

Abstract geometric lines in the top left corner, consisting of several overlapping, irregular polygons and lines that create a complex, layered effect.

APRESENTAÇÃO– TRABALHO APLICADO

DECISÕES BASEADAS EM DADOS II

PRISCILA YUMI SASAKI MARTINS

DATA: 14/07/2025

CONTEÚDO

Entendimento do Negócio

Entendimento dos Dados

Preparação dos Dados

Modelagem

Avaliação

Implantação– Preparação

Recomendações

PLANO DE EXECUÇÃO – DEFINIÇÕES E PREMISSAS

1. Definição de escopo com orientadores:

- Objetivos;
- Restrições;
- Requisitos de precisão.

2. Aprovação de cronograma

3. Linguagem:

- Python

4. Infraestrutura física:

- GPU
- Notebook
- Servidor Remoto FGV.

1. ENTENDIMENTO DO NEGÓCIO

Objetivo:

- Modelo de previsão da inflação de bens industriais brasileira utilizando métodos de machine learning e redes neurais artificiais a partir de outros indicadores de mercado.
- Futuro desenvolvimento de um nowcasting da inflação de bens industriais

Saídas do modelo:

- Previsão da inflação de bens industriais;
- Tempo de repasse de um choque nos custos de produção;
- Indiretamente: avaliação da pressão na cadeia produtiva.

Importância:

- Adoção ou ajustes de políticas monetárias por Bancos Centrais para controlar a inflação, visando a estabilidade econômica;
- Ajuste e negociação de contratos por empresas, assim como para avaliação de investimento em aumento da capacidade produtiva e/ou instalação de novas tecnologias;
- Avaliação de aplicações financeiras por investidores, que podem antecipar suas ações de compra e venda a partir da previsão de inflação para determinado setor.

1. ENTENDIMENTO DO NEGÓCIO

Definição de Horizonte de Predição:

- 3 meses

Requisitos de Precisão:

- A série de Trabalhos para Discussão nº 561, de julho de 2022, do Banco de Central do Brasil: “*Machine Learning Methods for Inflation Forecasting in Brazil: new contenders versus classical models*”, de autoria de Gustavo Silva Araujo e Wagner Piazza Gaglianone, avaliou 50 modelos de previsão da inflação, com diferentes horizontes de previsão. Neste trabalho apresentam uma tabela de comparação de performance dos modelos, utilizando o MSE (erro quadrático médio) como métrica de avaliação de performance de modelos.
- A seguir é apresentada uma tabela resumo dos 3 melhores modelos para horizonte de predição = 1, 3 e 6 meses

1. ENTENDIMENTO DO NEGÓCIO

Horizonte	Posição	Modelo	MSE	RMSE (aprox.)
h = 1 mês	1º	Comb3 Median (Combinação)	0.07	0.27%
	2º	Comb3 GR (Combinação)	0.071	0.27%
	3º	Comb2 GR (Combinação)	0.077	0.28%
h = 3 meses	1º	BEI (Expectativa do Mercado)	0.589	0.77%
	2º	Comb2 Median (Combinação)	0.612	0.78%
	3º	Comb2 GR (Combinação)	0.615	0.78%
h = 6 meses	1º	Focus (Pesquisa Focus)	2.624	1.62%
	2º	BEI (Expectativa do Mercado)	2.825	1.68%
	3º	Random Walk (RW)	2.898	1.70%

Tabela de Comparação de Performance dos 3 Melhores Modelos para h = 1, 3 e 6 meses

Notas:

- (1) Comb3 Median: modelo de combinação de previsões, por exemplo: AR, VAR, XGBoost, Redes Neurais, etc.. Método estatístico usado para a agregação: mediana.
- (2) Comb3 GR: modelo de combinação de previsões, por exemplo: AR, VAR, XGBoost, Redes Neurais, etc.. Método estatístico usado para a agregação: mediana, porém o GR é a sigla para "Guidolin and Rossi". Refere-se ao trabalho dos pesquisadores Tullio Guidolin e Barbara Rossi, que propuseram metodologias para selecionar e combinar previsões. No contexto do artigo do Banco Central, "GR" indica que o conjunto de modelos a serem combinados foi selecionado usando o "Model Confidence Set (MCS)", uma técnica estatística popularizada por esses autores para identificar o conjunto dos melhores modelos.
- (3) BEI: Breakeven Inflation, que em português é traduzido como Inflação Implícita. Este não é um modelo de previsão no sentido tradicional (como uma rede neural), mas sim uma medida da expectativa de inflação do próprio mercado financeiro

1. ENTENDIMENTO DO NEGÓCIO

Requisitos de Precisão:

- No desenvolvimento dos modelos de previsão da inflação de bens industriais, a métrica utilizada foi o RMSE (raiz do erro quadrático médio). **Por isso, é adotado como valor de referência o RMSE da tabela anterior para um horizonte de predição de 3 meses: 0,77%.**

PLANO DE EXECUÇÃO - ETAPAS

1. Entendimento do Negócio:

- Alinhamento com interessados:
 - Objetivos;
 - Restrições;
 - Escopo;
 - Requisitos de precisão;
 - Aprovação de cronograma.
- Consulta em artigos e mídias sobre métodos em uso atualmente (benchmark)

2. Entendimento dos Dados:

- Coleta de dados históricos;
- Exploração inicial (análise de tendências, sazonalidade, estatística descritiva, faltantes, estacionariedade, quebras estruturais).

3. Preparação dos Dados:

- Limpeza de dados;
- Feature engineering: definição de defasagens, aplicação de log-linearização, variação, primeiras diferenças, criação de dummies, etc.;
- Análise de Componentes Principais.

PLANO DE EXECUÇÃO - ETAPAS

4. Modelagem:

- Definição de modelos de machine learning e de redes neurais;
- Separação de base de dados de treinamento e teste;
- Normalização da base de dados;
- Definição de métrica de performance;
- Modelagem dos dados;
- Comparação com modelo de base (modelo obtido por machine learning).

5. Avaliação:

- Avaliação de performance do modelo utilizando base de dados de teste;
- Avaliação de capacidade de explicação do modelo alinhado com conceitos econômicos (verificar importância correta de variáveis econômicas e se reflete relações econômicas esperadas);
- Teste de Stress.

6. Deployment:

- Preparar plano para disponibilização do modelo em ambiente amigável ao usuário;
- Criação de rotina de monitoramento da performance do modelo;
- Criação do ambiente para usuário;
- Treinamento de usuário. (não será realizado)

MATRIZ DE RISCOS E CONTINGÊNCIAS (EXCERTO)

Fase	Atividade	Risco-Chave		Risco Inerente				Contingência
		Cód.	Descrição	Impacto	Probabilidade	Nível de Risco		
Entendimento dos Dados	Coleta de Dados	NA	Não disponibilidade dos indicadores listados	5	5	25	Médio	Prover lista de indicadores substitutos
Entendimento dos Dados	Coleta de Dados	NA	Indicadores com frequências diferentes e/ou séries com data de início diferentes	10	5	50	Alto	Tratamento de dados na fase de Preparação de Dados
Entendimento dos Dados	Coleta de Dados	NA	Quebras estruturais passadas	8	5	40	Alto	Definição de período de tempo excluindo quebras estruturais
Modelagem	Geração do modelo	NA	Quebras estruturais	10	2	20	Médio	Redefinição de base de dados, tratamento de dados para remover efeito da quebra estrutural
Avaliação do Modelo	Avaliação dos resultados	NA	Quebras estruturais	10	2	20	Médio	Re-iniciar modelagem considerando redefinição de base de dados, tratamento de dados para remover efeito da quebra estrutural
Deployment	Disponibilização de plataforma para uso do modelo	NA	Quebras estruturais	10	2	20	Médio	Re-iniciar modelagem considerando redefinição de base de dados, tratamento de dados para remover efeito da quebra estrutural

PLANEJAMENTO DE ATIVIDADES (EXCERTO)

ATIVIDADE	INÍCIO DO PLANO	DURAÇÃO DO PLANO	INÍCIO REAL	DURAÇÃO REAL	PORCENTAGEM M CONCLUÍDA
1. Coleta de Dados e Tratamento de Base Dados	15/11/2024	30/06/2025	15/11/2024	30/06/2025	16%
2. Recuperação dados sintéticos - desagregados IPP (2010 a 2019)	16/04/2025	19/04/2025	16/04/2025	19/04/2025	100%
3. Análise de PCA	01/07/2025	08/07/2025	01/07/2025	08/07/2025	0%
4. Modelos Econométricos IPCA x IPP	20/04/2025	27/05/2025	20/04/2025	27/05/2025	60%
5. Modelos Econométricos IPCA x variáveis PCA	09/07/2025	05/08/2025	09/07/2025	05/08/2025	0%
Análise e Testes	09/07/2025	24/07/2025	09/07/2025	24/07/2025	0%
Modelo VAR (com HAC) com análise resultados	26/07/2025	05/08/2025	26/07/2025	05/08/2025	0%
6. Modelos Machine Learning IPCA x IPP	15/11/2024	16/05/2025	15/11/2024	16/05/2025	100%
Modelagem métodos diversos - base de dados observ. a partir de 2019	15/11/2024	15/12/2024	15/11/2024	15/12/2024	100%
Modelagem métodos diversos - base dados observ + dados sintéticos (2010 - 19)	12/05/2025	16/05/2025	12/05/2025	16/05/2025	100%
7. Modelos Machine Learning IPCA x variáveis PCA	05/07/2025	30/07/2025	05/07/2025	30/07/2025	0%
8. Avaliação Modelos VAR x Machine Learning	05/08/2025	20/08/2025	05/08/2025	20/08/2025	0%

2. ENTENDIMENTO DOS DADOS

1. Carregamento dos Dados no Python por:

- Arquivos em csv ou xlsx;
- Através de API (Sidra IBGE);

2. Análise dos Dados:

- Conversão de dados de string para float e datetime;
- Análise de comportamento das séries temporais – identificação de possíveis erros de dados;
- Remoção de linhas com dados faltantes;
- Teste de Estacionariedade;
- Decomposição das séries em tendência, sazonalidade e resíduos;
- Análise dos gráficos de autocorrelação e autocorrelação parcial;
- Algumas séries não apresentavam estacionariedade por se tratarem de números índices ou números brutos. Neste caso, foram incluídos os dados de variação (primeiras diferenças)

3. Desenvolvimento dos Códigos:

- Para cada um dos indicadores, exceto para ICI, NUCI e Nível de Estoques, foi criado um código separadamente chamado “Code_GetData_’nomedoindicador’”. Neste código são desenvolvidas as etapas (1) e (2) acima;
- Essa estratégia foi necessária devido à grande diversidade dos dados, do formato de disponibilização destes e fontes;
- Após todo o processo de tratamento dos dados, é gerado um arquivo csv;
- Finalmente, todos esses arquivos csv’s são compilados através de código num único arquivo csv para ser utilizado para o desenvolvimento dos modelos de previsão.

2. ENTENDIMENTO DOS DADOS

VARIÁVEL TARGET – INFLAÇÃO DE BENS INDUSTRIAIS

- Obtida através de proxy a partir do IPCA ou IPC

VARIÁVEIS EXPLICATIVAS – LISTA COMPLETA SEM TRATAMENTO DE PCA*

- IPCA – IBGE
- IPA – FGV
- IPC – FGV (possível substituto do IPCA)
- ICI – Índice Confiança da Indústria - FGV
- NUCI - Nível Utilização Capacidade Instalada (ICI) - FGV
- Nível Estoques (ICI) - FGV
- PIM-PF.T - Pesquisa Industrial – Produção da Indústria de Transformação - IBGE
- IBC-Br – Índice Atividade Econômica - BC
- IAE-Br – Monitor Atividade Econômica – FGV (possível substituto do IBC-Br)
- IPP – IBGE (possível substituto para o IPA)
- IC-Br – Índice de Commodities - BC
- Taxa Câmbio Efetiva Real – IPEA (já está contemplado no IC-Br)
- Consumo Aparente – Indústria de Transformação - IPEA
- Indicador IPEA de formação bruta de capital fixo – IPEA (possível substituto complementar ao ICI)
- ICEI - Índice Confiança Empresário Industrial Geral – CNI (possível substituto do ICI)
- Consumo Energia Elétrica – Indústria – Eletrobras
- Monitor da Inflação – FGV (necessário para desenvolvimento do nowcasting)

DADOS DE PAINEL:

- Matriz Insumo Produto - IBGE

2. ENTENDIMENTO DOS DADOS

Disponibilidade dos Dados - Fontes:

- IPEA (<http://www.ipeadata.gov.br/Default.aspx>) - arquivos csv
- API Sidra IBGE (<https://sidra.ibge.gov.br>) - códigos em Python para download dos dados através de api
- IBRE – FGV (através de orientador) - arquivos csv ou xlsx
- IBGE (<https://www.ibge.gov.br/pt/inicio.html>) - relatórios em pdf e arquivos em xls
- Banco Central (<https://www.bcb.gov.br/estatisticas/indicadoresselecionados>) - arquivos xlsx

2. ENTENDIMENTO DOS DADOS

	Arquivo	Entrada	Fonte	Arquivo de Saída	Observação
Code_GetData_	IPA	excel	Orientador	base_dados_ipa_desag.csv	Recuperados dados das desagregações "Produtos Alimentícios" e "Produtos Químicos" de jan/05 a dez/07 através de método ML.
Code_GetData_	PIM-PF	API	Sidra IBGE	base_dados_pimpf.csv	
Code_GetData_	ConsApar	csv	IPEA	base_dados_cons_apar.csv	
Code_GetData_	ConsEE	csv	IPEA	base_dados_cons_EE.csv	
Code_GetData_	IBC	csv	IPEA	base_dados_ibc.csv	
Code_GetData_	IC	excel	BC	base_dados_ic.csv	
Code_GetData_	ICI	excel	Orientador	base_dados_ici_ind.csv"	Base de dados contém as desagregações NUCI e Nível de Estoques
Code_GetData_	IPCA	csv	BC	IPCA_Var Mensal_BC/ipca_desag_var_mensal_mai o25.xlsx	
Code_GetData_	BaseCompleta	csv	arquivos acima	Dados/Base Completa/base_dados_2005.csv	

Tabela Resumo de Fonte de Dados, arquivos Code_GetData e arquivos de saída para uso em desenvolvimento de modelos

DADOS – PROXY CUSTO DE PRODUÇÃO INDÚSTRIA TRANSFORMAÇÃO

Para o cálculo da variação mensal da proxy do custo de produção da indústria de transformação, os coeficientes técnicos de transferências entre setores da indústria de transformação da tabela 14 da Matriz Insumo Produto do IBGE (2015) são utilizados como pesos. Cálculo feito no Python:

```
#Cálculo da Proxy do Custo da Industria de Transformação considerando os coeficientes técnicos da matriz insumo produto do IBGE 2015

df_proxy_custo_ind_transf = (0.22053*df_ipa_final["Produtos Alimenticios"] +
0.02516*df_ipa_final["Bebidas"] +
0.00522*df_ipa_final["Fumo Processado e Produtos do Fumo"]+
0.01523*df_ipa_final["Produtos Texteis"]+
0.01633*df_ipa_final["Artigos de Vestuario"]+
0.01249*df_ipa_final["Couros, Artigos para Viagem e Calçados"]+
0.00775*df_ipa_final["Madeira Desdobrada e Produtos de Madeira"]+
0.03218*df_ipa_final["Celulose, Papel e Produtos de Papel"]+
0.16493*df_ipa_final["Produtos Derivados do Petroleo e Biocombustiveis"]+
0.09938*df_ipa_final["Produtos Quimicos"]+
0.01640*df_ipa_final["Produtos Farmaceuticos"]+
0.03506*df_ipa_final["Artigos de Borracha e de Material Plastico"]+
0.02915*df_ipa_final["Produtos de Minerais Nao-Metalicos"]+
0.05898*df_ipa_final["Metalurgia Basica"]+
0.02659*df_ipa_final["Produtos de Metal"]+
0.03322*df_ipa_final["Equipamentos de Informatica, Produtos Eletronicos e Opticos"]+
0.02579*df_ipa_final["Maquinas, Aparelhos e Materiais Eletricos"]+
0.03862*df_ipa_final["Maquinas e Equipamentos"]+
0.08191*df_ipa_final["Veiculos Automotores, Reboques, Carrocerias e Autopecas"]+
0.01828*df_ipa_final["Embarcacoes, Veiculos Ferroviarios e Outros Equipamentos de Transporte"]+
0.01815*df_ipa_final["Moveis"]+
0.01865*df_ipa_final["IPCA Serv Var Mensal (%)"])
```


3. PREPARAÇÃO DOS DADOS

- Seleção e inclusão de dados das variáveis de interesse no mesmo dataframe – tarefa realizada no final da etapa anterior;
- Definição de data de início das séries a serem utilizadas;
- Criação de Dummies:
 - Período de Regime de Exceção – período de março/2020 a dezembro/2022 contemplando pandemia e 1º ano da Guerra da Rússia / Ucrânia;
 - Dummies para cada mês do ano - necessárias para Modelos de Machine Learning, como o Random Forest, que não "enxergam" a sazonalidade apenas pelo índice de data.
- Criação de 4 defasagens para todas as variáveis explicativas e da variável target (modelo deve considerar a endogeneidade da série);
- Análises exploratórias;
- Análise de Componentes Principais: redução de dimensionalidade da base de dados (análise de dados exploratória, visualização e pré-processamento de dados) – base de dados mais refinada para modelagem econométrica e modelos de machine learning / redes neurais;
- Separação da base de dados de treinamento e da de teste (80% para treinamento);
- Normalização de dados para ficarem no mesmo intervalo [0,1]

3. PREPARAÇÃO DOS DADOS

- A aplicação do PCA foi removida dos códigos para desenvolvimento dos modelos de previsão.
- Inicialmente os códigos foram desenvolvidos incluindo a transformação dos dados utilizando o PCA, porém removidos ao longo do processo devido à presença de possível overfitting nos modelos.
- Após uma extensa fase de experimentação, que incluiu a validação cruzada de múltiplas arquiteturas de redes neurais (SimpleRNN, LSTM, GRU, incluindo modelos empilhados) e o ajuste fino de hiperparâmetros (como dropout, batch_size e window_size), foram observados os seguintes resultados:
 - Bom Desempenho na Validação, Ruim no Teste: Os modelos treinados com os dados transformados pelo PCA apresentaram um bom desempenho durante a validação cruzada (com o melhor modelo, Stacked_GRU_64x64, atingindo um RMSE médio de ~0.26). No entanto, ao avaliar este mesmo modelo no conjunto de teste (dados nunca vistos), o desempenho caiu drasticamente, resultando em um RMSE de 0.45.
 - Sinal de Overfitting e Falha de Generalização: A grande discrepância entre o erro de validação e o erro de teste é um sintoma clássico de que o modelo não está generalizando bem. Ele aprendeu os padrões específicos dos componentes principais do conjunto de treino, mas esses padrões não se mostraram robustos o suficiente para o conjunto de teste.
 - Instabilidade nos Resultados: Múltiplas execuções do mesmo script de validação cruzada produziram resultados ligeiramente diferentes, com diferentes arquiteturas sendo coroadas "campeãs" a cada vez. Isso sugere que a performance dos modelos era muito sensível às condições iniciais quando treinados com os componentes do PCA.
- Modelos como o Random Forest e Redes Neurais são naturalmente bons em lidar com as situações que o PCA tenta resolver, tais como as listadas abaixo
 - Grande Número de Features: O Random Forest foi projetado para lidar com muitas colunas. Ele tem um mecanismo interno de seleção de features (em cada divisão de cada árvore, ele escolhe a feature mais importante), então ele consegue "ignorar" as variáveis menos relevantes.
 - Features Correlacionadas: Ele também é muito mais robusto à multicolinearidade (features correlacionadas entre si) do que modelos de regressão linear, por exemplo.
 - Vantagem Extra: Interpretabilidade - ao não usar o PCA, há a vantagem da interpretabilidade. Por exemplo, no final do treinamento, pode-se usar o Random Forest para informar quais das features originais (como 'Nível de estoques', 'regime_excecao' ou 'IPCA_lag_2') foram as mais importantes para a previsão. Isso já não seria possível de se fazer com os componentes abstratos do PCA.

4. MODELAGEM

- Separação da base de dados de treinamento e de teste realizada na etapa anterior.
- Definido horizonte de predição = 3 meses.
- Foram desenvolvidos 2 códigos:
 - Pred_IPCA_BaseCompleta_sempca_dummy_ann_rev1:
 - Testadas diversas configurações de hiperparâmetros para redes do tipo Simple RNN, LST, GRU e Stacked;
 - Pred_IPCA_BaseCompleta_machinelearning:
 - Testadas diversas configurações de hiperparâmetros para modelos de Machine Learning: Random Forest, AdaBoost, Gradient Boosting e XG Boost.
- Métrica para avaliação de performance de modelos: RMSE.
- Após a definição das configurações com melhor performance utilizando redes neurais e utilizando ML, os modelos de treinamento são gerados para serem testados na próxima etapa utilizando a base de dados de testes.

4. MODELAGEM

Modelos com Melhor Performance:

- Redes Neurais Artificiais: Modelo: GRU_256_base_d0.2_b16
 - Tipo de Camada: GRU (Gated Recurrent Unit)
 - Unidades na Camada: 256
 - Taxa de Dropout: 0.2
 - Tamanho do Lote (Batch Size): 16
 - RMSE na Validação Cruzada: 0.24559 % - variação mensal
- Machine Learning: XGBoost Regressor
 - n_estimators: 120
 - learning_rate: 0.05
 - max_depth: 4
 - subsample: 0.8 (Usa 80% dos dados por árvore)
 - colsample_bytree: 0.4 (Usa 40% das features por árvore)
 - gamma: 0
 - RMSE na Validação Cruzada: 0.2559 % - variação mensal

5. AVALIAÇÃO

- Com os dados da base de teste, são utilizados os modelos gerados para prever os valores da variável target (IPCA de bens industriais). Estes valores são, então, comparados com os valores reais da base de teste.
- Observou-se que ocorre um aumento significativo do RMSE na base de teste em ambos os modelos

Modelo	RMSE %	
	Validação	Teste
Redes Neurais - Modelo: GRU_256_base_d0.2_b16	0,25	0,38
Machine Learning - XGBoost Regressor	0,26	0,42

5. AVALIAÇÃO

- O modelo foi treinado majoritariamente com dados de um período (2005-2020) que, apesar de suas próprias crises, tinha uma certa dinâmica. O conjunto de teste (2021-2024), por outro lado, foi dominado por um regime completamente novo e de altíssima volatilidade, influenciado pelos efeitos da pandemia e da guerra na Ucrânia. O modelo aprendeu os padrões do "mundo antigo" e teve dificuldade em aplicá-los perfeitamente ao "mundo novo".
- Existe uma profunda quebra estrutural na série da inflação de bens industriais a partir de 2020. Os padrões e as relações entre as variáveis no período pré-pandemia não são mais suficientes para prever com acurácia o período durante pandemia.
- A análise dos resultados dos modelos indica que há uma pior performance no período de exceção. A exclusão dos dados do período de exceção pode ser a chave para se entender o verdadeiro potencial do modelo em "tempos normais".
 - Ao fazer isso, há a separação efetiva da avaliação em duas partes:
 - Qual a performance do modelo em "tempos normais"?
 - Qual o impacto da crise na performance?

5. AVALIAÇÃO

- O desafio na previsão, portanto, não advém de uma limitação de um tipo de modelo específico (redes neurais vs. árvores), mas sim uma característica dos dados: o período de teste (2021-2024) representa um regime econômico muito diferente daquele em que os modelos foram predominantemente treinados (2005-2020).
- Foi realizada a mesma análise, separando o período de exceção do período “normal” para comparar os dois modelos e entender o verdadeiro potencial de cada um.

Métrica	Melhor Rede Neural (GRU)	Melhor Machine Learning (XGBoost)	Melhor Modelo
RMSE (Validação Cruzada) %	~0.246	0.256	Rede Neural - GRU
RMSE (Teste - Completo) %	0.38	0.42	Rede Neural - GRU
RMSE (Teste - Período de Exceção) %	0.43	0.53	Rede Neural - GRU
RMSE (Teste - Período Normal) %	0.33	0.28	ML - XGBoost

- Comparando com o requisito de precisão definido na 1ª Etapa do Projeto para um horizonte de predição de 3 meses (0,77), os modelos apresentaram performance superior ao requisito.

5. AVALIAÇÃO

- A superioridade do XGBoost no teste do "período normal" pode ser atribuída a algumas hipóteses relacionadas à natureza dos modelos e dos dados:
- Robustez a Mudanças de Regime (Generalização)
 - Rede Neural (GRU): Redes neurais, especialmente as com alta capacidade como a GRU(256), são "máquinas de aprendizado" extremamente potentes, aprendendo padrões complexos e não-lineares nos dados de treinamento. A desvantagem é que elas podem ser mais suscetíveis ao overfitting: é provável que o GRU tenha se "especializado" demais nos padrões do período de treino (2005-2020), e quando confrontado com o "novo normal" de 2023-2024, sua performance se degradou.
 - XGBoost: Modelos baseados em árvores, como o XGBoost, aprendem através de uma série de regras de divisão. Essas regras, especialmente com a forte regularização aplicada podem ser mais simples e robustas. É possível que as regras aprendidas pelo XGBoost fossem mais gerais e, portanto, mais resistentes à mudança de dinâmica observada no conjunto de teste.
- Natureza dos Dados: Tabular vs. Sequencial:
 - XGBoost: É um algoritmo especialista em dados tabulares. O processo de engenharia de features transformou a série temporal em uma grande tabela, com dezenas de colunas. O XGBoost é mestre em explorar as interações entre essas colunas de forma explícita.
 - Rede Neural (GRU): Embora o GRU seja excelente com sequências, foi alimentado com uma "sequência de tabelas" (vetores de features a cada passo de tempo). Ele aprendeu a dinâmica sequencial desses vetores, mas talvez não tenha sido tão eficaz quanto o XGBoost em explorar as relações dentro de cada vetor de features.
- Conclusão: A melhor abordagem depende da natureza dos dados e do problema. Enquanto a Rede Neural GRU mostrou um potencial de aprendizado altíssimo na fase de validação, o XGBoost provou ser o modelo mais robusto e confiável quando confrontado com a tarefa mais difícil: generalizar para um futuro com dinâmicas diferentes das do passado

6. IMPLANTAÇÃO – ETAPA SUBSEQUENTE

- Qualquer um dos modelos (redes neurais ou ML) ainda não está apto para passar para a fase de deployment - fase em que o modelo é integrado aos processos de negócio. Para tanto, as próximas etapas são:
- Etapa 1: Empacotamento e Criação de API
 - Objetivo: Transformar o modelo de um script de análise em um serviço de previsão funcional e reutilizável.
 - Ações de Empacotamento do Modelo (Model Packaging):
 - Centralizar a Lógica: Criar um script Python (predict.py) com uma única função que recebe os dados de entrada e executa todo o pipeline:
 - Carregar o modelo salvo e os scalers.
 - Realizar a mesma engenharia de features (lags, dummies).
 - Normalizar os dados, fazer a previsão e reverter a previsão para a escala original.
 - Gerenciar Dependências: Criar um arquivo requirements.txt para listar todas as bibliotecas necessárias, garantindo que o ambiente seja 100% reproduzível.
 - Exposição do Modelo via API (Interface):
 - Desenvolver uma API REST: Usar um framework como Flask ou FastAPI para criar um "serviço" que outros sistemas possam chamar.
 - Criar um Endpoint de Previsão (/predict): Este endereço recebe os novos dados e retorna a previsão do modelo, tornando-o acessível para outras aplicações.

6. IMPLANTAÇÃO – ETAPA SUBSEQUENTE

- Etapa 2: Containerização, Implantação e Ciclo de Vida (MLOps)
 - Objetivo: Garantir que o serviço de previsão seja portátil, escalável e sustentável em um ambiente de produção.
 - Containerização com Docker:
 - Criar uma "Caixa" para a Aplicação: Desenvolver um Dockerfile que define um ambiente isolado e padronizado para a API.
 - Garantir Portabilidade: O contêiner empacota a aplicação, as dependências e os artefatos do modelo, permitindo que ele rode de forma idêntica em qualquer lugar (localmente ou na nuvem).
 - Implantação na Nuvem (Cloud Deployment):
 - Hospedar o Serviço: Fazer o deploy do contêiner em uma plataforma de nuvem como AWS, Google Cloud Platform (GCP) ou Azure.
 - Serviços Gerenciados: Utilizar serviços como Google Cloud Run ou AWS App Runner para gerenciar a infraestrutura e a escalabilidade automaticamente.
 - Manutenção e Monitoramento (MLOps):
 - Monitorar a Performance: Acompanhar a precisão do modelo ao longo do tempo para detectar o decaimento de performance (model drift).
 - Planejar o Retreinamento: Definir uma estratégia para retreinar o modelo periodicamente com novos dados, garantindo que ele se mantenha relevante e preciso.



OBRIGADA

Priscila Yumi Sasaki Martins

<https://www.linkedin.com/in/priscila-martins-90773215/>

priscila.sasaki@gmail.com