Pós-graduação Lato Sensu em Ciência de Dados e Big Data - 2021

Desempenho de modelos de *gradient boosting* para predição de inflação mensal

Pedro Gabriel Lima de Andrade

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Contextualização

REPORTAGEM DE CAPA

O Departamento de Pesquisa do Banco Central emprega um modelo econômico tropical e um batalhão de Ph.Ds para fazer suas estimativas econômicas. Por Alex Ribeiro, de Brasília

O Samba dos **juros**

Quando o sistema de metas de inflação foi adotado, há mais de uma década, o Banco Central abriu um concurso público para selecionar um time de economistas de primeira linha para o seu recém-criado departamento de pesquisas econômicas. Foi um grande fiasco: das 30 vagas em disputa, apenas uma foi preenchida. Todos os demais candidatos foram reprovados.

De lá para cá, outros concursos tiveram mais êxito em atrair talentos, e muitos dos funcionários da casa cursaram pós-graduação em centros de excelência. Assim, o Departamento de Pesquisa Econômica (Depep), o cérebro do Banco Central, firmou-se como uma referência na produção de conhecimento sobre política monetária, finanças e economia bancária no Brasil.

Um dos marcos é o desenvolvimento de um modelo de projeção econômica de última geração, batizado como Samba, que usa técnicas da chamada "economia artificial" e que coloca o Banco Central do Brasil no primeiro pelotão entre países emergentes.

"Nos primeiros anos do regime de meta de inflação, os modelos eram bem básicos", afirma o professor Fábio Kanczuk, da Universidade de São Paulo (USP), especialista em estudo de política monetária com formação em economia pela Universidade da Califórnia, em Los Angeles, e por Harvard. "Com o Samba, o Depep deu um salto. Não deve nada aos seus pares."

Nesses anos, o Depep subiu ao poder no Banco Central. Seu primeiro chefe, o economista Alexandre Tombini, é hoje o presidente da instituição. De seu s quadros também saiu o diretor de política econômica, Carlos Hamilton de Araújo, doutor pela Escola de Pós-Graduação em Economia (EPGE) da Fundação Getulio Vargas (FGV). Pela primeira vez, houve o reconhecimento de que um funcionário do próprio Banco Central estava preparado para assumir a cadeira central na gestão do regime de metas de inflação, depois de uma linhagem de economistas vindos de fora que inclui Sérgio Werlang (Princeton), Ilan Goldfajn (MIT), Afonso Bevilaqua (Berkeley) e Mário Mesquita (Oxford).

O Depep ganhou o reconhecimento até de um dos maiores críticos da abordagem excessivamente científica da economia — o ex-ministro da Fazenda, do Planejamento e da Agricultura Antonio Delfim Netto. "Nem nossos mais sofisticados economistas [do mercado] ou da academia podem competir com as informações armazenadas nas cabeças dos profissionais que habitam o Departamento de Estudos e Pesquisa (Depep) do Banco Central", escreveu Delfim, em artigo recente no **Valor**.

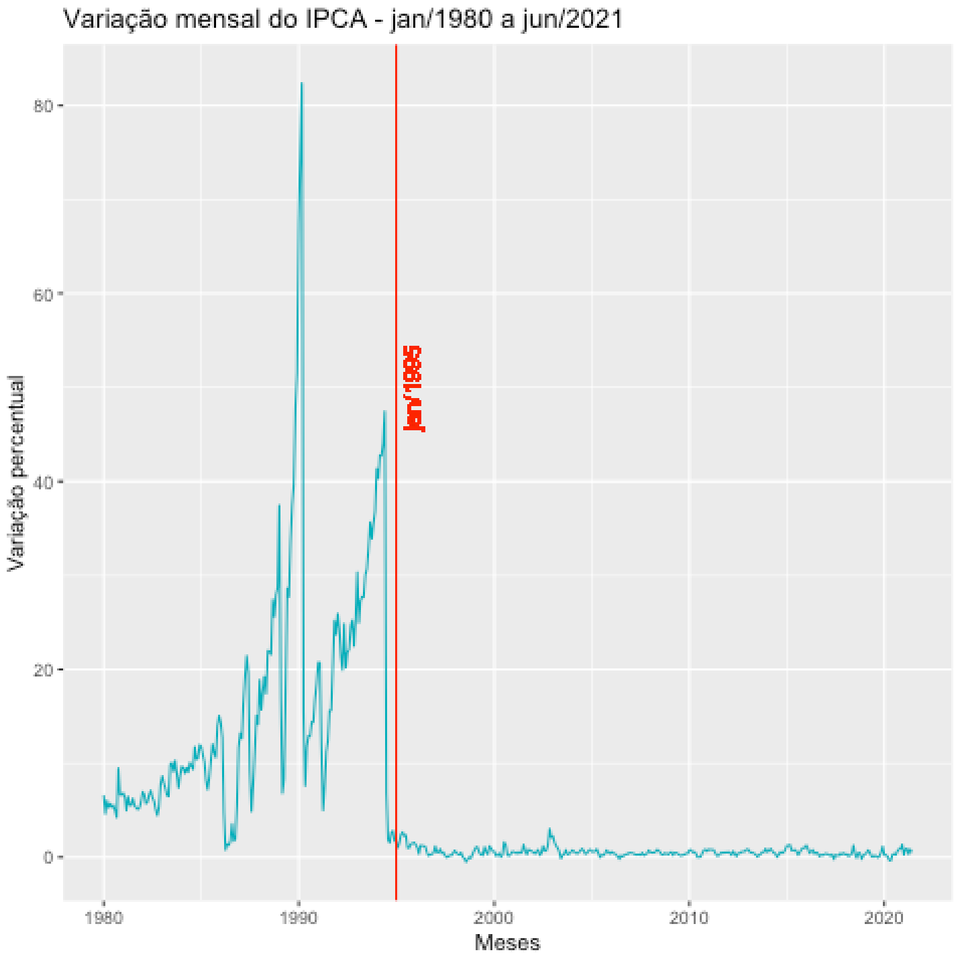
Neste exato momento, as projeções de in-

PUC Minas

Contextualização

o

- O Brasil conviveu durante décadas com inflação alta, o que trazia diversos problemas, entre eles a pouca previsibilidade de variáveis macroeconômicas
- Desafio, a partir do plano real em 1994, é ter previsão mais assertiva dessas variáveis, principalmente da inflação;
- Empresas, governos e sociedade em geral, precisam de previsões de inflação para poder planejar orçamentos - comportamento denominado **ancoragem de expectativas**.



PUC Minas

Problema proposto

O objetivo central deste trabalho é identificar **a performance dos melhores modelos de gradient boosting** na previsão da inflação mensal do Brasil.

Os objetivos secundários são:

- Criar um **scrapping** para coletar os dados necessários a partir das APIs do Banco Central;
- Demonstrar os **métodos de feature engineering** adequados para os modelos;
- Criar os **modelos de gradient boosting** ajustando hiperparâmetros através da otimização bayesiana;
- Obter os **resultados de predição do IPCA** de até 4 meses posteriores com a métrica do erro quadrático médio;
- Analisar os **resultados** com a biblioteca SHAP;

Problema proposto

- A **escolha dos modelos de gradient boosting** para esse problema se deve ao fato que esses modelos são o estado da arte para solução de problemas que envolvem dados tabulares;
- Esses modelos conciliam de maneira satisfatória **performance e tempo de processamento**;

Stack de **ferramentas e técnicas** utilizada:

- Linguagem python para coleta e modelagem;
- Linguagem R para visualização dos dados;
- SQLite para armazenar os dados coletados no *scrapping;*
- Otimização bayesiana para otimizar os hiperparâmetros dos modelos;
- Catboost, XGBoost e LightGBM para modelos de *gradient boosting*;
- Biblioteca SHAP para extrair as features mais importantes;

PUC Minas

Coleta dos dados

Estudo que embasou a coleta inicial dos dados:

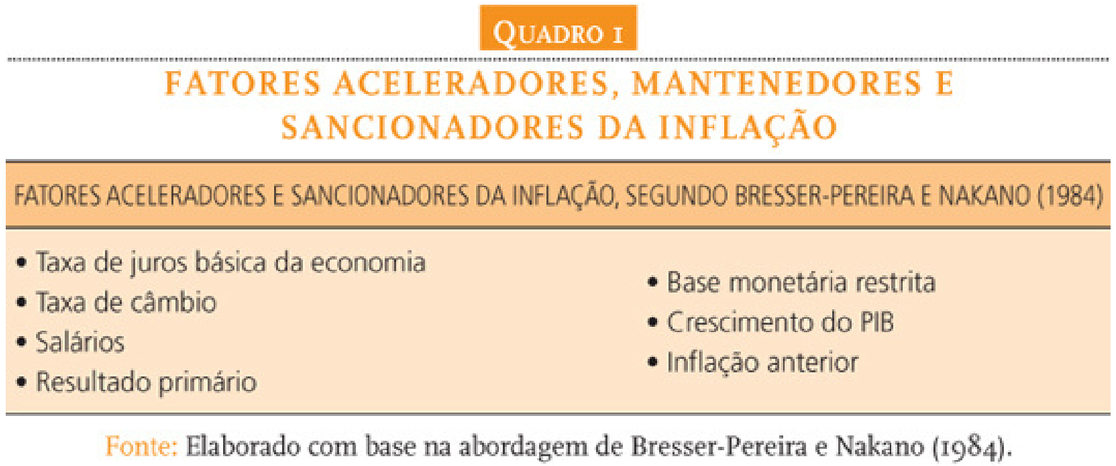
Finanças Estratégicas • RAM, Rev. Adm. Mackenzie 13 (1) • Fev 2012 • https://doi.org/10.1590/S1678-69712012000100004 COPIAR

Redes neurais artificiais na previsão da inflação: aplicação como ferramenta de apoio à análise de decisões financeiras em organizações de pequeno porte

Redes neuronales artificiales en el pronóstico de la inflación: la aplicación como una herramienta para apoyar el análisis de las decisiones financieras en organizaciones pequeñas

Artificial neural networks in inflation prediction: application like analysis tool for financial decisions at small organizations

Leonardo Augusto Amaral Terra João Luiz Passador SOBRE OS AUTORES



QUADRO 1

FATORES ACELERADORES, MANTENEDORES E SANCIONADORES DA INFLAÇÃO

FATORES ACELERADORES E SANCIONADORES DA INFLAÇÃO, SEGUNDO BRESSER-PEREIRA E NAKANO (1984)	
• Taxa de juros básica da economia	• Base monetária restrita
• Taxa de câmbio	• Crescimento do PIB
• Salários	• Inflação anterior
• Resultado primário	

Fonte: Elaborado com base na abordagem de Bresser-Pereira e Nakano (1984).

PUC Minas

Coleta dos dados

- Os dados foram coletados em **duas bases disponíveis pelo Banco Central:** o Sistema de Gerenciamento de Séries Temporais e o Sistema de Expectativas de Inflação.
- Os dados históricos escolhidos foram majoritariamente de índice de preços diversos, mas também teve dados de finanças públicas (NFSP), dados de base monetária, atividade econômica e entre outros;
- Os **dados históricos** são tanto de séries mensais quanto diárias - para esta foi colhida a média do mês;
- Os **dados de expectativas de mercado** para as inflações dos meses posteriores foram para seguintes índices de inflação: IPCA, INPC e IGP-M;
- O script de carga de dados históricos consta no repositório com a função "carga_dados_historicos(mes, ano, n_serie_m, n_serie_d)"

Coleta dos dados

- O nome e código das séries temporais históricas constam no seguinte dicionário:

```
num_series_bacen_m = {"index_exp_futuras": 4395, "index_confianca": 4393,
"index_cond_econ_atuais": 4394, "ibc_br": 24363, "nfsp_rp": 4649, "ipca_comerc": 4447,"ipca_nao_comerc": 4448,
"ipca_itens_livres": 11428, "ipca_servicos": 10844, "ipca_duraveis": 10843, "ipca_bens_semidur": 10842,
"ipca_nao_duraveis": 10841, "inpc_index_dif": 21379, "inpc_nucleo_suav": 4466, "igpm": 189, "igpm_di": 190,
"ipc_br": 191, "incc": 192, "ipa": 225, "ipc_nucleo": 4467, "igp10": 7447, "igpm_1decendio": 7448, "igpm_2decendio": 7449,
"ipam": 7450, "ipam_1decendio": 7451, "ipam_2decendio": 7452, "ipcm": 7453, "ipcm_1decendio": 7454, "ipcm_2decendio": 7455,
"incc": 7456, "incc_1decendio": 7457, "incc_2decendio": 7458, "ipaog_prod_indus": 7459, "ipaog_prod_agro": 7460,
"inpc": 188, "ipc_fipe": 193, "ipc_fipe_2quadrisemana": 272, "ipca": 433, "ipca_alim_beb": 1635, "ipca_habit": 1636,
"ipca_art_habit": 1637, "ipca_vestuario": 1638, "ipca_transportes": 1639, "ipca_comunicacao": 1640, "ipca_saude": 1641,
"ipca_desp_pes": 1642, "ipca_educacao": 1643, "inpc_alim_beb": 1644, "inpc_habit": 1645, "inpc_art_habit": 1646,
"inpc_vestuario": 1647, "inpc_transporte": 1648, "inpc_comunicacao": 1649, "inpc_saude": 1650, "inpc_desp_pes": 1651,
"inpc_educacao": 1652, "ipca_monitorados": 4449, "ipc_fipe_1quadrisemana": 7463, "ipc_fipe_3quadrisemana": 7464,
"ipc_fipe_aliment": 7465, "ipc_fipe_indust": 7467, "ipc_fipe_innatura": 7468, "ipc_fipe_habit": 7469,
"ipc_fipe_transp": 7470, "ipc_fipe_desp_pes": 7471, "ipc_fipe_vest": 7472, "ipc_fipe_saude": 7473,
"ipc_fipe_educacao": 7474, "ipc_fipe_comerc": 7475, "ipc_fipe_nao_comerc": 7476, "ipc_fipe_monit": 7477,
"ipca15": 7478, "ipca_e": 10764, "ipca_12meses": 13522, "ipca_industriais": 27863, "ipca_alim_dom": 27864,
"custo_cesta_aracaju": 7479, "custo_cesta_belem": 7480, "custo_cesta_bh": 7481, "custo_cesta_brasilia": 7482,
"custo_cesta_curitiba": 7483, "custo_cesta_floripa": 7484, "custo_cesta_fortaleza": 7485, "custo_cesta_goiania": 7486,
"custo_cesta_jp": 7487, "custo_cesta_natal": 7488, "custo_cesta_poa": 7489, "custo_cesta_recife": 7490,
"custo_cesta_rj": 7491, "custo_cesta_salvador": 7492, "custo_cesta_sp": 7493, "custo_cesta_vitoria": 7494,
"pib_acum_12meses": 4192, "pib_mensal": 4380, "pib_acum_ult12meses": 4382}

num_series_bacen_d = {"selic_diaria": 11, "selic_acum": 1178, "meta_selic": 432}
```

PUC Minas

Coleta dos dados

- Para os dados de **expectativas de mercado**, coletou-se a expectativa de mercado para os seguintes índices de preços: IPCA, IGP-M e INPC;
- O horizonte de expectativas foi de 5 meses, ou seja, coletou-se os dados de expectativas dos agentes econômicos para os 5 meses posteriores;
- Das expectativas, coletou-se a média, mediana e desvio padrão do mercado;
- No entanto, os dados só são disponíveis a partir de 2000, ano que se criou o Sistema de Expectativas;
- O script de carga dos dados constam no repositório com o nome da função "carga_dados_expectativas(mes, ano)"

PUC Minas

Modelagem

- Para testagem de performance dos algortimos gradient boosting, escolheu-se os três algoritmos de **regressão** *state of art* da família de *decision tree*: **CatBoost, XGBoost e LightGBM**;
- XGBoost foi o primeiro e talvez o mais conhecido;
- O objetivo do LightGBM é acelerar o treinamento se comparado com XGBoost;
- O CatBoost, por sua vez, tem dois grandes objetivos: evitar o *overfitting* e fornecer bons hiperparâmetros padrão;
- Estes algoritmos possuem diferenças na amostragem e na forma de fazer os *ensembles* ao longo da árvore de decisão;

XGBoost CatBoost LightGBM

PUC Minas

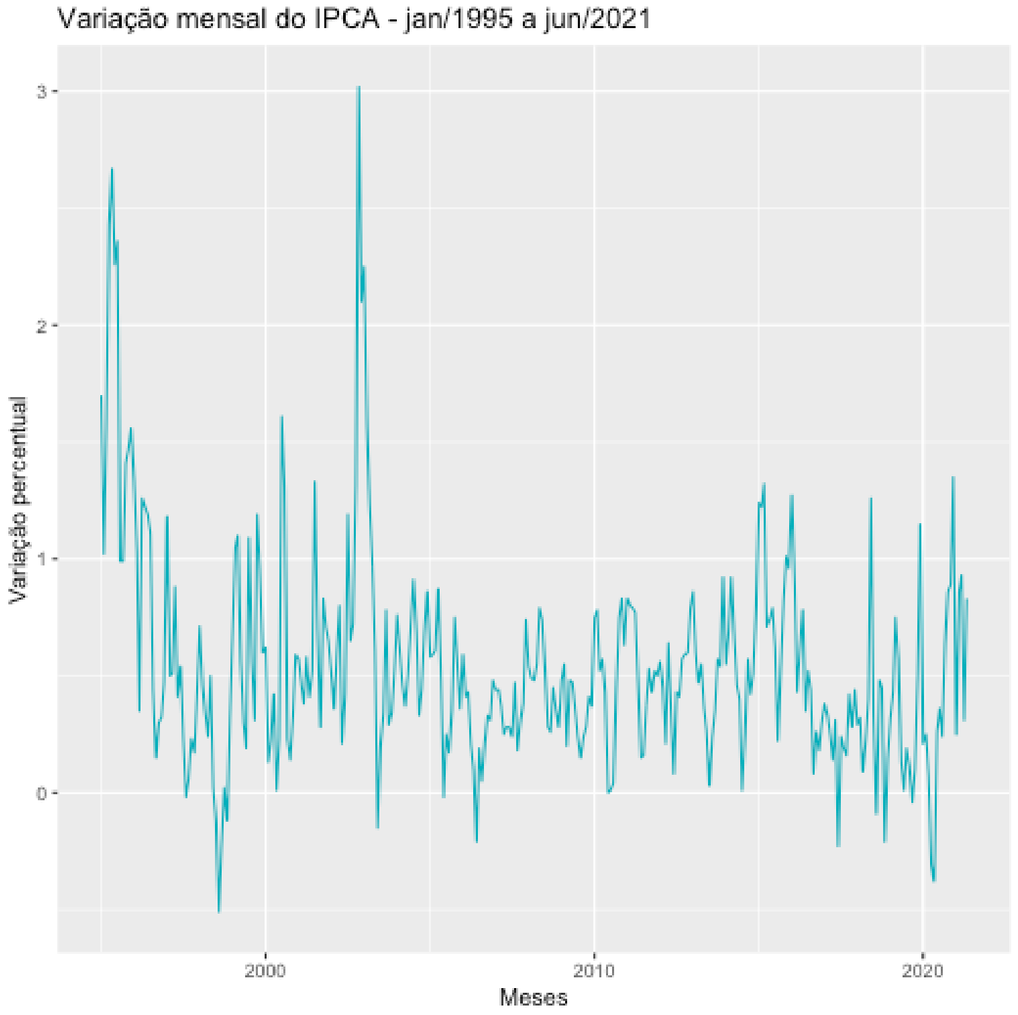
Modelagem

- O **processo de feature engineering** foi relativamente simples, usando valores exorbitantes (-9999) para dados faltantes e dados do mês passado para principais séries
- Foram feitos **testes** com diversas formas de tratar melhor os dados tabulares de séries temporais, como média móvel e diferenças, e não lograram êxito;
- Os **dados de treino e teste** foram dividido por critério do tempo. Dados de janeiro de 1995 até dezembro de 2014 foram usados para teste, e dados de janeiro de 2015 até junho de 2021 foram usados para teste;
- Outras técnicas de divisão da base de dados não foram utilizadas devido à baixa dimensionalidade dos dados;

PUC Minas

Modelagem

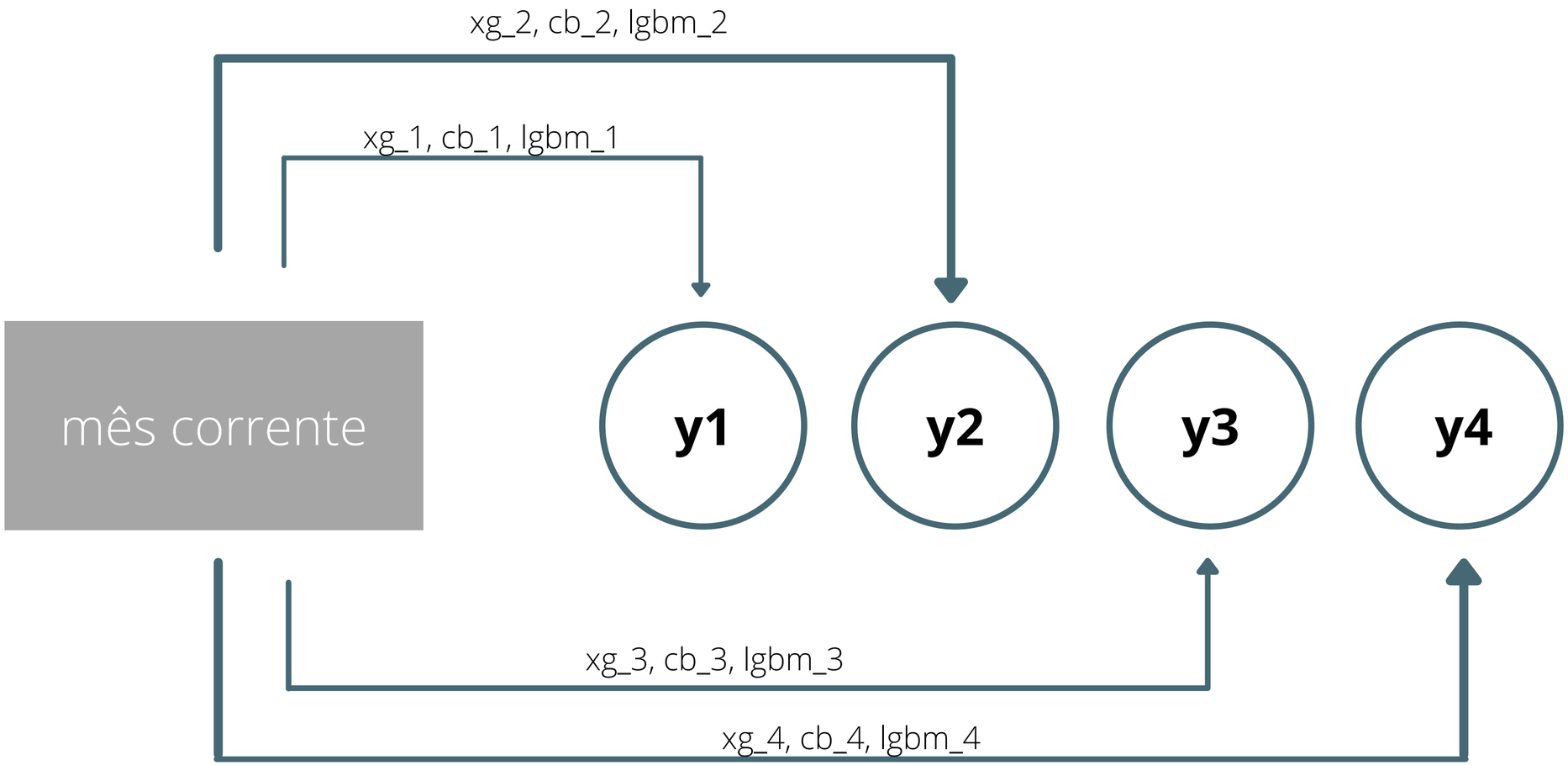
- A série temporal do IPCA, que é o principal índice de inflação, consta
- Para se criar o target dos modelos, utilizou-se da função "shift" da biblioteca pandas;
- O target possui o nome de y1, y2, y3 e y4;
- Com isso, criou-se para cada mês posterior um modelo, de cada algoritmo de gradient boosting diferente.



PUC Minas

Modelagem

A **"arquitetura de modelagem"** deste problema foi concebida da seguinte forma:



PUC Minas

Modelagem

- Para otimizar os hiperparâmetros dos modelos, utilizou-se da abordagem de **otimização bayesiana**;
- Essa abordagem **procura melhores combinações de hiperparâmetros em um determinado espaço amostral**, levando-se em consideração a performance da busca em eventos passados;
- A otimização bayesiana demonstrou superar outros algoritmos de otimização global, pois possui bons resultados e tem melhor custo/benefício se comparado por exemplo ao grid search;
- Nesta abordagem, ela procura, **durante 300 tentativas**, minimizar o erro utilizando a biblioteca **scikit-optimize**.

Modelagem

- Os hiperparâmetros para otimização foram escolhidos com base em artigos publicados, que constam nas referências do trabalho;
- Espaço de otimização dos três modelos:

XGBoost

```
space = [(0.6, 0.7), # colsample_bylevel
        (0.6, 0.7), # colsample_bytree
        (0.01, 1), # gamma
        (0.0001, 1), # learning_rate
        (0.1, 10), # max_delta_step
        (6, 15), # max_depth
        ]
```

Catboost

```
space = [(10, 300), # iterations
        (1, 8), # depth
        (0.01, 1.0), # learning_rate
        (1e-9, 10), # random_strength
        (0.0, 1.0), # bagging_temperature
        (1, 255), # border_count
        (2, 30), # l2_leaf_reg
        ]
```

LGBM

```
space = [(1e-3, 1e-1, 'log-uniform'), #learning_rate
        (2, 128), # num_leaves
        (1, 100), # min_child_samples
        (0.05, 1.0), # subsamples
        (0.1, 1.0), # colsample_bytree
        (0.1, 0.9), # feature_fraction
        (0.8, 1), # bagging_fraction
        (17, 25), # max_depth
        (0.001, 0.1), # min_split_gain
        (10, 25) # min_child_weight
        ]
```

Modelagem

A métrica utilizada para mensurar o erro da regressão foi o **erro quadrático médio**, que possui a seguinte forma de cálculo:

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$

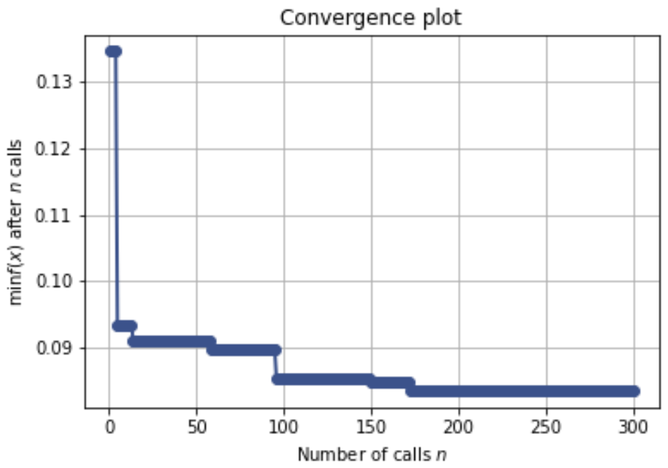
O erro quadrático médio tem a vantagem de **penalizar erros maiores**, o que, em um processo de otimização de hiperparâmetros, pode ajudar a tornar os erros menores.

PUC Minas

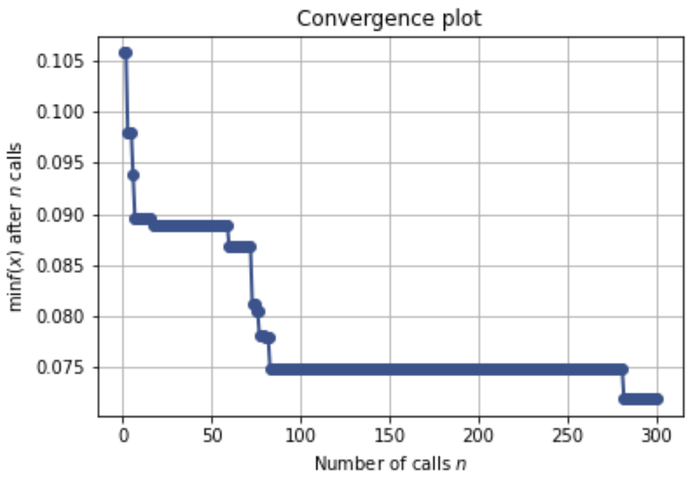
Modelagem

Os gráficos de convergência dos hiperparâmetros ao erro mínimo para estimação de inflação de **um mês**:

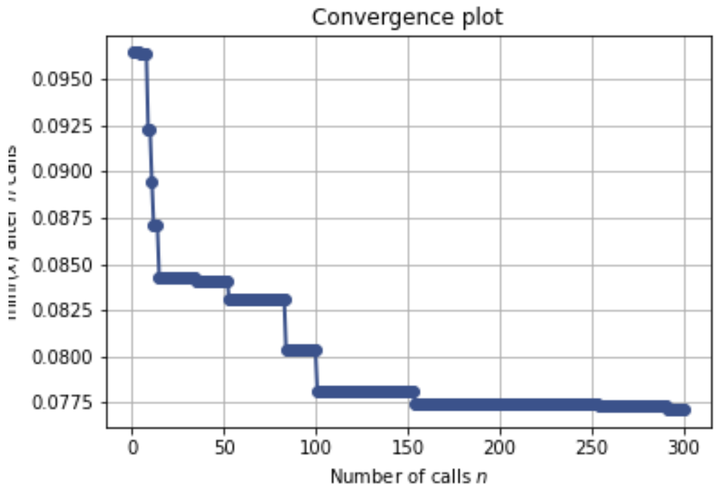
XGBoost para um mês



Catboost para um mês



LGBM para um mês

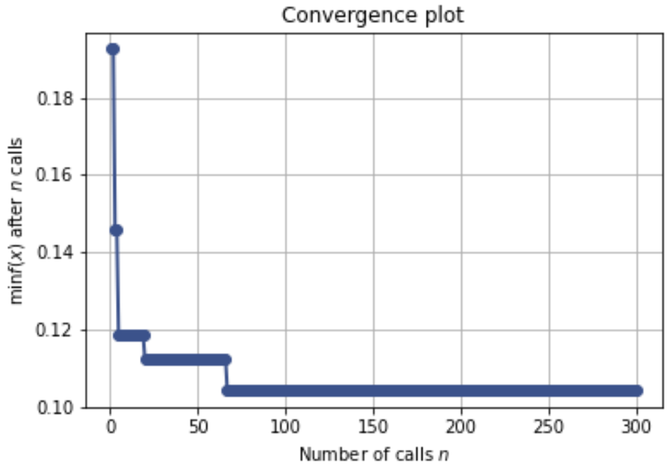


PUC Minas

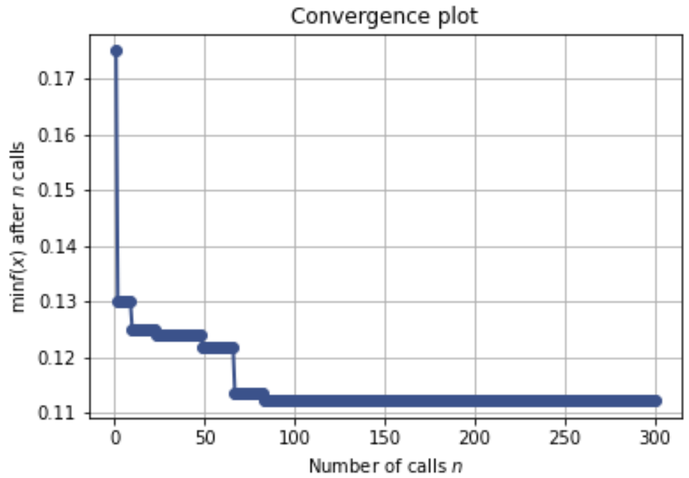
Modelagem

Os gráficos de convergência dos hiperparâmetros ao erro mínimo para estimação de inflação de **quatro meses**:

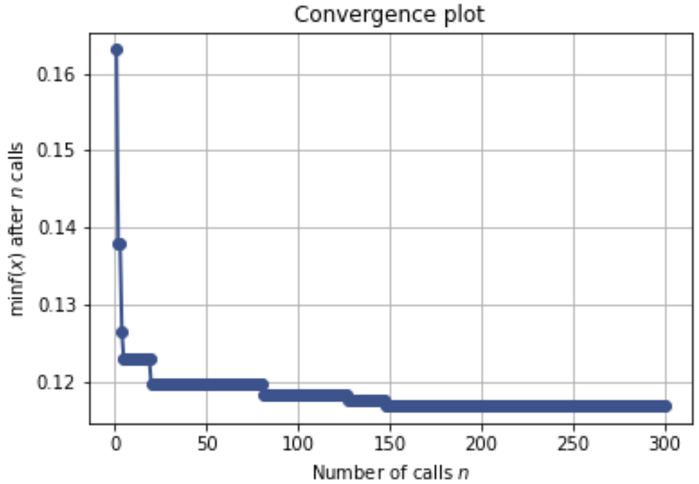
XGBoost para quatro meses



Catboost para quatro meses



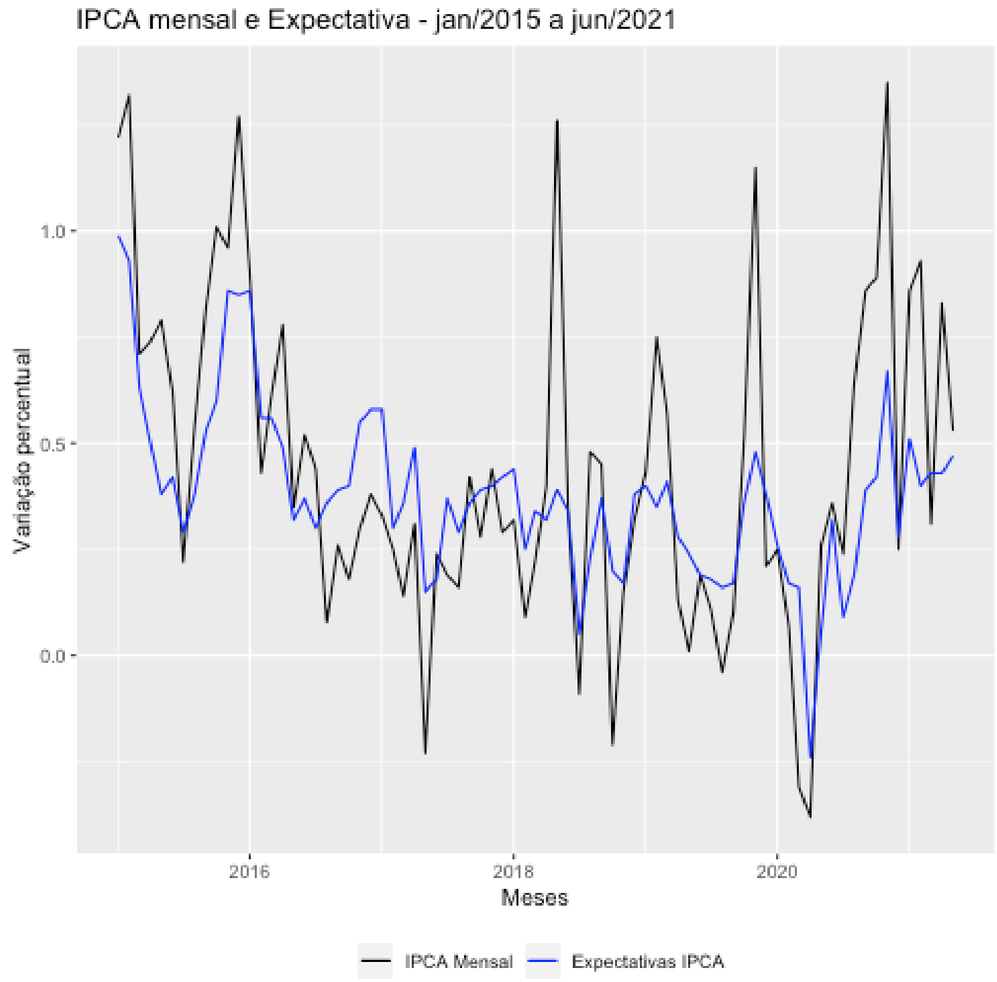
LGBM para quatro meses



PUC Minas

Desempenho dos modelos

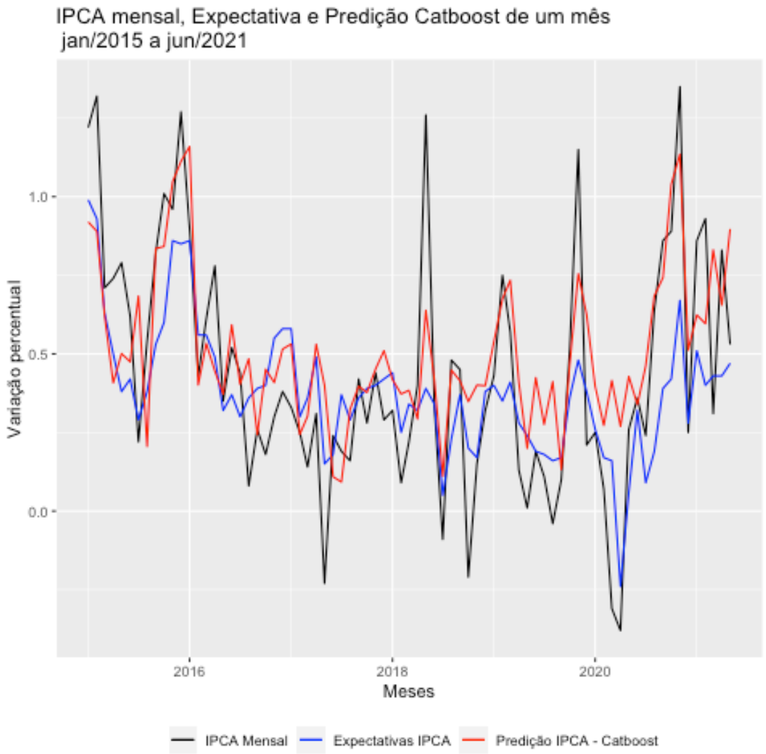
- Para comparação com os valores previstos por cada modelo (*baseline*), utilizou-se os valores reais e os valores do sistema de expectativas de mercado;

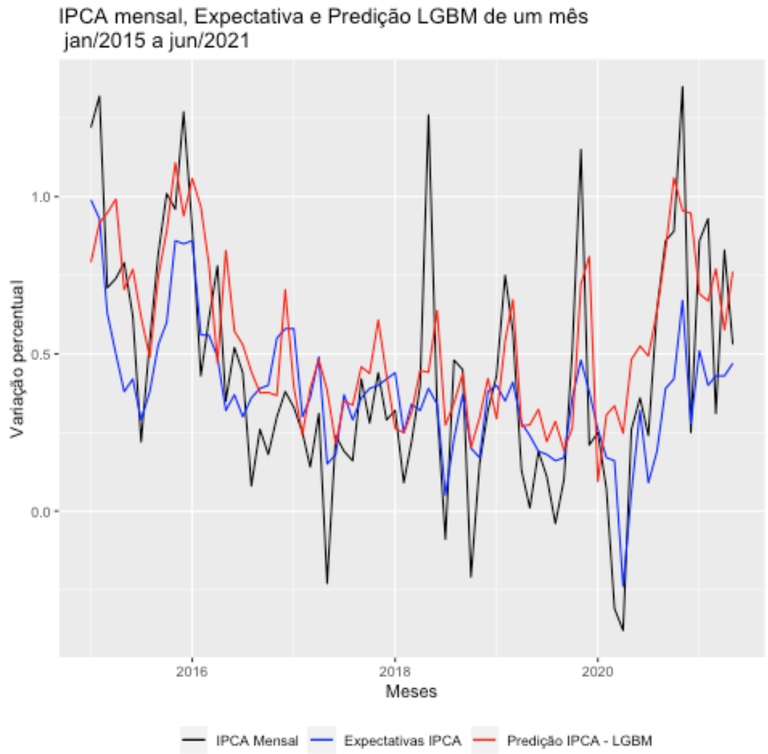


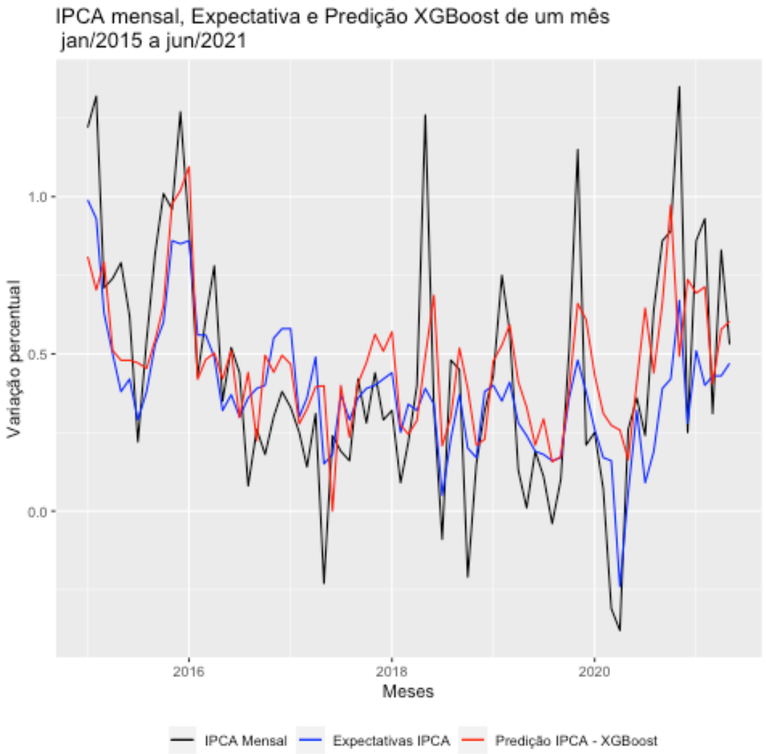
PUC Minas

Desempenho dos modelos

- Os resultados dos modelos para um mês foram:



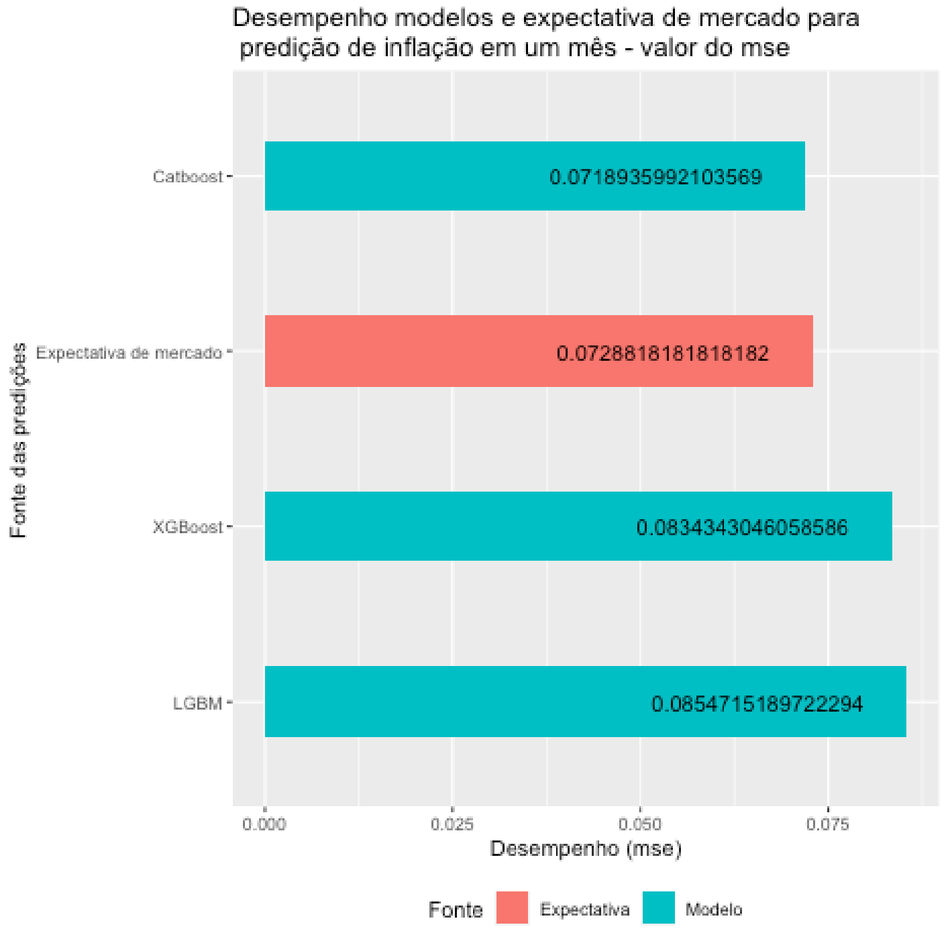


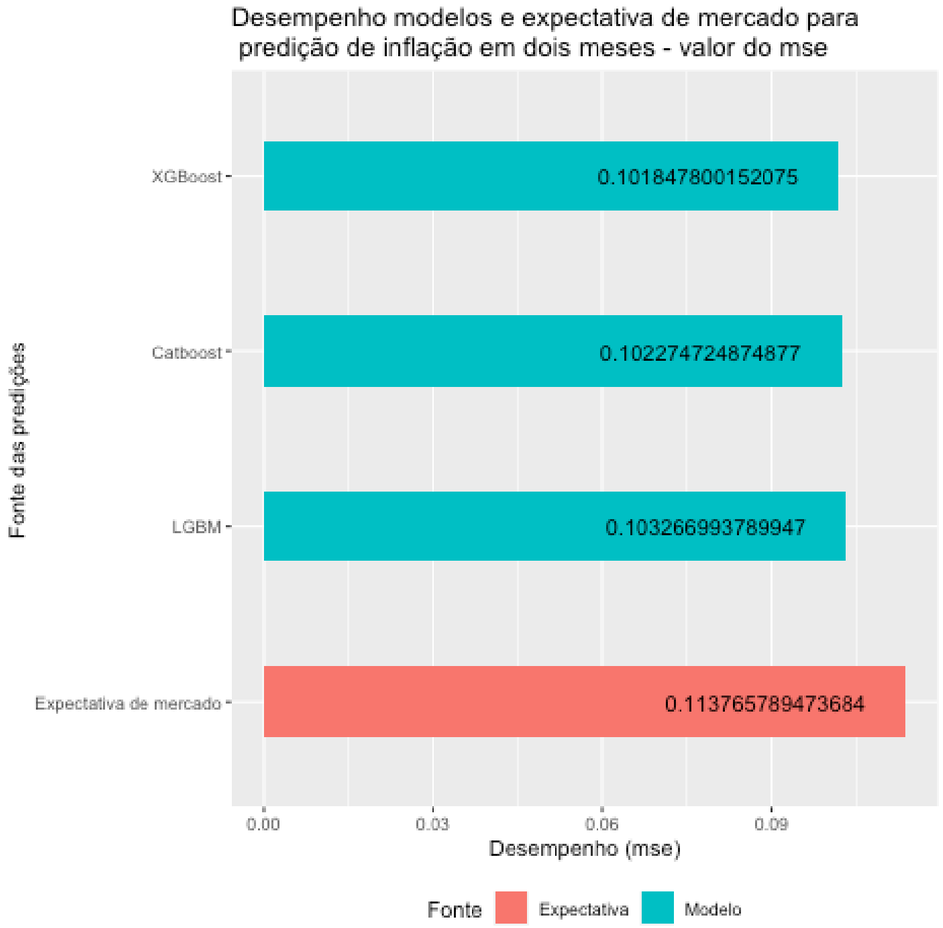


PUC Minas

Desempenho dos modelos

- Os resultados dos modelos para um mês foram:

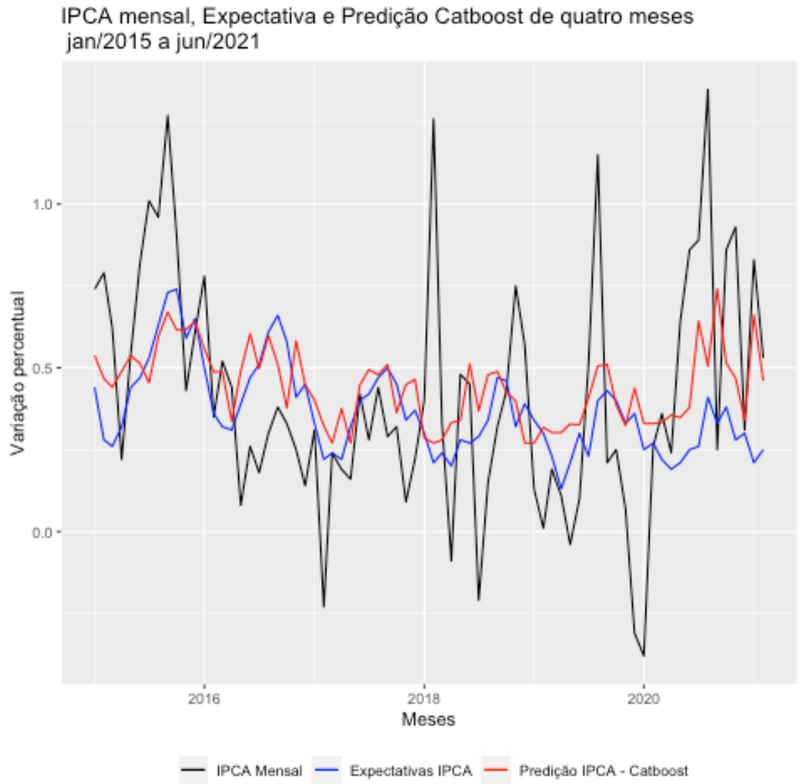


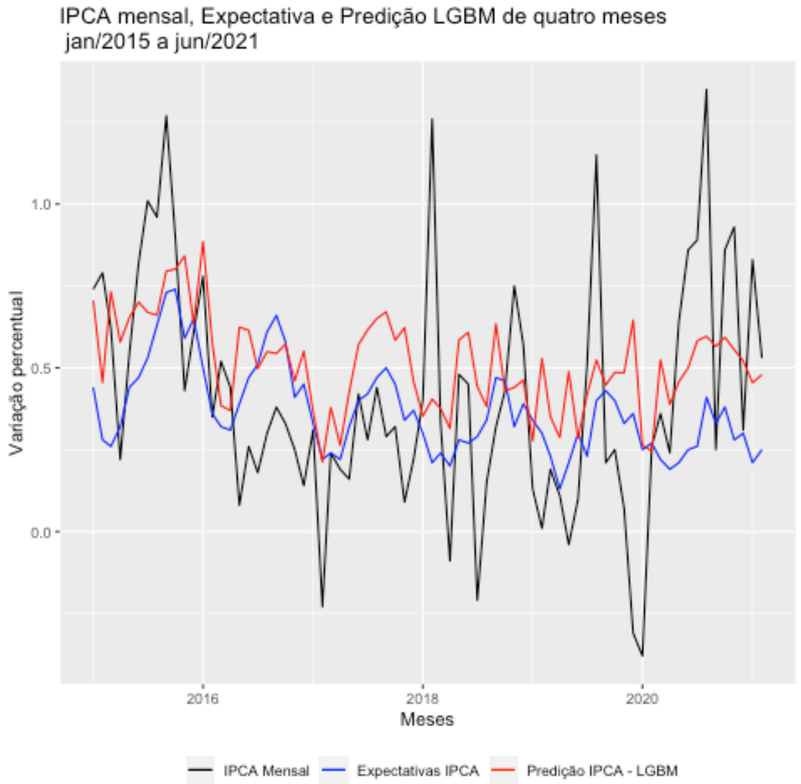


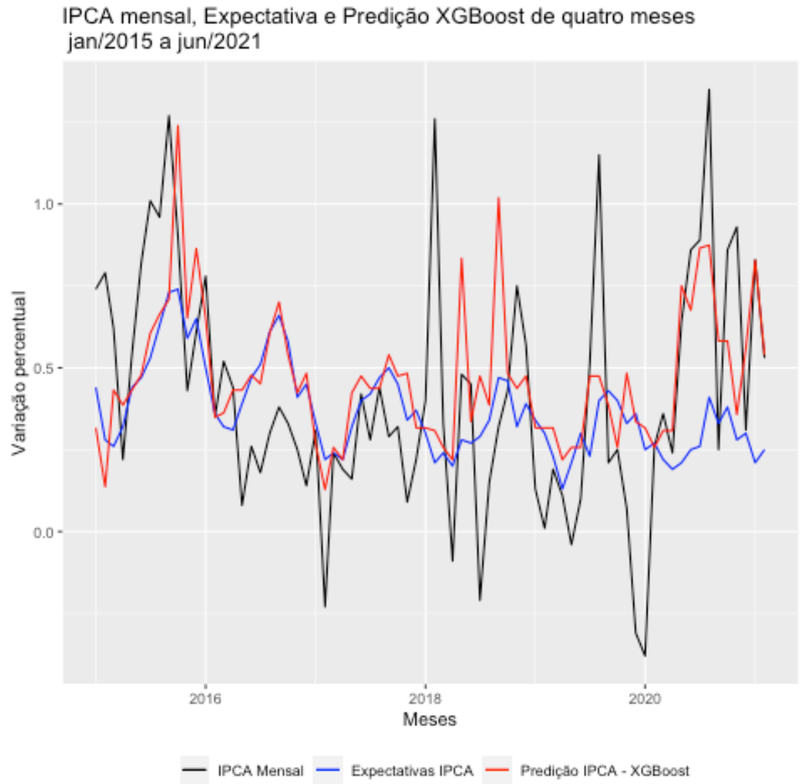
PUC Minas

Desempenho dos modelos

- Os resultados dos modelos para um mês foram:



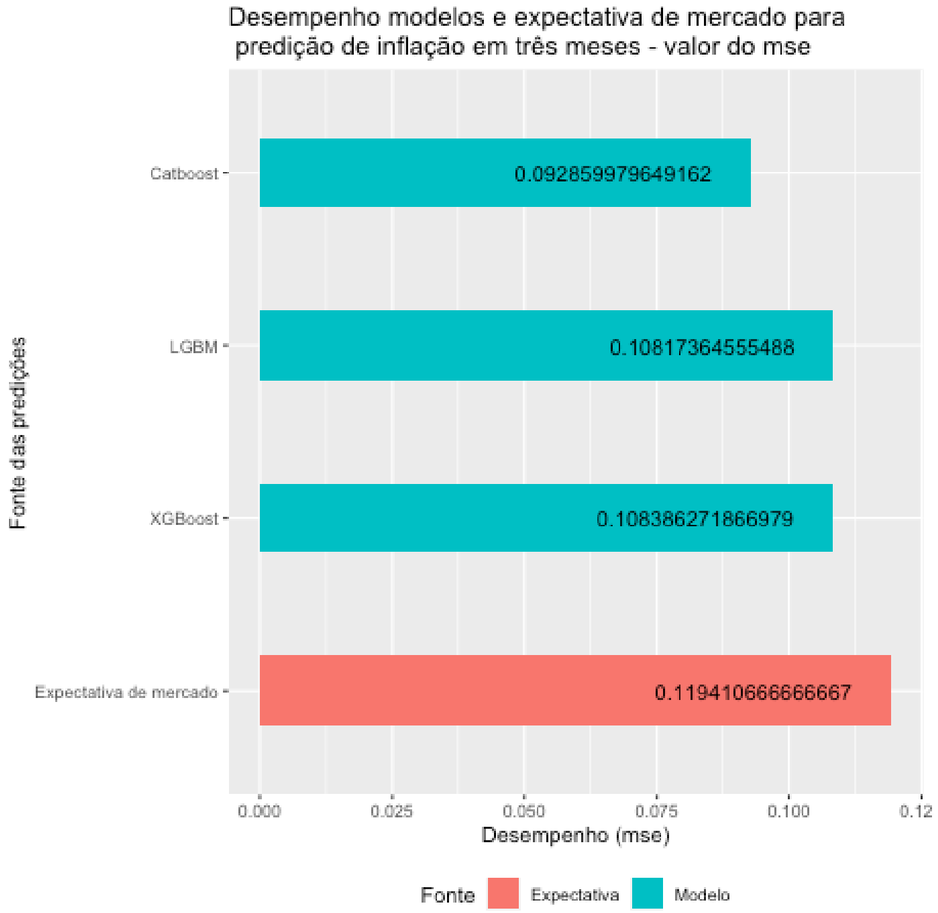


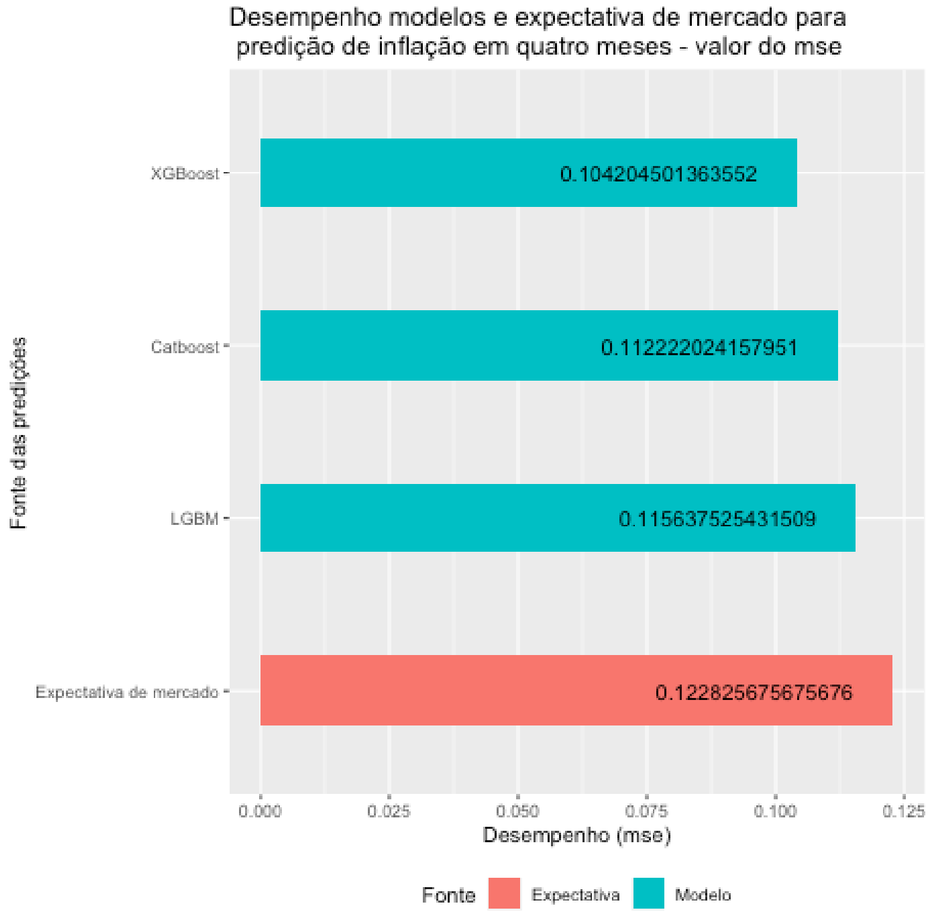


PUC Minas

Desempenho dos modelos

- Os resultados dos modelos para um mês foram:





PUC Minas

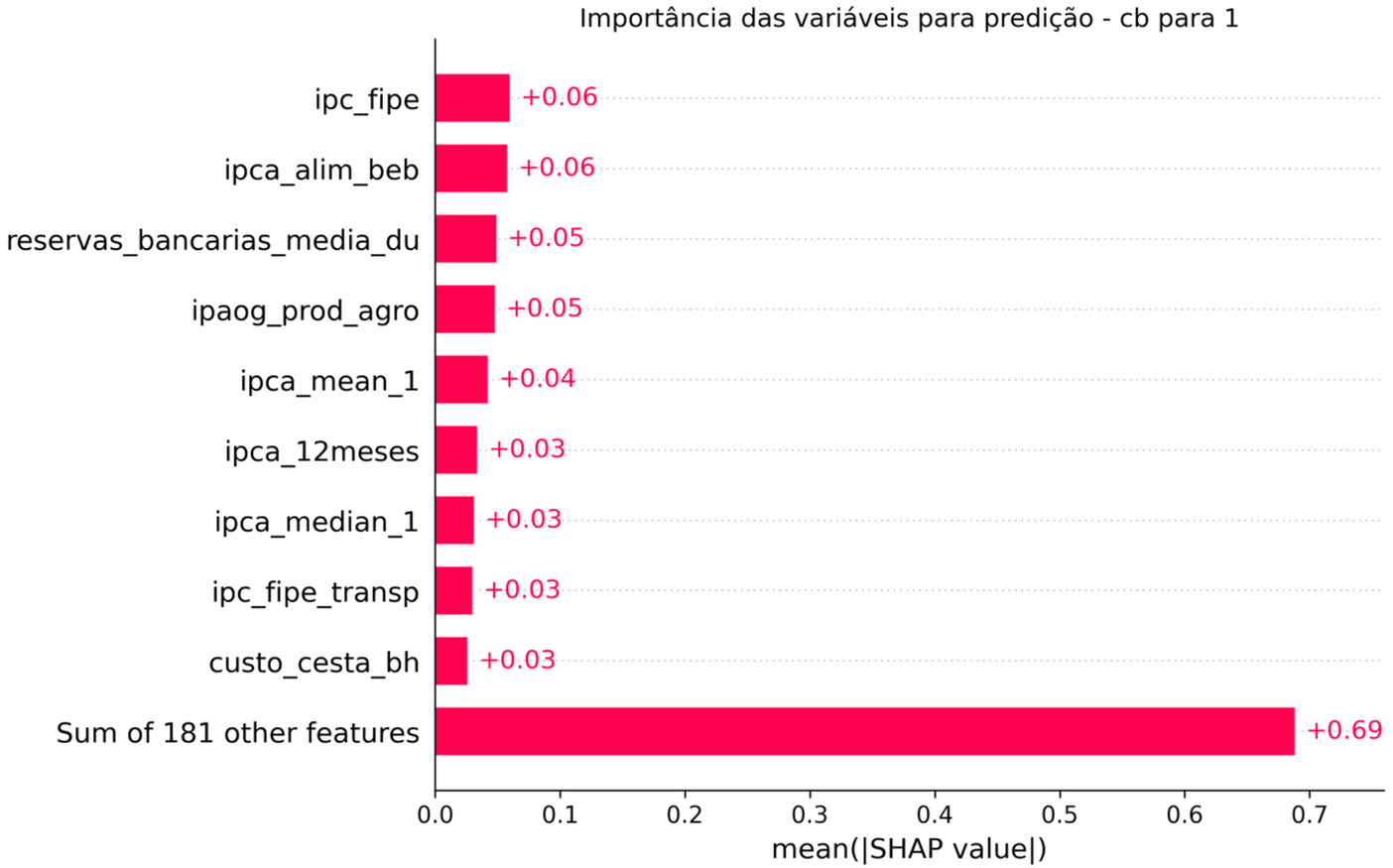
Análise dos resultados

- Na tentativa de esmiuçar melhor os resultados, foi usado a biblioteca SHAP para identificar as features mais importantes para cada resultado;
- A biblioteca SHAP usa da abordagem de **teoria dos jogos** para explicar como os algoritmos tomam decisões;
- Para esse trabalho, extraiu-se o comportamento de cada algoritmo para cada intervalo de mês de predição de inflação.

PUC Minas

Análise dos resultados

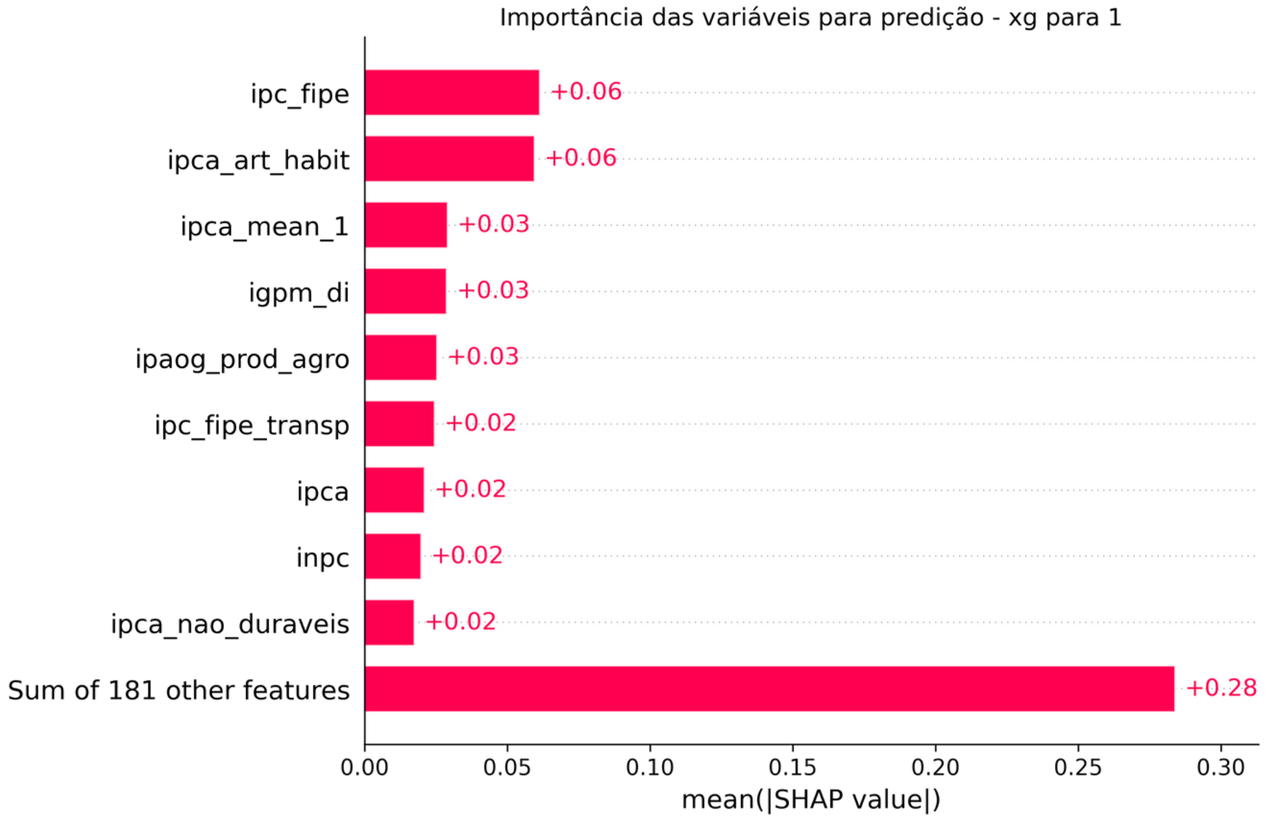
- Para previsão do Catboost para um inflação de um mês, identificou-se que as principais variáveis foram:



PUC Minas

Análise dos resultados

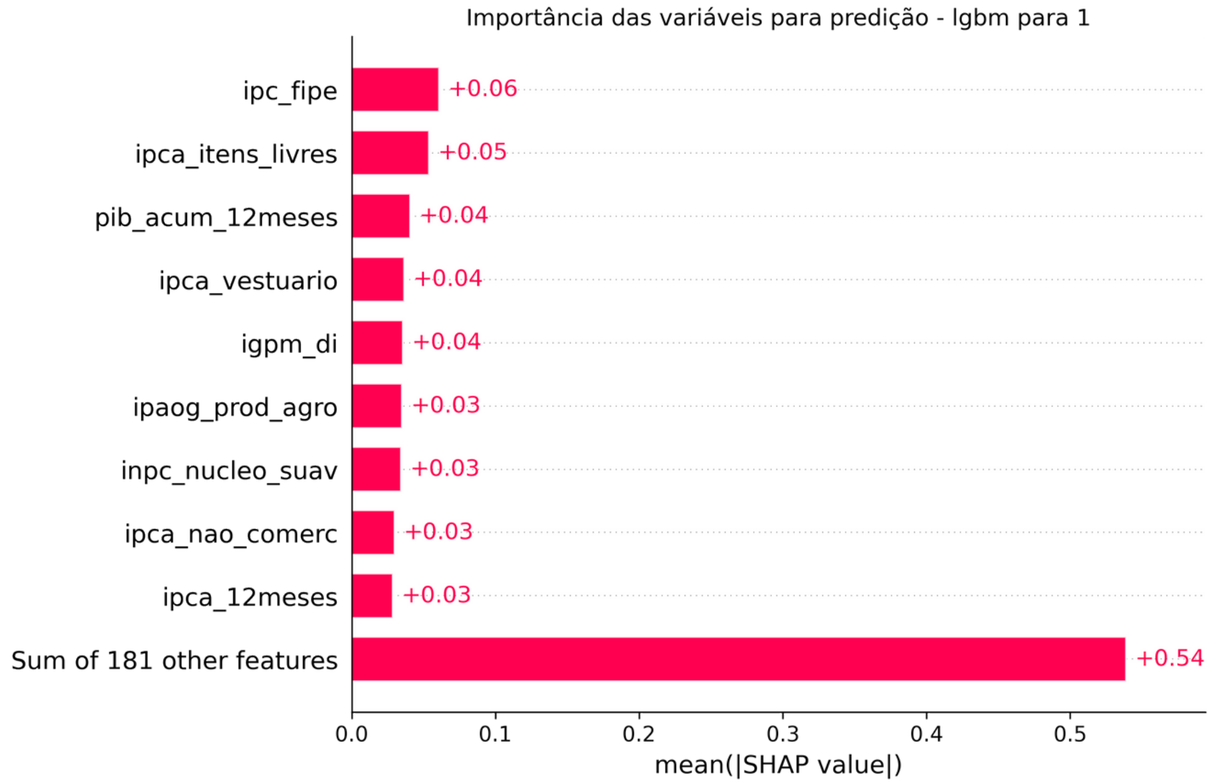
- Para previsão do XGboost para um inflação de um mês, identificou-se que as principais variáveis foram:



PUC Minas

Análise dos resultados

- Para previsão do LGBM para um inflação de um mês, identificou-se que as principais variáveis foram:



PUC Minas

Conclusões

- Os **modelos são úteis** e performam bem para operações que exijam previsão uma vez a cada mês;
- Ainda possui **espaço de melhoria** se buscar novos dados, como novas bases de atividade econômica e comércio exterior;
- O **processo de validação pode ser aprimorado**, enquanto os dados tiverem mais dimensionalidade;
- Mesmo com a utilização da biblioteca SHAP para identificar feature importance, ainda é necessário **mais informações sobre os resultados**;

PUC Minas

Obrigado!

pedrokeyloger@gmail.com

kaggle https://www.kaggle.com/pbizil

PUC Minas