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

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.

PUC Minas

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:

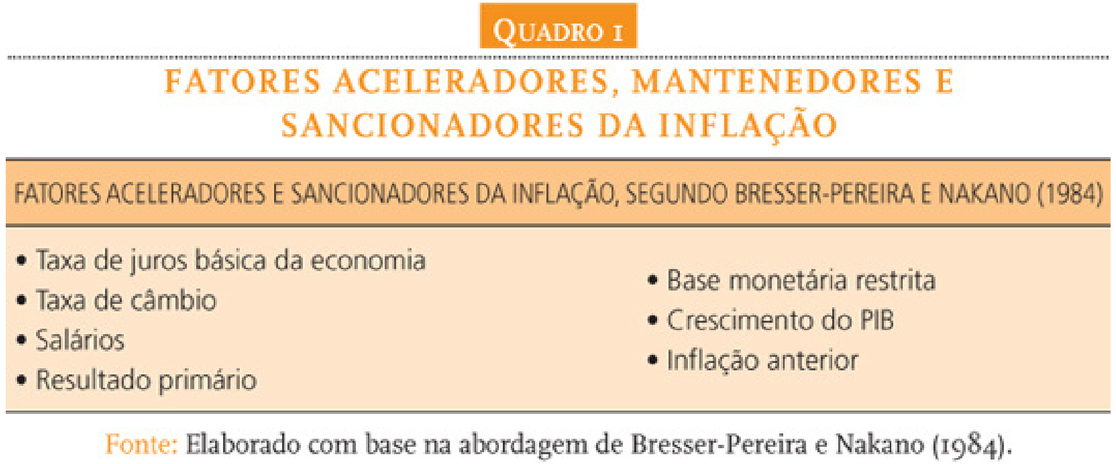
Finanças Estratégicas • RAM, Rev. Adm. Mackenzie 13 (1) • Fev 2012 • https://doi.org/10.1590/S1678-69712012000100004 COPIAR

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)"

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

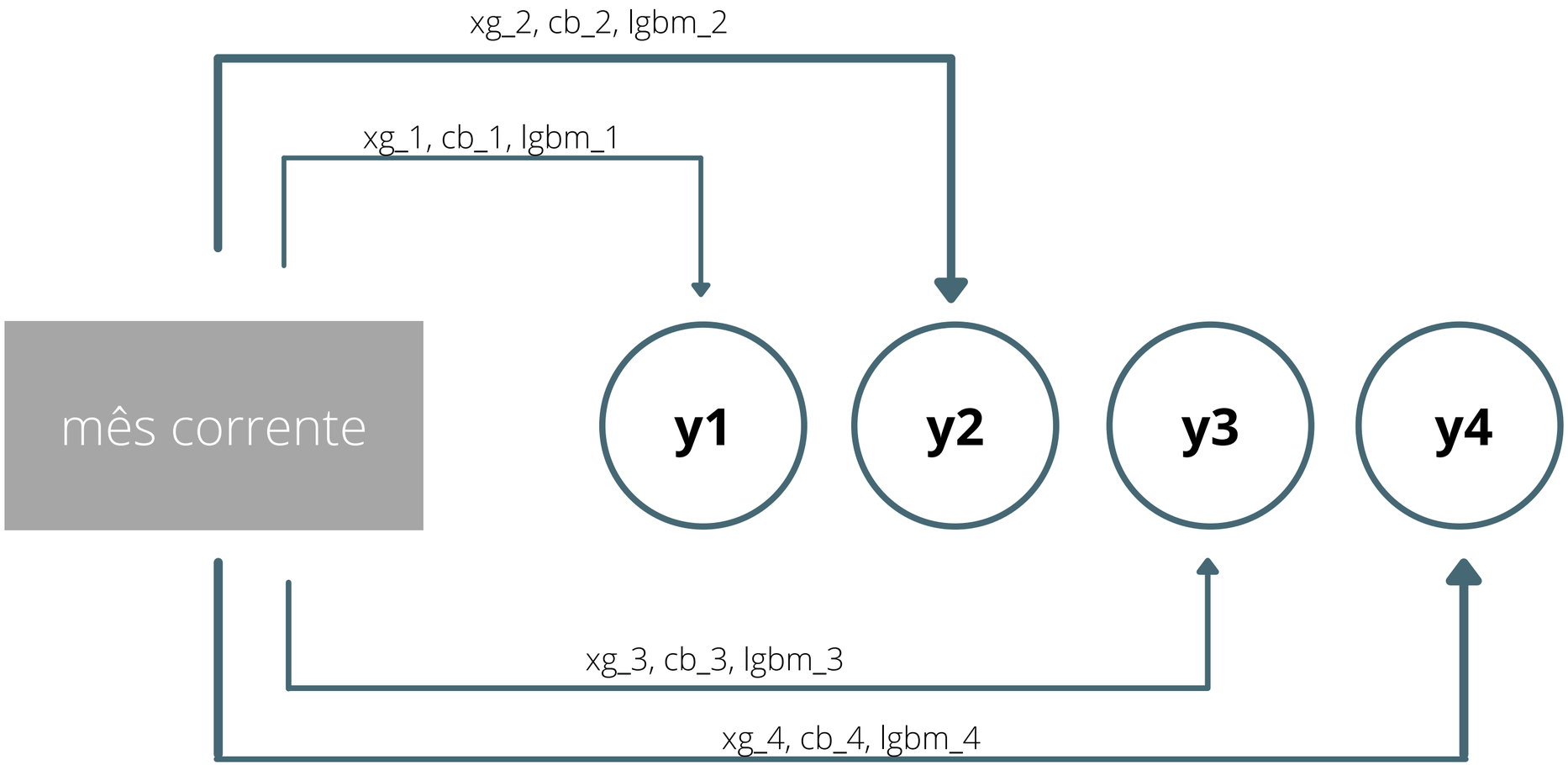
PUC Minas

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;

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**.

PUC Minas

Modelagem

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

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)}^{2}$$

The square of the difference between actual and predicted

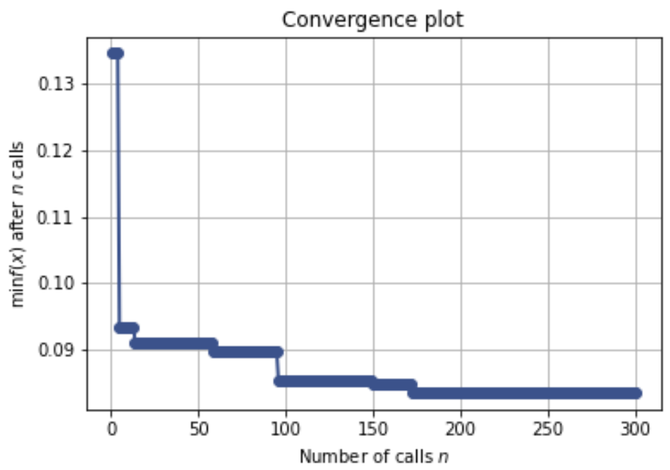
O erro quadrático médio tem a vantagem de **penalizar erros maiores**, o que, em um processo de otimização de hiperparâmetros, pode ajudar a tornar os erros menores.

PUC Minas

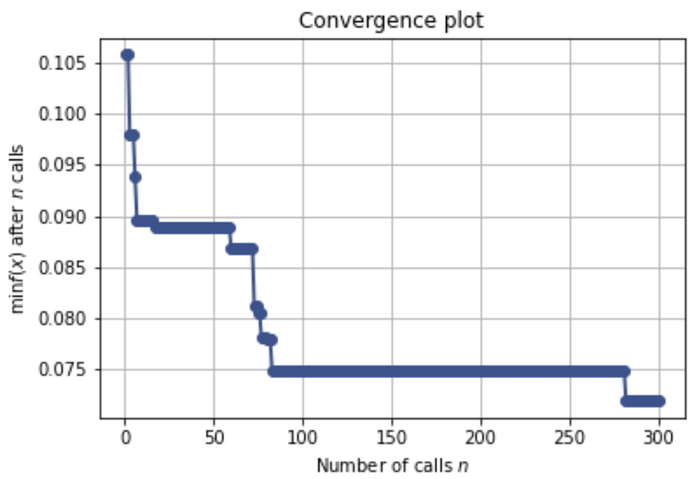
Modelagem

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

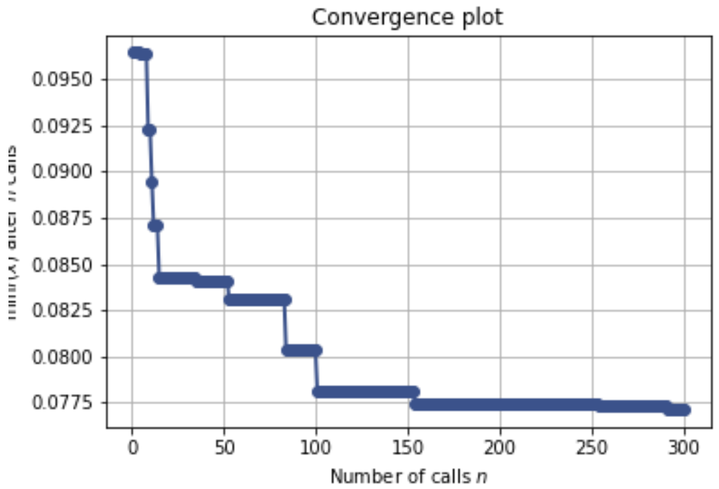
XGBoost para um mês



Catboost para um mês



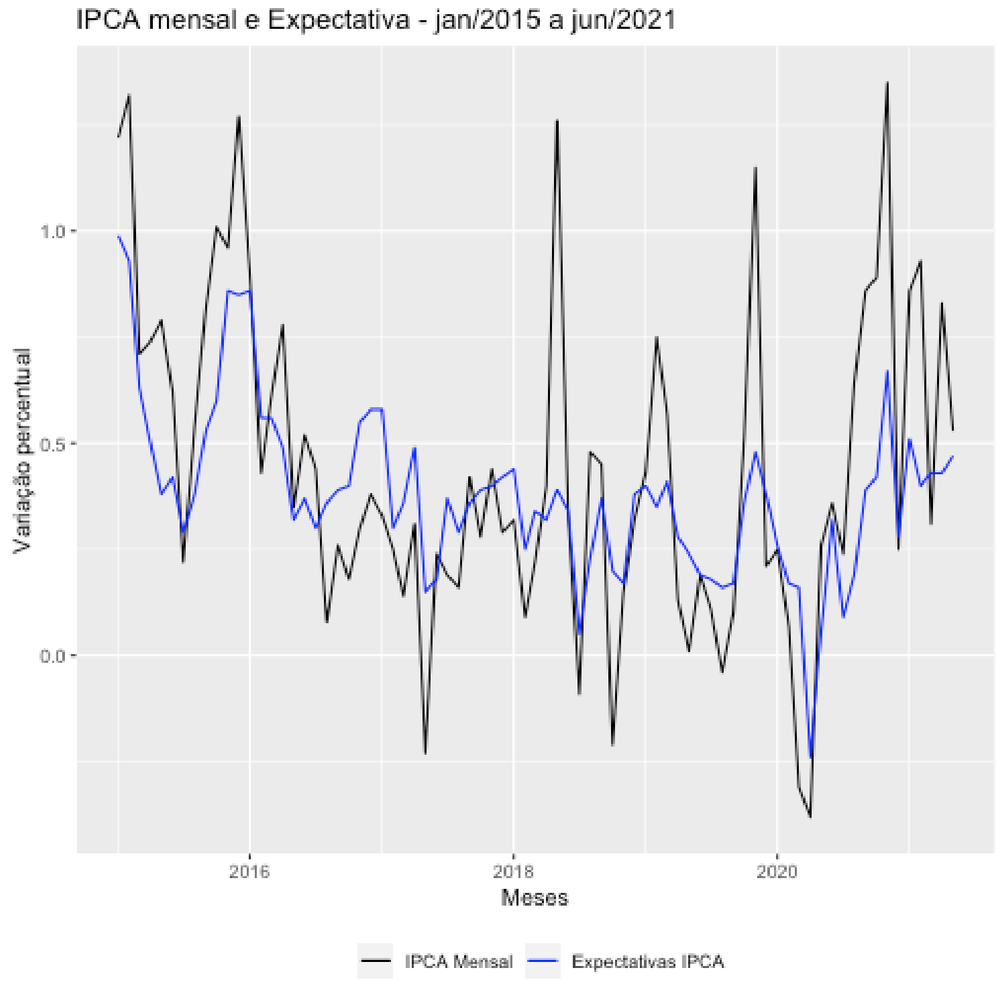
LGBM para um mês



PUC Minas

Desempenho dos modelos

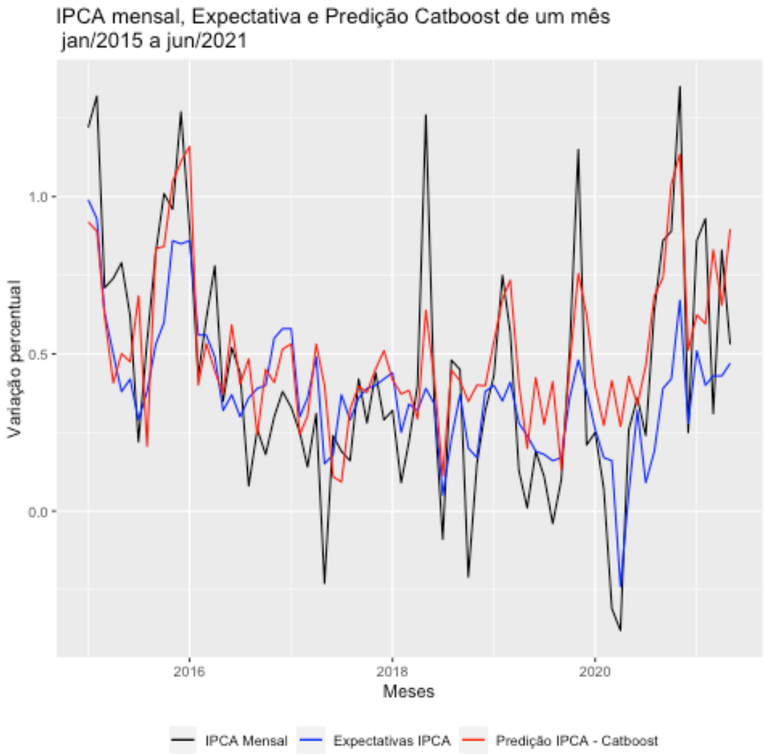
- Para comparação com os valores previstos por cada modelo (*baseline*), utilizou-se os valores reais e os valores do sistema de expectativas de mercado;

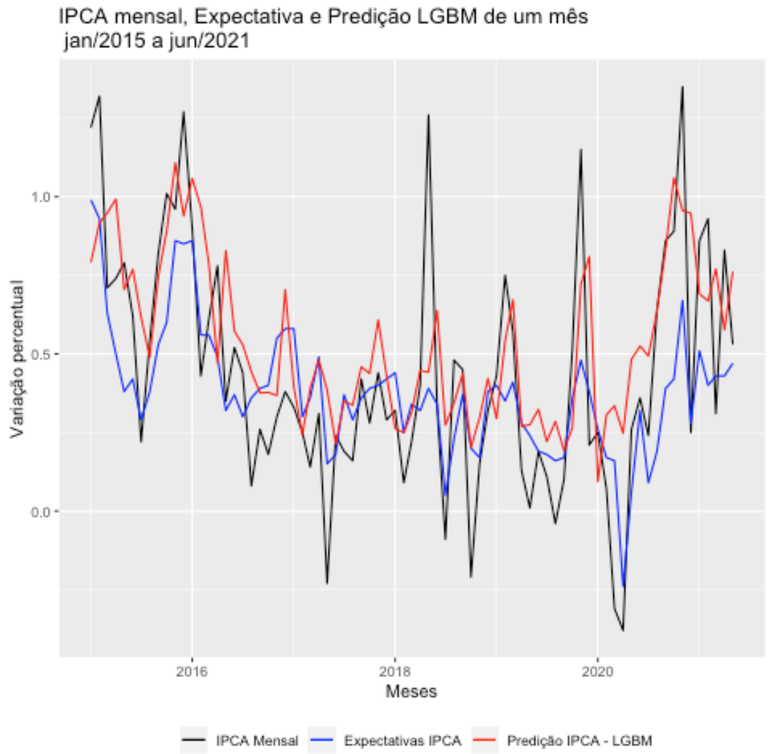


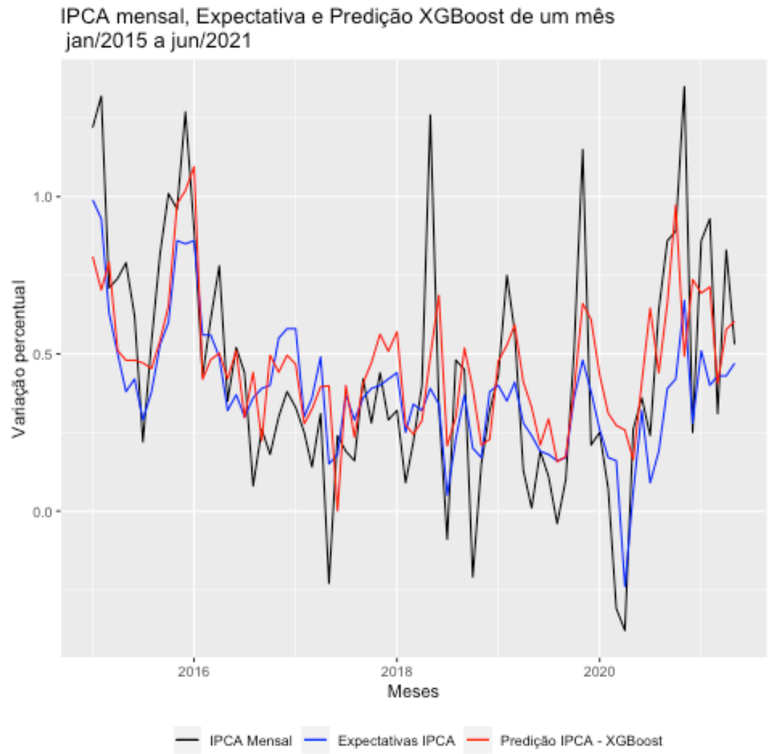
PUC Minas

Desempenho dos modelos

- Os resultados dos modelos para um mês foram:



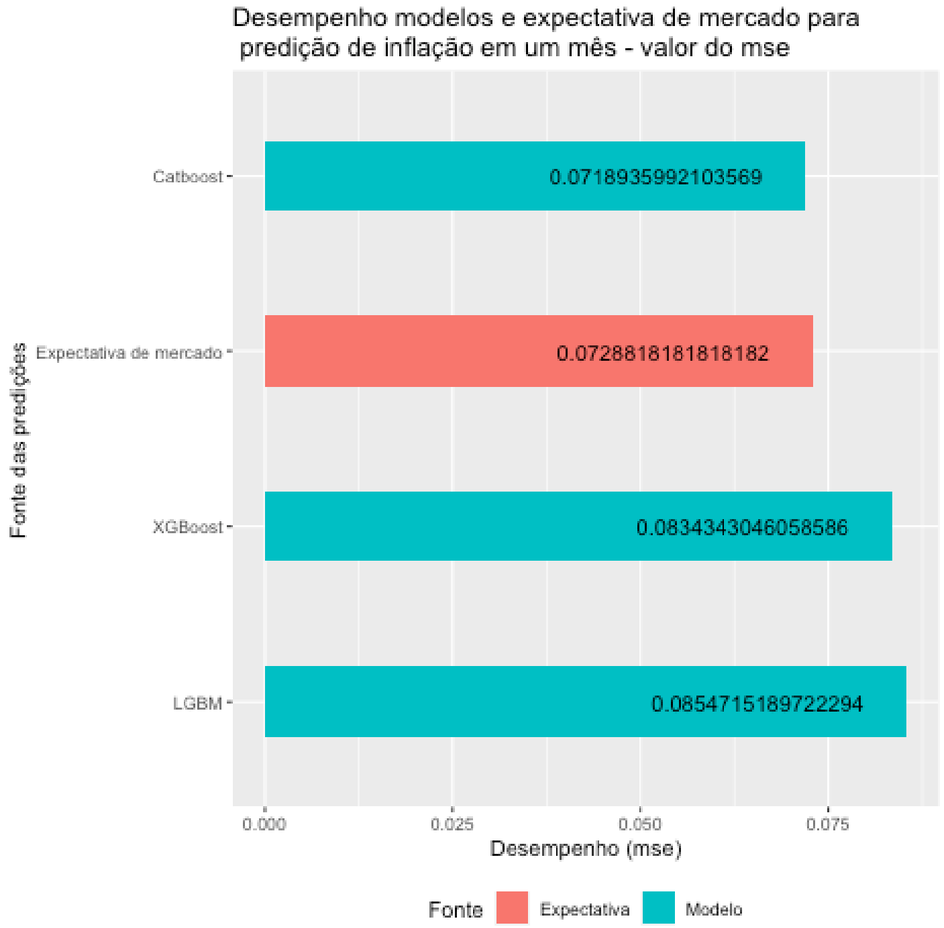


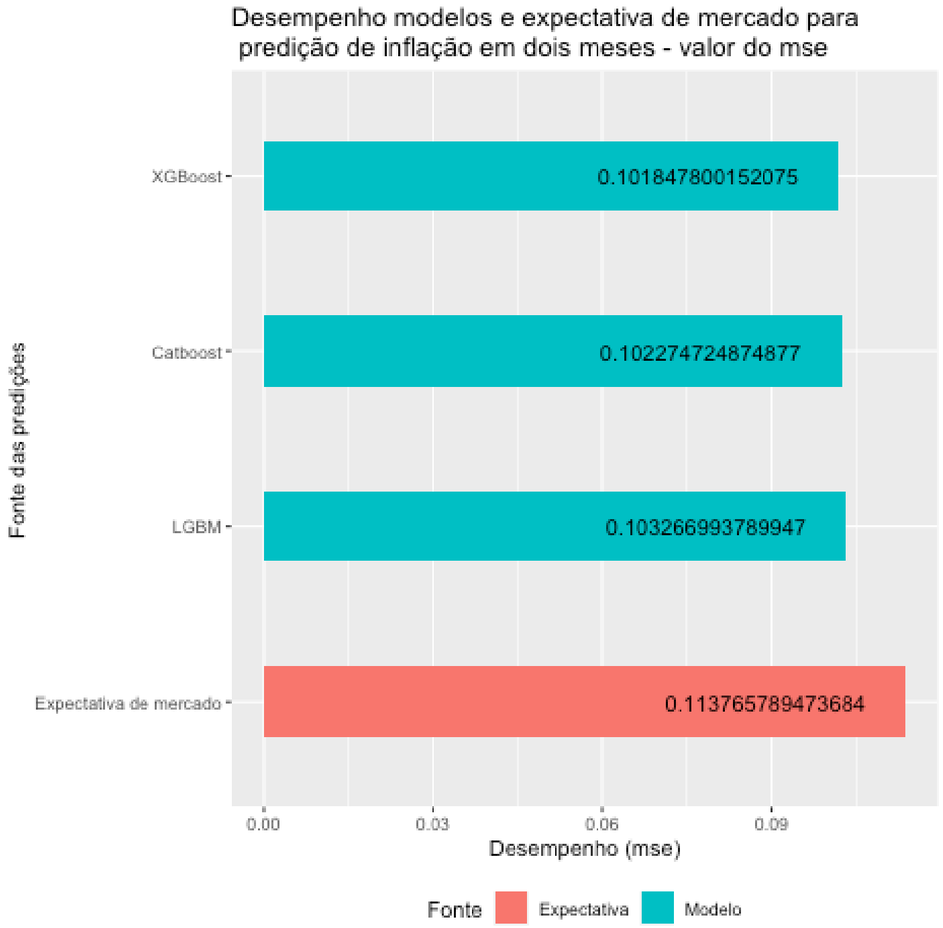


PUC Minas

Desempenho dos modelos

- Os resultados dos modelos para um mês foram:





PUC Minas

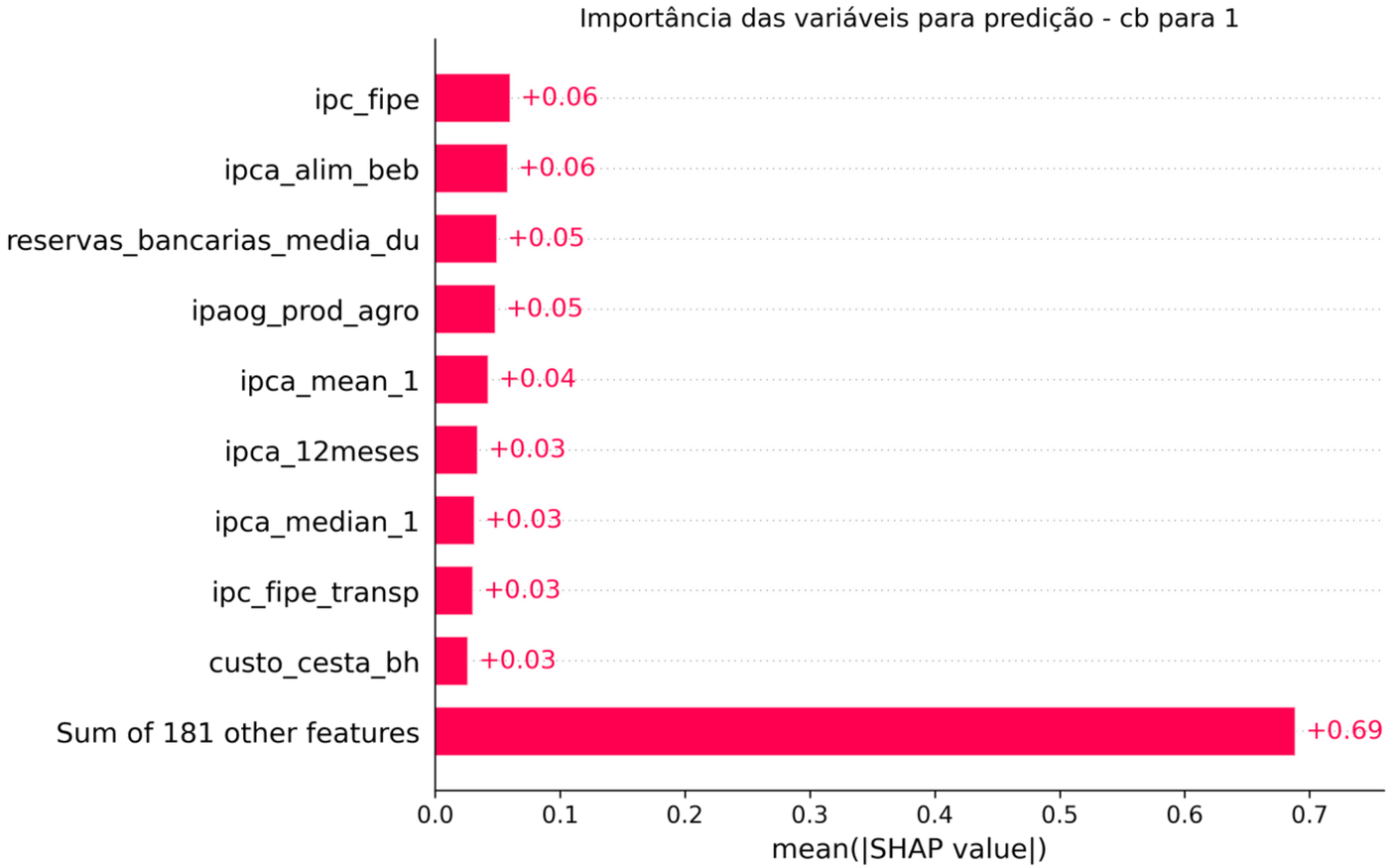
Análise dos resultados

- 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

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