Predição da Taxa de Desemprego no Brasil: Séries Temporais com Dados do CAGED, PNAD e SELIC

Aline Correa de Araújo[1]
Franciele Paterni[1]
Giovanna Sobral da Silva[1]
Guilherme Soares Frota[1]

[1]Faculdade de Computação e Informática (FCI)
Universidade Presbiteriana Mackenzie, São Paulo, Brasil
Emails: 10414773@mackenzista.com.br, 10414598@mackenzista.com.br,
10424600@mackenzista.com.br, 10416060@mackenzista.com.br

2025

Resumo

Este artigo apresenta o desenvolvimento de um modelo preditivo para a taxa de desocupação no Brasil, utilizando séries temporais e variáveis exógenas provenientes de bases de dados oficiais. A pesquisa integra informações do CAGED, PNAD Contínua e taxa SELIC, abrangendo o período de 2012 a 2025. O objetivo é construir um pipeline reprodutível capaz de gerar previsões de curto prazo com rigor estatístico e utilidade prática para órgãos públicos, pesquisadores e analistas econômicos.

A metodologia inclui coleta e padronização dos dados, análise exploratória (EDA), modelagem com LGBM e SARIMAX, além de validação temporal por *rolling-origin*. Os resultados mostram que o modelo SARIMAX com o saldo de empregos formais (CAGED) apresentou o melhor desempenho, com MAPE de 4,6%, enquanto o LGBM baseline teve MAPE de cerca de 40%. A inclusão da SELIC não melhorou o desempenho, indicando ausência de contribuição informativa relevante na especificação utilizada. A análise dos resíduos e da validação temporal reforça a adequação do modelo para previsão de curto prazo.

Conclui-se que o SARIMAX com CAGED fornece previsões estáveis e interpretáveis, sendo uma abordagem eficiente para monitoramento da dinâmica recente do mercado de trabalho brasileiro. São sugeridas extensões futuras envolvendo novas variáveis, granularidades regionais e modelos híbridos.

Palavras-chave: desemprego; séries temporais; SARIMAX; CAGED; PNAD.

Abstract

This paper presents the development of a predictive model for Brazil's unemployment rate using time series and exogenous variables from official public datasets. The study integrates data from the CAGED, PNAD Continuous Survey and the SELIC interest rate, covering the period from 2012 to 2025. The objective is to build a reproducible pipeline capable of delivering short-term forecasts with statistical rigor and practical utility for public agencies, researchers, and economic analysts.

The methodology includes data preprocessing, exploratory data analysis (EDA), modeling with LGBM and SARIMAX, and temporal validation through rolling-origin. Results show that the SARIMAX model with CAGED performed best, achieving a 4.6% MAPE, while the baseline LGBM model presented a MAPE of approximately 40%. The inclusion of SELIC did not improve performance, indicating limited predictive value in the tested specification. Residual analysis and temporal validation support the model's adequacy for short-term forecasting.

The study concludes that the SARIMAX model with CAGED provides stable and interpretable forecasts, making it an effective approach for monitoring the recent dynamics of the Brazilian labor market. Future work includes incorporating additional exogenous variables, regional granularity, and hybrid models.

Keywords: unemployment; time series; SARIMAX; CAGED; PNAD.

1 Introdução

O mercado de trabalho brasileiro é influenciado por oscilações macroeconômicas, políticas públicas e ciclos econômicos, refletindo-se tanto na geração de empregos quanto na taxa de desocupação (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 2025). Variáveis como o saldo de empregos formais e a taxa básica de juros (SELIC) desempenham papel central nesse processo, pois afetam o nível de atividade econômica, o consumo e o investimento (BANCO CENTRAL DO BRASIL, 2025). Compreender essas dinâmicas é fundamental para apoiar políticas públicas, planejamento econômico e decisões estratégicas.

A previsão da taxa de desemprego constitui ferramenta relevante em contextos de incerteza econômica, pois permite antecipar tendências de curto prazo e orientar ações governamentais voltadas ao trabalho decente e ao crescimento econômico, em alinhamento ao ODS 8. A utilização de séries temporais e de indicadores oficiais, como CAGED e PNAD Contínua, possibilita capturar padrões estruturais do mercado de trabalho brasileiro (BEIRãO; GONçALVES; NETO, 2021).

A lacuna que motiva este estudo reside na necessidade de avaliar, com rigor estatístico, a capacidade preditiva de modelos estruturados frente a abordagens modernas de aprendizagem de máquina, em um contexto de séries relativamente curtas e marcadas por choques econômicos, como a pandemia de COVID-19. Assim, o problema de pesquisa pode ser sintetizado na seguinte questão: *quais modelos apresentam desempenho mais robusto e interpretável para prever a taxa de desocupação no Brasil utilizando dados oficiais e variáveis exógenas?*

O objetivo geral deste trabalho é desenvolver um pipeline reprodutível para previsão da taxa de desocupação brasileira, analisando séries temporais provenientes do CAGED, PNAD Contínua e SELIC. Os objetivos específicos incluem:

- realizar o tratamento, padronização e integração das bases de dados (CAGED, PNAD e SELIC);
- analisar padrões de tendência, sazonalidade e estacionariedade das séries;
- construir e comparar modelos preditivos, como SARIMAX e LGBM;
- avaliar o desempenho dos modelos em métricas e validação temporal;
- apresentar previsões e *insights* que possam orientar decisões no âmbito de políticas públicas.

Bases de Dados Utilizadas

Foram utilizadas três bases oficiais e públicas. O CAGED fornece dados mensais e trimestrais de admissões, desligamentos e saldo de empregos formais (MINISTÉRIO DO TRABALHO E EMPREGO, 2025). A PNAD Contínua (Tabela 4099) apresenta taxas trimestrais de desocupação da população com 14 anos ou mais (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 2025). A taxa SELIC foi obtida por meio do Sistema Gerenciador de Séries Temporais (SGS) do Banco Central do Brasil, agregada de forma trimestral para compatibilização (BANCO CENTRAL DO BRASIL, 2025). As séries abrangem o período de 2012 a 2025 e foram tratadas em formato CSV, com granularidade transformada para o calendário trimestral.

Essas bases permitem uma análise abrangente e reprodutível do comportamento do mercado de trabalho brasileiro, tornando possível o desenvolvimento de modelos estatísticos e computacionais adequados ao objetivo da pesquisa.

2 Referencial Teórico

As séries temporais constituem uma classe de dados em que a dependência temporal é o principal fator de análise, permitindo identificar tendências, ciclos e padrões sazonais (ENDERS, 2015). Os modelos tradicionais para lidar com esse tipo de dado incluem o ARIMA (*AutoRegressive Integrated Moving Average*), que combina componentes autorregressivos (AR), de integração (I) e de médias móveis (MA). Quando as séries apresentam padrões sazonais, recorre-se a modelos estendidos como o SARIMA, que incorpora a sazonalidade, e o SARIMAX, que permite a inclusão de variáveis exógenas relevantes, como indicadores macroeconômicos (BEIRãO; GONçALVES; NETO, 2021; BOX et al., 2015).

Um conceito central em séries temporais é a estacionariedade, que ocorre quando a média e a variância permanecem constantes ao longo do tempo. Séries com forte tendência ou sazonalidade não são estacionárias e precisam ser transformadas, geralmente por diferenciação ou aplicação de transformações como logaritmos, para que métodos como ARIMA e SARIMA sejam aplicáveis de forma adequada (ENDERS, 2015). Testes formais de raiz unitária, como o teste de Dickey e Fuller (DICKEY; FULLER, 1979), são amplamente utilizados para diagnosticar a presença de não estacionariedade.

No contexto do desemprego, diversos estudos no Brasil têm demonstrado a relevância da modelagem estatística. Becker (2010) analisou a taxa de desemprego da Região Metropolitana de Porto Alegre e avaliou a aplicação de modelos como Holt Winters e SARIMA, concluindo que modelos sazonais oferecem melhor ajuste em séries com ciclos

econômicos regulares. Beirão, Gonçalves e Neto (2021), ao estudar a taxa de desemprego brasileira com a metodologia Box Jenkins, confirmam a necessidade de diferenciação para lidar com a não estacionariedade da série.

Além dos métodos estatísticos clássicos, há uma crescente utilização de algoritmos de aprendizagem de máquina. Trabalhos recentes apontam que abordagens híbridas, que combinam modelos de séries temporais com *machine learning*, podem apresentar robustez maior e menor erro de previsão, sobretudo em cenários com muitas variáveis explicativas. Estudos nacionais publicados em periódicos de economia têm explorado modelos fatoriais de alta dimensão aliados a técnicas de seleção de variáveis para prever indicadores macroeconômicos, incluindo a taxa de desemprego, alcançando ganhos significativos de precisão.

A literatura indica, portanto, que a modelagem de séries temporais voltada para a previsão de indicadores de mercado de trabalho no Brasil é um tema relevante, com aplicações práticas em políticas públicas e gestão econômica. O uso combinado de técnicas estatísticas clássicas e métodos modernos de aprendizagem de máquina surge como alternativa promissora, unindo a interpretabilidade dos primeiros à flexibilidade dos segundos.

3 Metodologia

A metodologia adotada neste trabalho foi estruturada como um fluxo sequencial de etapas, que garante reprodutibilidade, clareza e coerência com boas práticas de modelagem de séries temporais. O objetivo foi prever a taxa trimestral de desocupação no Brasil a partir da integração de dados oficiais do CAGED, PNAD Contínua e taxa SELIC. O fluxo metodológico seguiu as etapas descritas a seguir.

3.1 Coleta e Integração das Bases

Foram reunidas três fontes oficiais de dados: (i) microdados mensais do CAGED, consolidados em frequência trimestral; (ii) taxa de desocupação da PNAD Contínua (tabela 4099), já em periodicidade trimestral; e (iii) série diária da taxa SELIC disponibilizada pelo Sistema Gerenciador de Séries Temporais (SGS), posteriormente agregada para médias trimestrais. Todas as bases foram padronizadas no calendário Q-DEC, assegurando alinhamento temporal entre as séries.

3.2 Pré-processamento e Engenharia de Atributos

O processo de pré-processamento incluiu limpeza de valores ausentes, padronização de tipos e identificação de inconsistências. A engenharia de atributos contemplou: (i) construção do saldo do CAGED (admissões menos desligamentos); (ii) criação da média móvel trimestral *roll3*; (iii) transformação *asinh* para estabilização de variância; (iv) geração de dummies sazonais para os quatro trimestres; e (v) inclusão de um indicador estrutural para o período pós-pandemia (a partir de 2021). As variáveis exógenas foram alinhadas e normalizadas quando necessário.

3.3 Análise de Séries Temporais

A série-alvo da taxa de desocupação passou por decomposição aditiva para identificação dos componentes de tendência, sazonalidade e ruído. Testes de estacionariedade foram conduzidos, com destaque para o teste ADF. As funções de autocorrelação (ACF) e

autocorrelação parcial (PACF) auxiliaram na definição das ordens iniciais da modelagem, indicando necessidade de diferenciação regular ($d = 1$) e estrutura sazonal trimestral ($s = 4$).

3.4 Análise Exploratória

Foram avaliadas a evolução histórica da taxa de desocupação, os ciclos de emprego formal capturados pelo CAGED, a trajetória da taxa SELIC e suas possíveis relações com variáveis do mercado de trabalho. Também foram identificados regimes distintos — sobretudo durante o período de 2020–2021 — além de padrões sazonais e correlações relevantes entre as variáveis.

3.5 Modelagem

A etapa de modelagem avaliou duas famílias de modelos. Como *baseline*, utilizou-se o LGBMRegressor, sem calibração manual e com variáveis derivadas do CAGED e dummies sazonais. A modelagem principal seguiu a estrutura SARIMAX, com duas especificações: (i) SARIMAX utilizando o CAGED como variável exógena; e (ii) SARIMAX com CAGED e SELIC em experimento de ablação. A especificação final adotada foi $(1, 1, 2) \times (1, 0, 1)_4$. As estimativas foram avaliadas por critérios de significância e diagnóstico dos resíduos, incluindo testes de Ljung-Box, ARCH-LM e inspeção visual.

3.6 Validação

A avaliação preditiva utilizou duas estratégias complementares. A primeira consistiu em um *holdout* temporal, no qual o conjunto de teste abrangeu o período de 2024Q3 a 2025Q2. A segunda aplicou validação *rolling-origin* com janela expandida, totalizando 13 *folds*, permitindo observar a estabilidade dos modelos ao longo do tempo. As métricas utilizadas foram MAE, RMSE e MAPE, comparadas ao modelo ingênuo sazonal (*sNaive*).

3.7 Avaliação Crítica

Os resultados das duas abordagens foram comparados à luz de desempenho preditivo, estabilidade temporal e interpretabilidade. A deterioração do desempenho ao incluir a SELIC foi discutida com base em possíveis fontes de colinearidade, instabilidade da política monetária e perda de granularidade após agregação da taxa diária.

3.8 Produto Final

O fluxo metodológico resultou em um pipeline completamente reprodutível, com notebook executável, gráficos de previsão com intervalos de confiança, tabela comparativa final dos modelos e relatório técnico em conformidade com a norma ABNT. A solução proposta encontra-se alinhada ao ODS 8, ao oferecer ferramenta analítica para monitoramento do mercado de trabalho brasileiro.

4 Resultados

Os resultados dos modelos avaliados foram obtidos a partir do conjunto de teste correspondente ao período entre 2024Q3 e 2025Q2. A Tabela 1 apresenta uma comparação direta entre as abordagens, considerando as métricas MAE, RMSE e MAPE.

Tabela 1 – Desempenho dos modelos no conjunto de teste

Modelo	MAE	RMSE	MAPE (%)
LGBM (baseline)	2.5551	2.7175	40.7191
SARIMAX (CAGED)	**0.2895**	**0.3236**	**4.6493**
SARIMAX (CAGED + SELIC)	0.7134	0.8264	11.2236

O LGBM baseline apresentou desempenho inferior, produzindo previsões excessivamente suavizadas e pouco responsivas às inflexões da série. O erro percentual superior a 40% evidencia a limitação do modelo para capturar a estrutura temporal presente em séries trimestrais curtas.

O modelo SARIMAX com CAGED foi o que apresentou melhor desempenho, com MAE inferior a 0,30 ponto percentual e MAPE abaixo de 5%. As previsões acompanharam de maneira adequada a trajetória do desemprego no período de teste, reproduzindo com precisão a tendência recente e a sazonalidade trimestral da série.

A inclusão da SELIC como segunda variável exógena resultou em piora do desempenho. Embora o modelo ainda apresente erros moderados, o aumento significativo de MAE e MAPE indica que a variável não agregou informação preditiva relevante na especificação utilizada.

A análise gráfica (Real vs. Previsto) confirma estes achados. O SARIMAX com CAGED mostra alta aderência à série real, enquanto o LGBM apresenta previsões mais lineares e o SARIMAX com SELIC demonstra leve enviesamento sistemático. Os diagnósticos de resíduos do modelo SARIMAX confirmam ausência de autocorrelação significativa e ausência de heterocedasticidade, indicando boa especificação estatística.

Por fim, a validação *rolling-origin* revelou convergência do erro nas janelas mais recentes, reforçando a capacidade do modelo SARIMAX de capturar a dinâmica recente do mercado de trabalho brasileiro.

5 Discussão

A comparação entre os modelos permite uma interpretação ampla sobre a dinâmica do desemprego e sobre os limites das abordagens avaliadas. Os resultados indicam que métodos estatísticos estruturados, como o SARIMAX, são mais adequados para séries trimestrais curtas e com forte dependência temporal do que modelos de aprendizagem de máquina genéricos, como o LGBM.

O excelente desempenho do SARIMAX com o CAGED como exógena está alinhado com a literatura que aponta o emprego formal como um indicador contemporâneo relevante da taxa de desocupação. O CAGED incorpora movimentos de contratação e demissão que refletem rapidamente alterações nas condições econômicas, tornando-se um preditor eficiente em horizontes curtos.

Por outro lado, a piora significativa no modelo SARIMAX ao incluir a SELIC traz uma contribuição analítica importante. Embora a taxa básica de juros seja um dos principais instrumentos de política monetária, seu efeito sobre o desemprego não é instantâneo e tampouco linear. Três fatores ajudam a explicar o resultado observado:

1. **Instabilidade estrutural da política monetária brasileira**: a SELIC apresenta ciclos abruptos entre 2015 e 2023, dificultando a modelagem de um efeito estatisticamente consistente na frequência trimestral.

2. **Colinearidade com ciclos já capturados pelo CAGED**: parte do impacto da SELIC sobre o emprego formal já está refletida nas variações do saldo de empregos, reduzindo a utilidade preditiva da variável no modelo.

3. **Perda de granularidade**: a agregação de uma série originalmente diária para médias trimestrais dilui movimentos importantes, prejudicando a captura de relações dinâmicas.

O LGBM baseline, por sua vez, apresentou limitações típicas de modelos de *machine learning* aplicados a séries temporais curtas: forte suavização, incapacidade de capturar regimes distintos e alta sensibilidade ao tamanho da base. Seu MAPE acima de 40% reforça que abordagens baseadas unicamente em aprendizado de máquina não são adequadas para esse problema sem engenharia de atributos muito mais sofisticada.

A combinação entre resultados quantitativos, análise dos resíduos e validação expandida revela que o SARIMAX com CAGED é o modelo mais robusto, interpretável e estável para previsão de curto prazo. No entanto, reconhece-se que sua capacidade diminui em cenários de mudanças estruturais profundas, como crises econômicas ou choques pandêmicos.

No conjunto, os resultados reforçam que previsões econômicas requerem modelos que incorporem estrutura temporal explícita, conhecimento do domínio e avaliação crítica contínua. A discussão apresentada evidencia os pontos fortes do modelo final, suas limitações e o caminho para possíveis melhorias metodológicas futuras.

6 Análise Crítica

A avaliação dos modelos e do processo completo de desenvolvimento permite identificar não apenas os pontos fortes, mas também as limitações estruturais que impactam a capacidade preditiva da solução. Embora o SARIMAX com CAGED tenha apresentado desempenho superior, alguns aspectos merecem atenção crítica.

Primeiro, a série da taxa de desocupação é curta e trimestral, o que limita a diversidade de padrões disponíveis para o aprendizado dos modelos. Esse fator restringe o uso de abordagens mais complexas e reduz o potencial de generalização, especialmente em períodos com mudanças bruscas nas condições macroeconômicas.

Segundo, a instabilidade provocada por eventos extraordinários — como a pandemia de COVID-19 e as mudanças abruptas na política monetária entre 2015 e 2023 — introduz rupturas estruturais que desafiam modelos lineares. O SARIMAX, apesar de eficiente em períodos estáveis, tende a sofrer degradação quando ocorrem quebras de tendência ou alterações abruptas no comportamento das variáveis exógenas.

Terceiro, a utilização da SELIC como exógena evidenciou limitações metodológicas importantes. A agregação trimestral dilui movimentos de curto prazo, e a relação entre juros e emprego não é contemporânea. Assim, o modelo não capturou efeitos defasados ou não lineares da política monetária, o que explica a deterioração observada na ablação.

Quarto, o LGBM baseline demonstrou que modelos de *machine learning* dependem de maior granularidade, volume de dados e engenharia de atributos específica para séries temporais. Em bases curtas, esses modelos tendem a subaprender padrões e produzir previsões suavizadas, o que reduz sua utilidade prática.

Por fim, embora o pipeline desenvolvido seja reprodutível e coerente, uma aplicação real exigiria monitoramento contínuo, atualização trimestral dos dados e reavaliação da estrutura do modelo sempre que novos regimes econômicos surgirem. A solução não deve ser interpretada como previsor estrutural de longo prazo, mas como ferramenta operacional de curto prazo.

A análise crítica evidencia que, apesar dos bons resultados, há limitações inerentes à natureza das séries e ao contexto econômico brasileiro, reforçando a necessidade de aprimoramentos metodológicos futuros.

7 Conclusão

Este trabalho apresentou o desenvolvimento de um pipeline completo e reprodutível para previsão da taxa de desocupação no Brasil, integrando dados oficiais do CAGED, PNAD Contínua e taxa SELIC. A análise exploratória revelou tendência moderada, sazonalidade trimestral marcada e rupturas associadas à pandemia, aspectos que motivaram a adoção de modelos sazonais com diferenciação.

Os resultados mostraram que o SARIMAX com CAGED é o modelo mais eficaz para previsões trimestrais no curto prazo, alcançando MAPE inferior a 5% e apresentando excelente aderência à série real. A estabilidade nos testes *rolling-origin* e os diagnósticos de resíduos reforçaram a adequação estatística da especificação adotada.

A inclusão da SELIC não gerou ganhos preditivos e, ao contrário, deteriorou o desempenho do modelo. Esse resultado, embora inesperado, demonstrou a importância de avaliar empiricamente hipóteses macroeconômicas e reforçou que nem todas as variáveis teoricamente relevantes possuem sinal preditivo na frequência utilizada.

Apesar dos resultados positivos, reconhece-se que limitações permanecem: a série é curta, sujeita a choques estruturais e dependente de divulgações trimestrais, o que reduz a capacidade de generalização em horizontes mais longos. Assim, as previsões são mais adequadas para cenários de monitoramento recente do que para projeções estruturais de longo prazo.

Como desdobramentos futuros, recomenda-se incorporar novas variáveis macroeconômicas — como PIB, IPCA, indicadores de confiança e crédito —, testar especificações com defasagens mais profundas e explorar granularidades regionais. Modelos híbridos, que combinem estruturas econométricas e técnicas de *machine learning*, também representam caminho promissor.

No conjunto, o projeto entrega uma solução estatisticamente sólida, interpretável e operacionalizável, capaz de apoiar o monitoramento do mercado de trabalho brasileiro no curto prazo e fornecer subsídios quantitativos para análises econômicas e políticas públicas.

Referências

BANCO CENTRAL DO BRASIL. *Sistema Gerenciador de Séries Temporais – Taxa Selic*. 2025. <https://www.bcb.gov.br>. Acesso em: 25 set. 2025.

BECKER Áureo de P. *Modelos de previsão para a taxa de desemprego na Região Metropolitana de Porto Alegre*. Porto Alegre: UFRGS, 2010.

BEIRãO Éder de S.; GONçALVES, M. E.; NETO, D. R. da S. Desemprego no brasil: uma análise empírica de previsão baseada na metodologia box–jenkins. *Revista Economia e Políticas Públicas*, v. 9, n. 1, p. 131–160, 2021. Acesso em: 25 set. 2025. Disponível em: <https://doi.org/10.46551/epp2021917>.

BOX, G. E. P. et al. *Time Series Analysis: Forecasting and Control*. 5. ed. Hoboken: Wiley, 2015.

DICKEY, D. A.; FULLER, W. A. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, v. 74, n. 366, p. 427–431, 1979. Acesso em: 25 set. 2025. Disponível em: <https://doi.org/10.1080/01621459.1979.10482531>.

ENDERS, W. *Applied Econometric Time Series*. 4. ed. Hoboken: John Wiley & Sons, 2015. Acesso em: 25 set. 2025. Disponível em: <https://www.wiley.com/en-us/Applied+Econometric+Time+Series%2C+4th+Edition-p-9781118808566>.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. *Pesquisa Nacional por Amostra de Domicílios Contínua – PNAD Contínua*. 2025. <https://www.ibge.gov.br>. Acesso em: 25 set. 2025.

MINISTÉRIO DO TRABALHO E EMPREGO. *Cadastro Geral de Empregados e Desempregados – CAGED (Microdados)*. 2025. Ftp://ftp.mtps.gov.br/pdet/microdados/. Acesso em: 25 set. 2025.