamento de Finanças e Contabilidade – Centro de Ciências Sociais Aplicadas
Universidade Federal da Paraíba (UFPB)
João Pessoa – PB – Brasil

{lauryanesiqueira, lnmatos}@academico.ufs.br, psgmn@cin.ufpe.br, fcld@cin.ufpe.br

Abstract. *Forecast difficulty estimation is an important challenge for improving the reliability of time series forecasting systems. This work investigates whether latent representations extracted from the interpretable N-BEATS model, time series characteristics, and recent error history can be combined to estimate the difficulty of individual forecasts. We formulate this task as a supervised meta-learning problem in which each forecast is represented as an instance in a meta-dataset. Experiments on the M4 benchmark show that supervised meta-models consistently outperform latent-space and signal-based baselines, achieving Spearman correlations up to 0.895, ROC-AUC values up to 0.960, and F1-scores up to 0.780. Ablation analyses indicate that recent error history is the most informative source of information, while latent representations provide complementary signals. Furthermore, selective prediction experiments demonstrate that the proposed scores effectively prioritize reliable forecasts, reducing prediction risk under low-coverage scenarios.*

1. Introdução

A previsão de séries temporais é fundamental em domínios como economia, energia, saúde e planejamento urbano, auxiliando a tomada de decisão em diferentes contextos [Hyndman and Athanasopoulos 2021]. Apesar dos avanços recentes em aprendizado profundo, avaliar a confiabilidade das previsões permanece um desafio relevante, especialmente em aplicações sensíveis à qualidade das estimativas produzidas [Yunita et al. 2025]. Nesses cenários, antecipar quando uma previsão tende a ser menos confiável pode contribuir para decisões mais seguras, além de apoiar mecanismos de monitoramento e controle de risco.

Grande parte dos modelos neurais utilizados na previsão de séries temporais opera como uma caixa-preta, dificultando a compreensão dos fatores que influenciam suas previsões. Como consequência, técnicas de interpretabilidade *post hoc*, como SHAP e LIME, são frequentemente utilizadas para auxiliar a análise dos resultados produzidos pelos modelos [Freeborough and van Zyl 2022]. Em contrapartida, o modelo N-BEATS (*Neural*

Basis Expansion Analysis for Time Series Forecasting) destaca-se por combinar elevado desempenho preditivo com uma estrutura intrinsecamente interpretável baseada em expansões em funções de base [Oreshkin et al. 2020]. Essa arquitetura produz coeficientes internos associados à construção da previsão, permitindo analisar componentes relacionados à tendência e sazonalidade do sinal.

Apesar da interpretabilidade estrutural do N-BEATS, compreender quando uma previsão deve ser considerada confiável permanece uma questão em aberto. Embora diferentes abordagens de quantificação de incerteza, como métodos Bayesianos, *ensembles* e estratégias baseadas em múltiplas inferências, tenham sido propostas [Gawlikowski et al. 2023], muitas delas exigem modificações arquiteturais ou maior custo computacional. Nesse contexto, representações latentes aprendidas pelo próprio modelo surgem como uma alternativa promissora para estimar a dificuldade das previsões. Essas representações organizam informações relevantes em espaços compactos [Rongali et al. 2020] e podem capturar características estruturais do comportamento temporal que complementam a observação direta da série [Jing et al. 2022].

Diante desse cenário, este trabalho investiga a estimativa da dificuldade de previsão por meio da integração de representações latentes do N-BEATS, características da série temporal e histórico recente de erro em uma abordagem de meta-aprendizado supervisionado. Diferentemente das abordagens tradicionais de quantificação de incerteza, a metodologia proposta reutiliza informações produzidas pelo próprio modelo de previsão, sem exigir modificações em sua arquitetura. Como contribuições, propõe-se uma meta-base de previsões individuais para estimativa de dificuldade, compara-se o desempenho de meta-modelos supervisionados e métodos baseados em espaço latente e avalia-se a utilidade prática dos *scores* produzidos em cenários de previsão seletiva.

2. Trabalhos Relacionados

A estimativa da dificuldade de previsão relaciona-se a diferentes linhas de pesquisa em séries temporais e aprendizado de máquina, incluindo representações latentes, interpretabilidade, *forecastability*, quantificação de incerteza e predição seletiva. Essas abordagens fornecem perspectivas complementares para compreender e avaliar a confiabilidade das previsões produzidas pelos modelos.

2.1. Representações Latentes e Interpretabilidade em Previsão de Séries Temporais

Modelos de aprendizado profundo têm se consolidado como abordagens eficazes para previsão temporal devido à sua capacidade de aprender representações complexas automaticamente [Casolaro et al. 2023, Song et al. 2024]. Esse avanço motivou o desenvolvimento de arquiteturas interpretáveis [Guo et al. 2018] e técnicas de Inteligência Artificial Explicável (XAI) [Rojat et al. 2021]. Essas representações internas podem conter informações úteis para estimar a dificuldade das previsões.

2.2. Forecastability e Dificuldade de Previsão

A previsibilidade de séries temporais (*forecastability*) busca quantificar em que medida o comportamento futuro de uma série pode ser antecipado a partir de observações passadas. Nesse contexto, a relação entre propriedades observáveis das séries e seu desempenho preditivo. [Papacharalampous et al. 2022] demonstraram que características extraídas diretamente das séries podem antecipar sua *forecastability*, enquanto [Ponce-Flores et al. 2020]

observaram associações entre medidas de complexidade baseadas em entropia e o erro obtido por diferentes métodos de previsão.

Nesse contexto, o erro observado após a realização da previsão pode ser utilizado como uma *proxy* operacional da dificuldade preditiva, uma vez que erros elevados refletem situações em que o modelo apresentou maior dificuldade para representar a dinâmica da série temporal.

2.3. Quantificação de Incerteza e Previsão Seletiva

A quantificação de incerteza complementa previsões pontuais com estimativas de confiabilidade. Revisões recentes organizam essas abordagens em métodos Bayesianos, *ensembles*, estratégias de *test-time augmentation* e métodos determinísticos [Gawlikowski et al. 2023]. Além disso, abordagens como *Evidential Deep Learning* buscam modelar explicitamente diferentes fontes de incerteza sem a necessidade de múltiplas inferências ou *ensembles* [Ferchichi et al. 2025].

Uma linha complementar é a previsão seletiva (*selective prediction*), na qual *scores* de confiança são utilizados para selecionar previsões consideradas mais confiáveis [Geifman and El-Yaniv 2017]. Embora amplamente explorada em classificação, essa estratégia ainda é relativamente pouco investigada em tarefas de previsão temporal.

2.4. Representações Latentes e Estimativa de Dificuldade

Pesquisas recentes investigam representações latentes como fonte de informação para estimativa de confiança. O método DUQ (*Deterministic Uncertainty Quantification*), por exemplo, utiliza medidas de proximidade no espaço latente para identificar exemplos potencialmente fora da distribuição de treinamento [Van Amersfoort et al. 2020], enquanto propriedades geométricas dessas representações têm sido associadas à densidade dos dados e à cobertura do espaço de entrada [Postels et al. 2020].

Estudos em aprendizado de representações mostram que espaços latentes organizam fatores complexos em estruturas compactas e semanticamente significativas [Bojanowski et al. 2017, Arvanitidis et al. 2017], indicando seu potencial para estimar a dificuldade das previsões.

Em conjunto, esses estudos demonstram avanços importantes na utilização de representações latentes, na caracterização da *forecastability*, na quantificação de incerteza e na predição seletiva. No entanto, poucos trabalhos investigam a estimativa da dificuldade de previsões em séries temporais utilizando representações latentes interpretáveis extraídas do próprio modelo de previsão e combinadas com características da série e histórico recente de erro. Essa lacuna motiva a abordagem proposta neste trabalho.

3. Metodologia

Esta seção descreve a abordagem proposta para estimativa de dificuldade de previsão. São apresentados a construção da meta-base, a definição da variável de dificuldade, os conjuntos de características utilizados, os métodos avaliados e as estratégias de avaliação adotadas.

3.1. Visão Geral da Abordagem

A Figura 1 apresenta uma visão geral da metodologia proposta. Inicialmente, séries temporais são processadas pelo modelo N-BEATS interpretável. Para cada previsão realizada, são extraídas informações provenientes de três fontes complementares: (i) representações latentes associadas aos coeficientes interpretáveis do modelo, (ii) características estatísticas calculadas diretamente sobre a série temporal e (iii) informações derivadas do histórico recente de erro. Essas informações são então integradas na construção de uma meta-base, na qual cada previsão é representada por um conjunto de atributos utilizado posteriormente pelos métodos de estimativa de dificuldade.

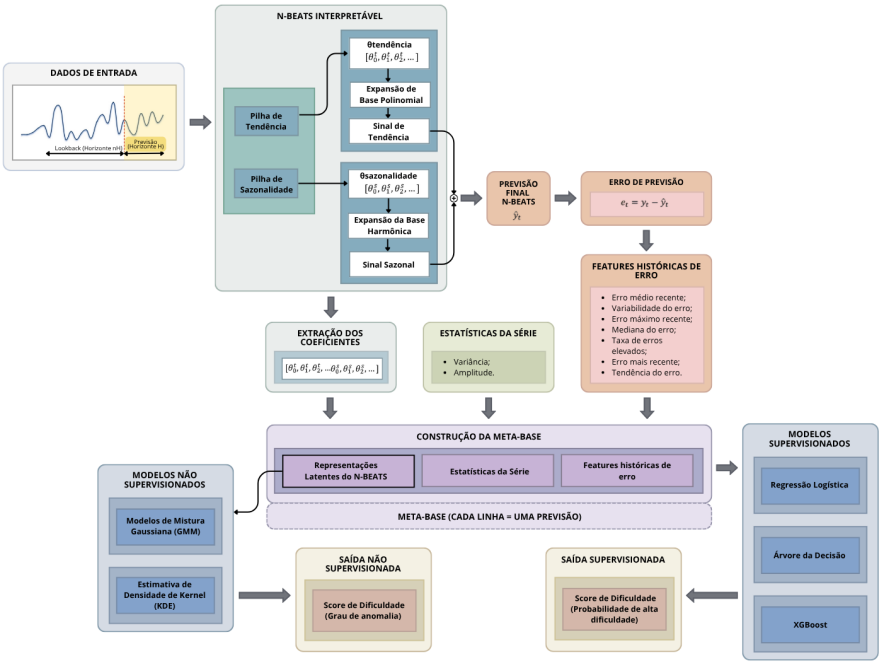


Figura 1. Pipeline da abordagem proposta. Representações latentes do N-BEATS, estatísticas da série temporal e histórico recente de erro são integrados em uma meta-base utilizada para estimar a dificuldade das previsões por métodos supervisionados e não supervisionados.

3.2. Representações Latentes Produzidas pelo N-BEATS

O modelo N-BEATS (*Neural Basis Expansion Analysis for Time Series Forecasting*) organiza a previsão como uma composição de blocos interpretáveis associados a componentes de tendência e sazonalidade [Oreshkin et al. 2020]. Em cada bloco, a previsão é construída a partir de coeficientes latentes que parametrizam funções de base.

Formalmente, seja $\theta_b^{(f)} \in R^{d_f}$ o vetor de coeficientes associado ao componente de *forecast* do bloco b. A contribuição desse bloco para a previsão pode ser expressa por

$$f_b = G_f \theta_b^{(f)}, \tag{1}$$

onde G_f representa a matriz de funções de base da arquitetura. Neste trabalho, os coeficientes de *forecast* dos três blocos de tendência e dos três blocos de sazonalidade são utilizados como representações latentes do estado interno do modelo.

Para obter descritores comparáveis entre previsões, calcula-se a norma Euclidiana dos coeficientes produzidos por cada bloco,

$$s_b = \|\theta_b^{(f)}\|_2, \tag{2}$$

resultando em seis atributos latentes correspondentes aos blocos de tendência e sazonalidade.

Adicionalmente, define-se o vetor

$$\mathbf{s} = [s_1, s_2, \ldots, s_6], \tag{3}$$

a partir do qual é calculada uma medida global da representação latente,

$$S = \|\mathbf{s}\|_2, \tag{4}$$

denominada *radial norm*, utilizada como atributo adicional na construção da meta-base.

3.3. Construção da Meta-Base e Definição da Dificuldade

A estimativa de dificuldade é formulada como um problema de meta-aprendizado supervisionado. Em vez de analisar diretamente as séries temporais, cada previsão produzida pelo N-BEATS é convertida em uma instância de uma meta-base,

$$M = (x_i, d_i)_{i=1}^{N}, \tag{5}$$

onde x_i representa o vetor de características associado à previsão i e d_i sua dificuldade observada.

A dificuldade é definida a partir do erro observado após a realização da previsão. Sejam y_i o valor real e $\hat{y}_i$ a previsão produzida pelo modelo. O erro associado à previsão é calculado por

$$e_i = |y_i - \hat{y}_i|. \tag{6}$$

Neste trabalho, o erro observado é utilizado como uma *proxy* operacional da dificuldade preditiva, assumindo que previsões associadas a maiores erros correspondem a situações mais difíceis para o modelo.

Além da análise contínua baseada em e_i, também são considerados cenários binários nos quais previsões pertencentes aos percentis superiores da distribuição de erros são classificadas como difíceis. Foi utilizado o limiar P90, calculado separadamente para cada frequência temporal para evitar efeitos de escala entre categorias do *benchmark*.

3.4. Conjunto de Características

As características utilizadas na meta-base foram organizadas em três grupos: (i) representações latentes extraídas dos coeficientes interpretáveis de tendência e sazonalidade do N-BEATS, resumidas pelas normas dos três blocos de cada componente e por uma norma radial global; (ii) características calculadas diretamente sobre a janela de entrada, incluindo amplitude e variância da série temporal, utilizadas como *baselines* de *forecastability*; e (iii) estatísticas derivadas das cinco previsões anteriores da mesma série, incluindo medidas de tendência central, dispersão e comportamento recente do erro. Essa organização permite avaliar a contribuição individual e combinada de diferentes fontes de informação para a estimativa da dificuldade preditiva.

3.5. Métodos para Estimativa de Dificuldade

Foram avaliadas duas categorias de métodos. A primeira inclui abordagens não supervisionadas baseadas em representações latentes e características da série temporal: Radial Norm, *Gaussian Mixture Models* (GMM) [Reynolds et al. 2009], *Kernel Density Estimation* (KDE) [Rosenblatt 1956], amplitude e variância. Esses métodos produzem scores de dificuldade diretamente a partir das representações latentes ou de propriedades observáveis da série temporal. A segunda categoria é composta por meta-modelos supervisionados treinados sobre a meta-base construída, incluindo Regressão Logística [Cox 1958], Árvore de Decisão [Quinlan 1996] e XGBoost [Chen et al. 2015]. Esses modelos recebem como entrada as características extraídas da meta-base e produzem scores contínuos de dificuldade por meio da integração das diferentes fontes de informação disponíveis.

4. Configuração Experimental

Esta seção descreve os dados utilizados, a construção da meta-base, os métodos avaliados e o protocolo experimental adotado. Além disso, são apresentadas as métricas empregadas para avaliar a capacidade dos modelos em estimar a dificuldade de previsão, identificar previsões difíceis e apoiar estratégias de previsão seletiva.

4.1. Dados e Meta-Base

Os experimentos utilizaram séries temporais do *benchmark* M4 [Makridakis et al. 2020], considerando as frequências *Monthly, Weekly e Yearly*, totalizando 64.670 séries temporais (41.393 *Monthly*, 359 *Weekly* e 22.918 *Yearly*). As previsões foram geradas pela versão interpretável do N-BEATS (N-BEATS-I) [Oreshkin et al. 2020], avaliando combinações de *lookback* $\in \ 2, 4, 6$ e funções de perda MAPE e MASE [Hyndman and Koehler 2006], com $repeat = 0$. A partir dessas previsões foi construída uma meta-base contendo inicialmente 751.885 previsões individuais. Após a remoção de instâncias sem histórico válido de erro, os experimentos foram conduzidos sobre 438.171 observações.

4.2. Protocolo Experimental

Foram avaliados três meta-modelos supervisionados (Regressão Logística, Árvore de Decisão e XGBoost) e cinco métodos de referência (Radial Norm, GMM, KDE, amplitude e variância). Os experimentos utilizaram particionamento por séries (*GroupShuffleSplit*),

com aproximadamente 70% das séries destinadas ao treinamento e 30% ao teste. Os meta-modelos foram treinados para prever a dificuldade definida pelo percentil P90. Para os métodos baseados em densidade, as representações latentes foram transformadas por Yeo-Johnson [Weisberg 2001] e posteriormente padronizadas. O *score Radial Norm* foi calculado a partir da magnitude das representações latentes, enquanto o número de componentes do GMM foi selecionado via Critério de Informação Bayesiano (BIC) e a largura de banda do KDE por validação cruzada. Todos os métodos foram ajustados separadamente para cada frequência temporal.

4.3. Métricas de Avaliação

A avaliação considerou três perspectivas complementares. A capacidade de ordenar previsões de acordo com sua dificuldade foi medida por meio da correlação de Spearman [Spearman 1961] entre os *scores* produzidos e o erro observado. Embora os meta-modelos sejam treinados utilizando a definição binária baseada no percentil P90, os *scores* produzidos são contínuos, tornando a correlação de Spearman adequada para avaliar sua capacidade de ordenar previsões segundo o risco estimado. A identificação de previsões difíceis foi avaliada utilizando a Área sob a Curva ROC (ROC-AUC) [Bradley 1997] e o *F1-score* [Chinchor and Sundheim 1993]. Por fim, a utilidade prática dos *scores* foi analisada em cenários de previsão seletiva por meio de curvas de risco seletivo e da Área sob a Curva de Risco Seletivo (AURC) [Zhou et al. 2024], permitindo avaliar a capacidade dos métodos em priorizar previsões potencialmente mais confiáveis.

5. Resultados e Discussão

Esta seção apresenta os resultados da estimativa de dificuldade de previsão. Inicialmente, é analisada a distribuição dos erros observados. Em seguida, são comparados os métodos avaliados, realizada uma análise de ablação dos grupos de características e, por fim, investigada a utilidade prática dos scores em cenários de previsão seletiva.

5.1. Caracterização da Dificuldade de Previsão

A Figura 2 apresenta a distribuição dos erros de previsão por frequência temporal em escala logarítmica. Observa-se que as séries *Weekly* tendem a apresentar menores erros, enquanto as séries *Monthly* exibem maior dispersão e valores típicos mais elevados. As séries *Yearly* apresentam comportamento intermediário.

Todas as frequências apresentam forte assimetria, com concentração de observações em baixos níveis de erro e uma pequena parcela de casos difíceis. Esse comportamento justifica a utilização do percentil P90 para identificação de previsões difíceis.

5.2. Comparação dos Métodos de Estimativa de Dificuldade

A Tabela 1 apresenta os melhores resultados obtidos para cada frequência temporal, considerando métodos supervisionados, métodos latentes não supervisionados e baselines derivados da série temporal. A seleção foi realizada utilizando o *F1-score* P90 como critério principal.

Os métodos supervisionados apresentaram os melhores resultados em todas as frequências, enquanto os métodos não supervisionados e os baselines permaneceram abaixo, indicando que as representações latentes e o histórico de erro fornecem informações mais discriminativas para estimar a dificuldade.

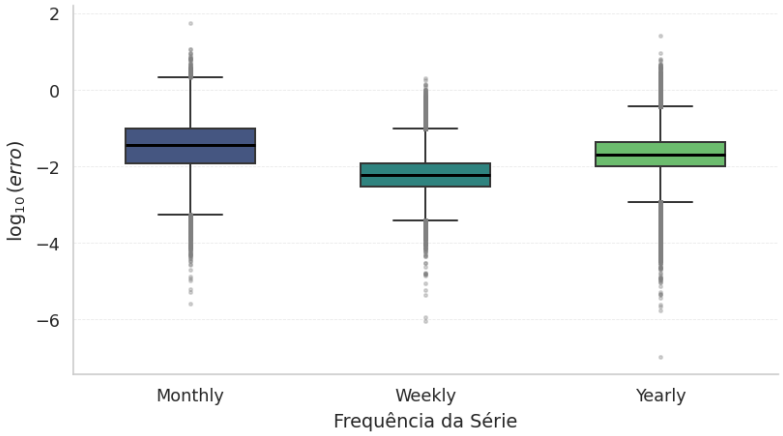


Figura 2. Distribuição do erro de previsão por frequência temporal em escala logarítmica. As linhas centrais representam as medianas dos grupos e os pontos correspondem a observações consideradas outliers segundo o critério do boxplot.

Esses resultados sugerem que a combinação de representações latentes, características estatísticas da série e histórico recente de erro constitui uma abordagem eficaz para estimar a dificuldade das previsões. Diferentemente de métodos tradicionais de quantificação de incerteza, frequentemente baseados em inferência Bayesiana, *ensembles* ou múltiplas passagens durante a inferência, nossa abordagem reutiliza informações produzidas pelo próprio modelo de previsão. Assim, os resultados complementam a discussão de [Gawlikowski et al. 2023], ao indicar uma alternativa de menor custo computacional para estimar a confiabilidade das previsões.

Tabela 1. Melhores resultados por método e frequência temporal. Os melhores resultados de cada frequência estão destacados em negrito.

	Monthly			Weekly			Yearly		
Método	Spearman	ROC-AUC	F1	Spearman	ROC-AUC	F1	Spearman	ROC-AUC	F1
Reg. Logística	**0.865**	**0.960**	**0.716**	0.933	0.955	0.776	**0.635**	**0.886**	**0.598**
Árvore	0.742	0.951	0.708	0.876	0.953	0.768	0.617	0.880	0.593
XGBoost	0.809	0.959	0.689	**0.895**	**0.957**	**0.780**	0.626	0.886	0.589
Amplitude	0.663	0.833	0.437	0.560	0.753	0.475	0.271	0.668	0.335
Variância	0.648	0.825	0.411	0.546	0.748	0.477	0.266	0.664	0.329
Norma Radial	0.481	0.745	0.343	0.469	0.714	0.470	0.275	0.660	0.324
KDE	0.250	0.659	0.245	0.281	0.599	0.401	0.188	0.646	0.325
GMM	0.376	0.675	0.229	0.268	0.615	0.403	0.181	0.642	0.317

5.3. Análise de Ablação

A Tabela 2 apresenta os resultados da análise de ablação considerando diferentes grupos de características. Os melhores desempenhos foram obtidos pelos grupos que incorporam informações do histórico de erro, alcançando valores de *F1-score* próximos de 0,60 e ROC-AUC superiores a 0,91. Em particular, o grupo composto exclusivamente pelo histórico de erro apresentou o maior valor de Spearman (0,709), indicando que essa é a principal fonte de informação para estimativa da dificuldade. As representações latentes superaram os atributos extraídos diretamente da série temporal, mas permaneceram

abaixo dos grupos que utilizam histórico de erro, sugerindo que fornecem sinais complementares para a estimativa da dificuldade.

Tabela 2. Análise de ablação dos grupos de características. Os valores correspondem ao desempenho médio obtido por cada grupo. Os melhores resultados de cada métrica estão destacados em negrito.

Grupo de características	Nº features	Spearman	ROC-AUC P90	F1 P90
Input + Histórico	10	0.706	0.913	**0.600**
Todas	18	0.699	**0.915**	**0.600**
Histórico de erro	8	**0.709**	0.914	0.599
Latente + Histórico	16	0.699	**0.915**	0.599
Latente + Input	10	0.505	0.784	0.379
Latente	8	0.463	0.755	0.349
Input	2	0.342	0.673	0.260

5.4. Predição Seletiva e Curvas de Risco

A Figura 3 apresenta as curvas de risco seletivo para as frequências *Monthly*, *Weekly* e *Yearly*. Nessa análise, os *scores* de dificuldade são utilizados para ordenar as previsões, mantendo-se apenas uma fração daquelas consideradas mais confiáveis. A cobertura representa a proporção de previsões selecionadas, enquanto o risco corresponde ao erro médio observado nesse subconjunto.

Os modelos supervisionados, especialmente a Regressão Logística e o XGBoost, produziram as menores curvas de risco em todas as frequências, indicando maior capacidade de identificar previsões potencialmente difíceis antes da observação do erro real. Essa vantagem é mais evidente em baixas coberturas, nas quais apenas as previsões de maior confiança são mantidas. Nesse cenário, os meta-modelos superam tanto os métodos baseados em representações latentes (Norma Radial, GMM e KDE) quanto os *baselines* derivados da série temporal (amplitude e variância), com diferenças mais acentuadas em *Monthly* e *Yearly*.

À medida que a cobertura aumenta, todas as curvas convergem para o erro médio global de cada frequência. Esse comportamento está de acordo com o princípio da predição seletiva descrito por [Geifman and El-Yaniv 2017], segundo o qual bons *scores* de confiança permitem reduzir o risco ao priorizar previsões mais confiáveis. Os resultados mostram que os *scores* produzidos pelos meta-modelos são eficazes para apoiar estratégias de previsão seletiva.

6. Conclusão

Este trabalho investigou a estimativa da dificuldade de previsão como um problema de meta-aprendizado supervisionado, integrando representações latentes extraídas do N-BEATS, características da série temporal e histórico recente de erro.

Os resultados mostraram que os meta-modelos supervisionados superam abordagens baseadas apenas em representações latentes ou características simples da série temporal. A análise de ablação indicou que o histórico recente de erro é a principal fonte de informação para a estimativa da dificuldade, enquanto as representações latentes fornecem informações complementares. Além disso, os experimentos de previsão seletiva

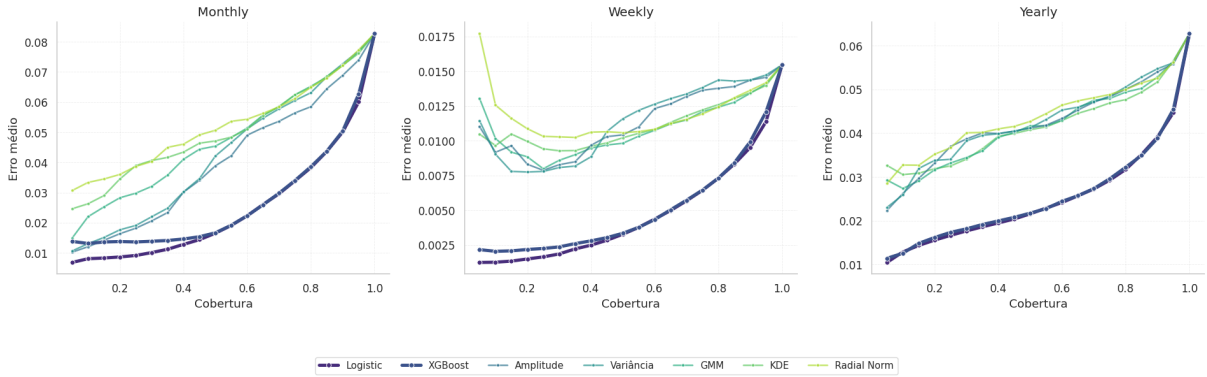


Figura 3. Curvas de risco seletivo para as frequências *Monthly, Weekly* e *Yearly*. Menores valores indicam maior capacidade de priorizar previsões confiáveis.

demonstraram que os *scores* produzidos permitem priorizar previsões mais confiáveis e reduzir o risco associado às previsões selecionadas.

Como limitações, a avaliação foi realizada apenas com o N-BEATS nas frequências *Monthly, Weekly* e *Yearly* do *benchmark* M4. Além disso, a abordagem depende do histórico recente de erro e utiliza uma definição binária de dificuldade baseada no percentil P90.

Como trabalhos futuros, pretende-se avaliar a metodologia em outras arquiteturas e conjuntos de dados, bem como investigar estratégias menos dependentes do histórico de erro e abordagens de regressão para estimar continuamente o grau de dificuldade das previsões.

Agradecimentos

O autor Leonardo Matos recebeu apoio da Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) por meio da bolsa Processo nº 88887.318495/2026-00, no âmbito do Programa PRAPG.

O autor Paulo Salgado Gomes de Mattos Neto agradece ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela bolsa de Produtividade em Pesquisa (Processo nº 306934/2025-6) e à Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE) pelo apoio financeiro concedido por meio do Processo APQ-0018-1.03/25.

1

Referências

Arvanitidis, G., Hansen, L. K., and Hauberg, S. (2017). Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*.

Bojanowski, P., Joulin, A., Lopez-Paz, D., and Szlam, A. (2017). Optimizing the latent space of generative networks. *arXiv preprint arXiv:1707.05776*.

[1]O código-fonte e os scripts utilizados nos experimentos estão disponíveis em: `https://github.com/Laury-SS/nbeats-forecast-difficulty`.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.

Casolaro, A., Capone, V., Iannuzzo, G., and Camastra, F. (2023). Deep learning for time series forecasting: Advances and open problems. *Information*, 14(11):598.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

Chinchor, N. and Sundheim, B. M. (1993). Muc-5 evaluation metrics. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20(2):215–232.

Ferchichi, A., Chihaoui, M., Toujani, R., Ferchichi, A., and Hendaoui, F. (2025). Evidential uncertainty quantification with multiple deep learning architectures for spatiotemporal drought forecasting. *Neural Computing and Applications*, 37(15):8773–8797.

Freeborough, W. and van Zyl, T. (2022). Investigating explainability methods in recurrent neural network architectures for financial time series data. *Applied Sciences*, 12(3):1427.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al. (2023). A survey of uncertainty in deep neural networks. *Artificial intelligence review*, 56(Suppl 1):1513–1589.

Geifman, Y. and El-Yaniv, R. (2017). Selective classification for deep neural networks.

Guo, T., Lin, T., and Lu, Y. (2018). An interpretable lstm neural network for autoregressive exogenous model. *arXiv preprint arXiv:1804.05251*.

Hyndman, R. J. and Athanasopoulos, G. (2021). *Forecasting: Principles and Practice*. OTexts, 3 edition.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688.

Jing, T., Zheng, P., Xia, L., and Liu, T. (2022). Transformer-based hierarchical latent space vae for interpretable remaining useful life prediction. *Advanced Engineering Informatics*, 54:101781.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2020). The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74. M4 Competition.

Oreshkin, B. N., Carpov, D., Chapados, N., and Bengio, Y. (2020). N-beats: Neural basis expansion analysis for interpretable time series forecasting.

Papacharalampous, G., Tyralis, H., Pechlivanidis, I. G., Grimaldi, S., and Volpi, E. (2022). Massive feature extraction for explaining and foretelling hydroclimatic time series forecastability at the global scale. *Geoscience Frontiers*, 13(3):101349.

Ponce-Flores, M., Frausto-Solís, J., Santamaría-Bonfil, G., Pérez-Ortega, J., and González-Barbosa, J. J. (2020). Time series complexities and their relationship to forecasting performance. *Entropy*, 22(1):89.

Postels, J., Blum, H., Cadena, C., Siegwart, R., Van Gool, L., and Tombari, F. (2020). Quantifying aleatoric and epistemic uncertainty using density estimation in latent space. *arXiv preprint arXiv:2012.03082*, 1.

Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, 28(1):71–72.

Reynolds, D. A. et al. (2009). Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663):3.

Rojat, T., Puget, R., Filliat, D., Del Ser, J., Gelin, R., and Díaz-Rodríguez, N. (2021). Explainable artificial intelligence (xai) on timeseries data: A survey. *arXiv preprint arXiv:2104.00950*.

Rongali, S., Rose, A. J., McManus, D. D., Bajracharya, A. S., Kapoor, A., Granillo, E., and Yu, H. (2020). Learning latent space representations to predict patient outcomes: Model development and validation. *Journal of Medical Internet Research*, 22(3):e16374.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, pages 832–837.

Song, X., Deng, L., Wang, H., Zhang, Y., He, Y., and Cao, W. (2024). Deep learning-based time series forecasting. *Artificial Intelligence Review*, 58(1):23.

Spearman, C. (1961). The proof and measurement of association between two things.

Van Amersfoort, J., Smith, L., Teh, Y. W., and Gal, Y. (2020). Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pages 9690–9700. PMLR.

Weisberg, S. (2001). Yeo-johnson power transformations. *Department of Applied Statistics, University of Minnesota. Retrieved June*, 1(2003):8.

Yunita, A., Pratama, M. I., Almuzakki, M. Z., Ramadhan, H., Akhir, E. A. P., Mansur, A. B. F., and Basori, A. H. (2025). Performance analysis of neural network architectures for time series forecasting: A comparative study of rnn, lstm, gru, and hybrid models. *MethodsX*, 15:103462.

Zhou, H., Van Landeghem, J., Popordanoska, T., and Blaschko, M. B. (2024). A novel characterization of the population area under the risk coverage curve (aurc) and rates of finite sample estimators. *arXiv preprint arXiv:2410.15361*.