Predição da Taxa de Desemprego no Brasil

Séries Temporais com Dados do CAGED, PNAD e SELIC

Aline Correa de Araújo
10414773@mackenzista.com.br

Franciele Paterni
10414598@mackenzista.com.br

Giovanna Sobral da Silva
10424600@mackenzista.com.br

Guilherme Soares Frota
10416060@mackenzista.com.br

Universidade Presbiteriana Mackenzie
Faculdade de Computação e Informática (FCI)
São Paulo, Brasil

2025

Agenda do Projeto

INTRODUÇÃO E FUNDAMENTOS

1. Contexto e Motivação
2. Problema de Pesquisa
3. Objetivos
4. Bases de Dados
5. Referencial Teórico

DESENVOLVIMENTO METODOLÓGICO

6. Metodologia
7. Pré-processamento
8. Análise de Séries Temporais

MODELAGEM E RESULTADOS

9. Modelagem
10. Resultados
11. Discussão e Diagnósticos
12. Limitações

CONCLUSÕES FINAIS

13. Conclusões
14. Trabalhos Futuros
15. Contribuições
16. Referências e Agradecimentos

Contexto e Motivação

Dinâmica do Mercado

O mercado de trabalho brasileiro é altamente sensível a ciclos macroeconômicos e choques estruturais, como a pandemia de COVID-19, exigindo monitoramento constante.

Período do Estudo

A pesquisa abrange séries temporais de **2012 a 2025**, período marcado por recessões, reformas trabalhistas e mudanças na política monetária.

Indicadores Chave

O saldo de empregos formais (**CAGED**) e a taxa básica de juros (**SELIC**) atuam como termômetros centrais da atividade econômica e do investimento.

Abordagem Híbrida

Motivação para comparar métodos econométricos tradicionais (SARIMAX) com algoritmos de Machine Learning (LGBM) em dados oficiais.

Políticas Públicas

Previsões precisas de curto prazo são ferramentas essenciais para orientar decisões governamentais e estratégias de planejamento econômico em cenários de incerteza.

8 **Objetivo de Desenvolvimento Sustentável**

Trabalho Decente e Crescimento Econômico

Alinhamento com a Agenda 2030 da ONU

Problema de Pesquisa

A Questão Central

Quais modelos apresentam desempenho mais robusto e interpretável para prever a taxa de desocupação no Brasil utilizando variáveis exógenas e dados oficiais?

Desafios dos Dados

A análise enfrenta limitações impostas por **séries temporais curtas** e forte **sazonalidade trimestral**, o que restringe a capacidade de aprendizado de padrões complexos.

Abordagem Comparativa

O estudo confronta a robustez dos modelos econométricos estruturados contra a flexibilidade dos algoritmos de Machine Learning.

Choques Estruturais

O período de estudo (2012-2025) é marcado por eventos disruptivos, como a pandemia de **COVID-19** e ciclos voláteis de política monetária, gerando quebras de tendência.

Dilema Metodológico

Avaliar se o rigor estatístico de métodos clássicos supera a capacidade de adaptação de modelos modernos em cenários de alta instabilidade.

LACUNA DA PESQUISA

Investigar a eficácia preditiva em contextos de séries curtas e instáveis, onde a literatura aponta trade-offs entre interpretabilidade e precisão.

Objetivos do Estudo

OBJETIVO GERAL

Desenvolver um pipeline reprodutível para a previsão trimestral da taxa de desocupação brasileira, analisando séries temporais e variáveis exógenas.

Objetivos Específicos

- **Tratar e integrar** bases de dados oficiais do CAGED, PNAD Contínua e Taxa SELIC.
- **Analisar padrões** de tendência, sazonalidade e estacionariedade das séries temporais.
- **Construir e comparar** modelos preditivos estruturados (SARIMAX) e de aprendizado de máquina (LGBM).
- **Validar o desempenho** utilizando estratégias robustas como holdout temporal e validação rolling-origin.
- **Entregar previsões e insights** acionáveis para auxiliar na tomada de decisão e formulação de políticas públicas.

Bases de Dados Utilizadas

CAGED

MINISTÉRIO DO TRABALHO

Dados de Emprego Formal
Admissões, desligamentos e saldo mensal consolidado.

Agregado para frequência trimestral.

PNAD Contínua

IBGE (TABELA 4099)

Taxa de Desocupação
Série trimestral da população desocupada (14+ anos).

Variável alvo (target) do modelo.

Taxa SELIC

BANCO CENTRAL (SGS)

Taxa Básica de Juros
Série diária obtida via SGS e transformada.

Média trimestral para compatibilização.

Período de Análise
2012 a 2025 • Calendário Q-DEC

Padronização

Formato CSV | Normalização | Alinhamento Temporal

Referencial Teórico

Séries Temporais

Análise da dependência temporal identificando componentes estruturais como **tendência**, **sazonalidade** e **ciclos** econômicos (Enders, 2015).

Modelos ARIMA/SARIMA

Combinação de componentes Autorregressivos (AR), de Integração (I) e Médias Móveis (MA), com extensão para capturar sazonalidade (SARIMA).

Extensão SARIMAX

Incorporação de variáveis exógenas relevantes (X) ao modelo, como o saldo do CAGED e a taxa SELIC, para enriquecer a capacidade preditiva.

Estacionariedade

Conceito central para estabilidade de média e variância. Aplicação de testes de raiz unitária como **ADF (Dickey-Fuller)** para validar a necessidade de diferenciação.

Evidências na Literatura

- **Becker (2010):** Modelos sazonais oferecem melhor ajuste em séries com ciclos econômicos regulares.
- **Beirão et al. (2021):** Confirmação da necessidade de diferenciação para lidar com não estacionariedade no desemprego.
- **Box et al. (2015):** Fundamentação da metodologia Box-Jenkins para construção iterativa de modelos.

Metodologia - Visão Geral

Coleta e Integração

CAGED, PNAD, SELIC

Séries mensais e diárias

Pré-processamento

Limpeza e Tratamento

Padronização Q-DEC

Análise (EDA)

Decomposição STL

Testes de estacionariedade

Modelagem

SARIMAX vs LGBM

Séries vs Machine Learning

Validação

Rolling-Origin

Holdout temporal

Avaliação Final

Produto Final

Diagnóstico e Relatório

FREQUÊNCIA Trimestral (Q-DEC) Pipeline Reprodutível

MÉTRICAS DE AVALIAÇÃO MAE RMSE MAPE

Pré-processamento e Engenharia

TRATAMENTO DE DADOS

Pipeline robusto de limpeza, transformação e feature engineering para adequar as bases brutas aos requisitos dos modelos estatísticos.

Etapas do Pipeline

- **Limpeza de Dados**: Identificação e tratamento de valores ausentes e inconsistências.
- **Suavização**: Aplicação de média móvel trimestral (roll3) para redução de ruído.
- **Sazonalidade**: Criação de dummies trimestrais (Q1, Q2, Q3, Q4).
- **Saldo CAGED**: Construção da variável via cálculo (Admissões – Desligamentos).
- **Estabilização**: Transformação *asinh* para controle de variância nas séries.
- **Quebra Estrutural**: Indicador binário para o período pós-pandemia (≥ 2021).
- **Integração**: Alinhamento temporal e normalização das variáveis exógenas (CAGED e SELIC).

Análise de Séries Temporais

Perfil Estrutural da Série

Parâmetros identificados na análise exploratória

s = 4

SAZONALIDADE TRIMESTRAL

Aditiva

TIPO DE DECOMPOSIÇÃO

d = 1

ORDEM DE DIFERENCIAÇÃO

2020-21

CHOQUE ESTRUTURAL (COVID)

Base de dados: PNAD Contínua (2012-2025)

Decomposição Aditiva

Identificação clara de **tendência** de longo prazo e padrão de **sazonalidade** trimestral marcada.

Estacionariedade (Teste ADF)

Série não estacionária em nível. Necessidade de **diferenciação regular (d=1)** para estabilizar média e variância.

Funções ACF / PACF

Análise das funções de autocorrelação guiou a definição das ordens iniciais e confirmou a estrutura sazonal.

Regimes e Choques

Identificação de regimes distintos, com destaque para o **choque estrutural** no período 2020-2021 (Pandemia).

Modelagem e Especificação

LGBM Baseline

MACHINE LEARNING

ABORDAGEM GENÉRICA

Modelo LightGBM aplicado sem calibração manual (sem tuning) para estabelecer linha de base.

- Features: Defasagens do CAGED
- Sazonalidade: Dummies Trimestrais

SARIMAX (Principal)

MELHOR DESEMPENHO

MODELO ESTRUTURADO

Especificação final otimizada para capturar dinâmica sazonal e influência do emprego formal.

$$(1, 1, 2) \times (1, 0, 1)_4$$

- **Exógena:** Saldo CAGED
- Diferenciação Regular (d=1)

Experimento

ABLAÇÃO

TESTE DE HIPÓTESE

Avaliação da contribuição informativa da política monetária na predição.

- \+ Variação: Inclusão da **Taxa SELIC**
- → Objetivo: Testar causalidade preditiva

Critérios de Avaliação

Significância
Coeficientes ($p < 0.05$)

Autocorrelação
Teste Ljung-Box (Resíduos)

Homocedasticidade
Teste ARCH-LM

Resultados: Desempenho Preditivo

Conjunto de Teste: **2024Q3 — 2025Q2** (Holdout Temporal)

MODELO / ESPECIFICAÇÃO	MAE	RMSE	MAPE (%)
LGBM Baseline Sem tuning, features simples	2.56	2.72	40.72%
SARIMAX + CAGED MELHOR **Exógena: Saldo de Empregos**	**0.29**	**0.32**	**4.65%**
SARIMAX + CAGED + SELIC Teste de ablação	0.71	0.83	11.22%

Superioridade do SARIMAX: O erro percentual (MAPE) do melhor modelo é quase **10x menor** que o baseline de Machine Learning.

Impacto da SELIC: A inclusão da taxa de juros **piorou** as métricas (aumento de ~6.5 p.p. no MAPE), indicando ruído na série.

Resultados: Análise do Comportamento

SARIMAX + CAGED

Melhor Desempenho

Apresentou **alta aderência** à série real. O modelo conseguiu capturar corretamente a tendência de queda recente e reproduziu com precisão os movimentos da sazonalidade trimestral.

LGBM (Baseline)

Subajustado

Produziu previsões **excessivamente suavizadas**. Falhou em responder às inflexões da série e não capturou a dinâmica sazonal, comportando-se quase como uma média linear.

SARIMAX + SELIC

Viés Sistemático

Demonstrou um **leve viés sistemático** para cima. A inclusão da variável SELIC introduziu ruído ao modelo sem trazer ganho de informação preditiva relevante.

Precisão e Tendência

Enquanto o modelo com CAGED acompanhou a trajetória real, o baseline de Machine Learning (LGBM) mostrou-se incapaz de modelar a estrutura temporal complexa sem engenharia de atributos mais profunda.

Relevância das Variáveis

O saldo de empregos (CAGED) provou ser um previsor contemporâneo robusto. Em contrapartida, a SELIC não apresentou contribuição imediata, indicando possíveis efeitos defasados não capturados.

Diagnóstico Comparativo

A superioridade do SARIMAX+CAGED (MAPE ~4,6%) sobre o Baseline (~40%) confirma que, para séries curtas e sazonais, modelos econométricos estruturados superam algoritmos genéricos sem tuning.

Discussão: Por que a SELIC não ajudou?

1

Instabilidade Estrutural

A política monetária brasileira (2015–2023) apresentou ciclos abruptos e voláteis.

Essa inconsistência dificultou a modelagem estatística de um efeito regular em frequência trimestral.

2

Colinearidade com CAGED

Parte do impacto dos juros sobre a economia real já está refletida nas contratações.

Ao usar o **saldo de empregos**, o efeito da SELIC torna-se redundante no modelo.

3

Perda de Granularidade

A agregação de uma taxa originalmente diária para médias trimestrais dilui a informação.

Movimentos dinâmicos de curto prazo são suavizados, perdendo poder preditivo.

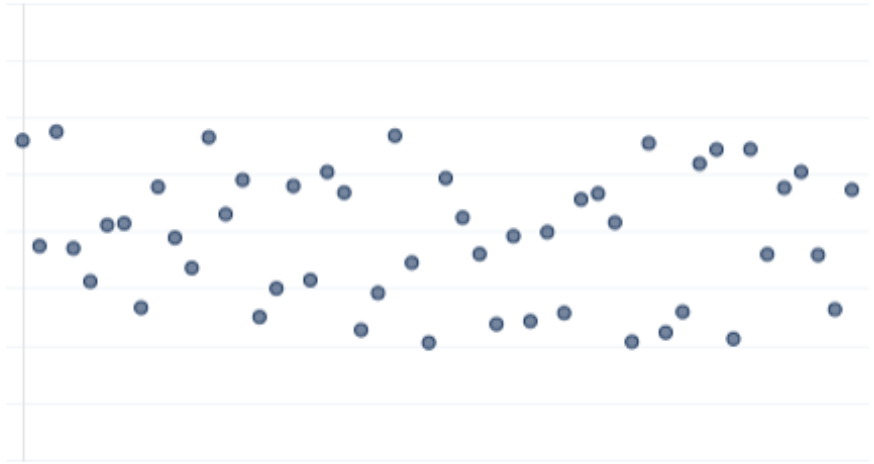
Conclusão Analítica

O efeito da taxa de juros sobre o desemprego não é instantâneo e nem linear. O modelo SARIMAX atual, focado em curto prazo, captura melhor a dinâmica através do fluxo direto de empregos (CAGED).

Diagnóstico do Modelo SARIMAX

Resíduos Padronizados

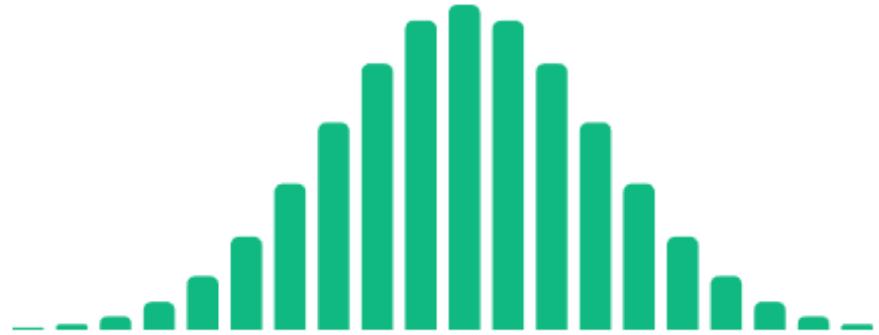
RUÍDO BRANCO



Distribuição aleatória em torno de zero sem padrões óbvios

Normalidade (Histograma)

NORMAL



Ausência de Autocorrelação

`p > 0.05`

Teste de **Ljung-Box** confirma que os resíduos se comportam como ruído branco, sem dependência temporal residual.

Homocedasticidade

`p > 0.05`

Teste **ARCH-LM** indica variância constante dos resíduos ao longo do tempo (ausência de efeitos ARCH).

Significância dos Parâmetros

`Sig. < 0.05`

Coeficientes AR, MA e da variável exógena (CAGED) são estatisticamente significativos e estáveis.

Especificação Estatística Válida

O modelo atende a todos os pressupostos teóricos, tornando as previsões e intervalos de confiança confiáveis para interpretação.

Análise Crítica e Limitações

RESTRIÇÃO DE DADOS

A série histórica curta e de frequência trimestral restringe a diversidade de padrões aprendíveis e limita a capacidade de generalização dos modelos em horizontes longos.

⚠ Desafios Estruturais e Conjunturais

Choques Estruturais

Rupturas como a pandemia de COVID-19 e ciclos monetários afetam a estabilidade dos parâmetros estatísticos.

Complexidade da SELIC

Os efeitos da política monetária não são contemporâneos e possuem não linearidades não capturadas pelo modelo.

Limitações do Machine Learning

Em bases curtas, modelos complexos como ML tendem a subaprender padrões (underfitting) e suavizar previsões.

Necessidade de Reavaliação

A solução exige monitoramento contínuo e recalibragem frequente para manter a aderência aos novos regimes.

Conclusões Principais

MELHOR MODELO

SARIMAX com CAGED

Apresentou o desempenho mais robusto e preciso entre todas as especificações testadas.

MAPE	4.6%
MAE	0.29 p.p.

- Previsões estáveis
- Alta aderência à série real

Curto Prazo e Interpretabilidade

O modelo demonstra ser altamente adequado para previsões de curto horizonte (1-4 trimestres), oferecendo interpretabilidade superior a métodos de caixa-preta (black-box).

SELIC: Sem Ganho Preditivo

A inclusão da taxa de juros não agregou valor na especificação testada, sugerindo que os efeitos da política monetária são defasados ou já absorvidos pelo emprego formal.

Pipeline Reprodutível

Desenvolvimento de um fluxo de trabalho transparente com rigor estatístico, validado por testes de resíduos e cross-validation (rolling-origin).

Monitoramento e ODS 8

A ferramenta contribui efetivamente para o monitoramento do mercado de trabalho brasileiro, apoiando decisões alinhadas ao crescimento econômico e trabalho decente.

Trabalhos Futuros

1

Novas Variáveis Exógenas

Incorporação de indicadores macroeconômicos como PIB, IPCA, índices de confiança do consumidor e volume de crédito.

2

Dinâmica Não Linear

Investigação de defasagens temporais mais profundas e efeitos de limiar (threshold) na relação entre juros e emprego.

3

Granularidade

Expansão da análise para recortes regionais (estados/regiões) e setoriais para capturar heterogeneidades do mercado.

4

Modelos Híbridos

Combinação da estrutura SARIMAX com algoritmos de Machine Learning (LSTM, XGBoost) para capturar resíduos complexos.

5

Monitoramento

Estabelecimento de rotina de atualização trimestral e reavaliação de quebras estruturais (regimes) pós-2025.

O roadmap visa aumentar a robustez preditiva e a utilidade prática da ferramenta para formuladores de políticas públicas.

Contribuições do Trabalho

Contribuições Técnicas

Rigor Metodológico e Estatístico

Pipeline Reprodutível

Código aberto estruturado, permitindo replicação completa da coleta de dados até a geração de previsões.

Comparação Rigorosa

Avaliação sistemática entre modelos estruturados (SARIMAX) e Machine Learning (LGBM) em séries curtas.

Validação Temporal Robusta

Utilização de estratégia *Rolling-Origin* com 13 janelas para garantir estabilidade das previsões ao longo do tempo.

Diagnóstico Completo

Verificação exaustiva de pressupostos (Ljung-Box, ARCH-LM, Normalidade) assegurando confiabilidade estatística.

Contribuições Práticas

Impacto Social e Econômico

Ferramenta Operacional

Modelo pronto para uso em monitoramento de curto prazo, oferecendo sinais rápidos sobre a direção do desemprego.

Subsídios para Políticas Públicas

Insights quantitativos que apoiam a tomada de decisão governamental e o planejamento econômico estratégico.

Alinhamento com ODS 8

Contribuição direta para a meta de promover o crescimento econômico sustentado e o trabalho decente.

Valor para a Sociedade

Redução da incerteza econômica através de projeções acessíveis e transparentes.

Agradecimentos e Referências

Agradecimentos

- FCI - Universidade Presbiteriana Mackenzie
- IBGE - PNAD Contínua
- MTE - CAGED
- BCB - SGS (Taxa Selic)

Contato

10414773@mackenzista.com.br

10414598@mackenzista.com.br

10424600@mackenzista.com.br

10416060@mackenzista.com.br

Referências Bibliográficas

BANCO CENTRAL DO BRASIL. Sistema Gerenciador de Séries Temporais – Taxa Selic. 2025. Disponível em: https://www.bcb.gov.br.

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.

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, 1979.

ENDERS, W. Applied Econometric Time Series. 4. ed. Hoboken: John Wiley & Sons, 2015.

IBGE. Pesquisa Nacional por Amostra de Domicílios Contínua – PNAD Contínua. 2025. Disponível em: https://www.ibge.gov.br.

MINISTÉRIO DO TRABALHO E EMPREGO. Cadastro Geral de Empregados e Desempregados – CAGED. 2025. Disponível em: ftp://ftp.mtps.gov.br/pdet/microdados/.