

Machine Learning na Previsão de Consumo do Cimento, com Análise Interpretativa por Valores SHAP

Pedro Fernandes - Prof. Dr. Marcelo Finger - Instituto de Matemática e Estatística - Universidade de São Paulo



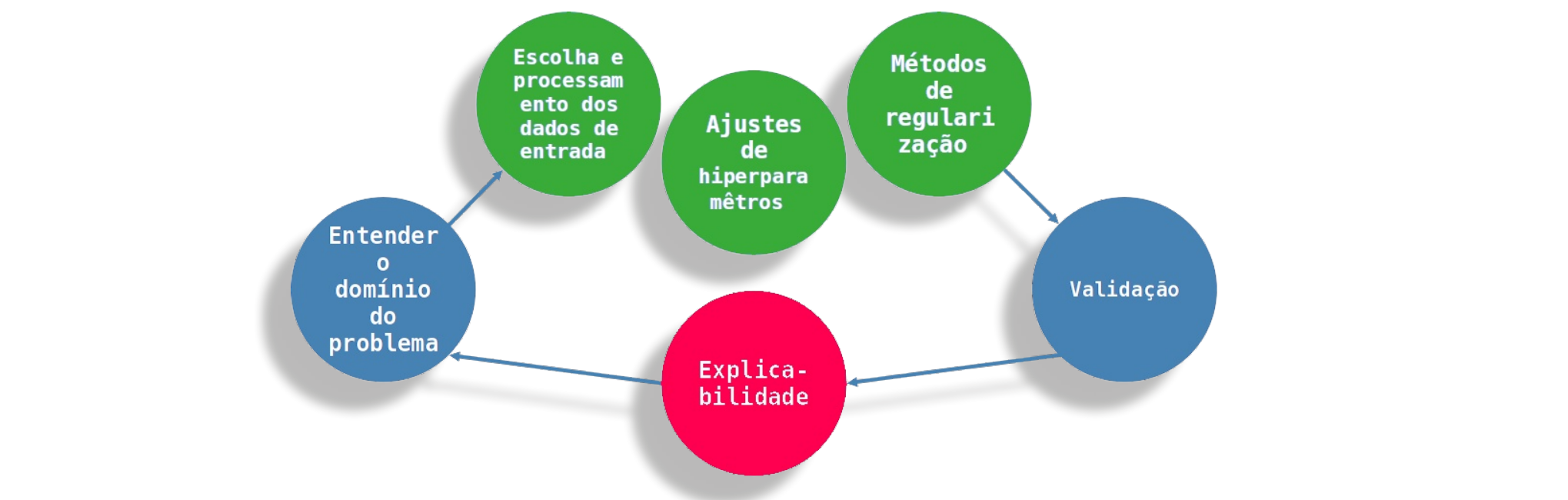
Machine Learning e Consumo de Cimento

A capacidade superior dos modelos baseados em aprendizado, especialmente os neurais, de capturar relações complexas entre os dados de entrada vem se tornando a primeira opção quando se trata de previsibilidade em domínios complexos cujos dados influenciadores são multivariados. Contudo, para construção desses modelos, deve haver atenção para a teoria base, pesquisas em andamento e conhecimento do domínio do problema. Este trabalho tem o consumo de cimento como estudo de caso.

A assertividade de sua previsão é de crítica importância para a indústria produtora e sociedade em geral, pois afetaria diretamente as estratégias de produção, logística, minimização de desperdícios e redução de poluição em todo processo.

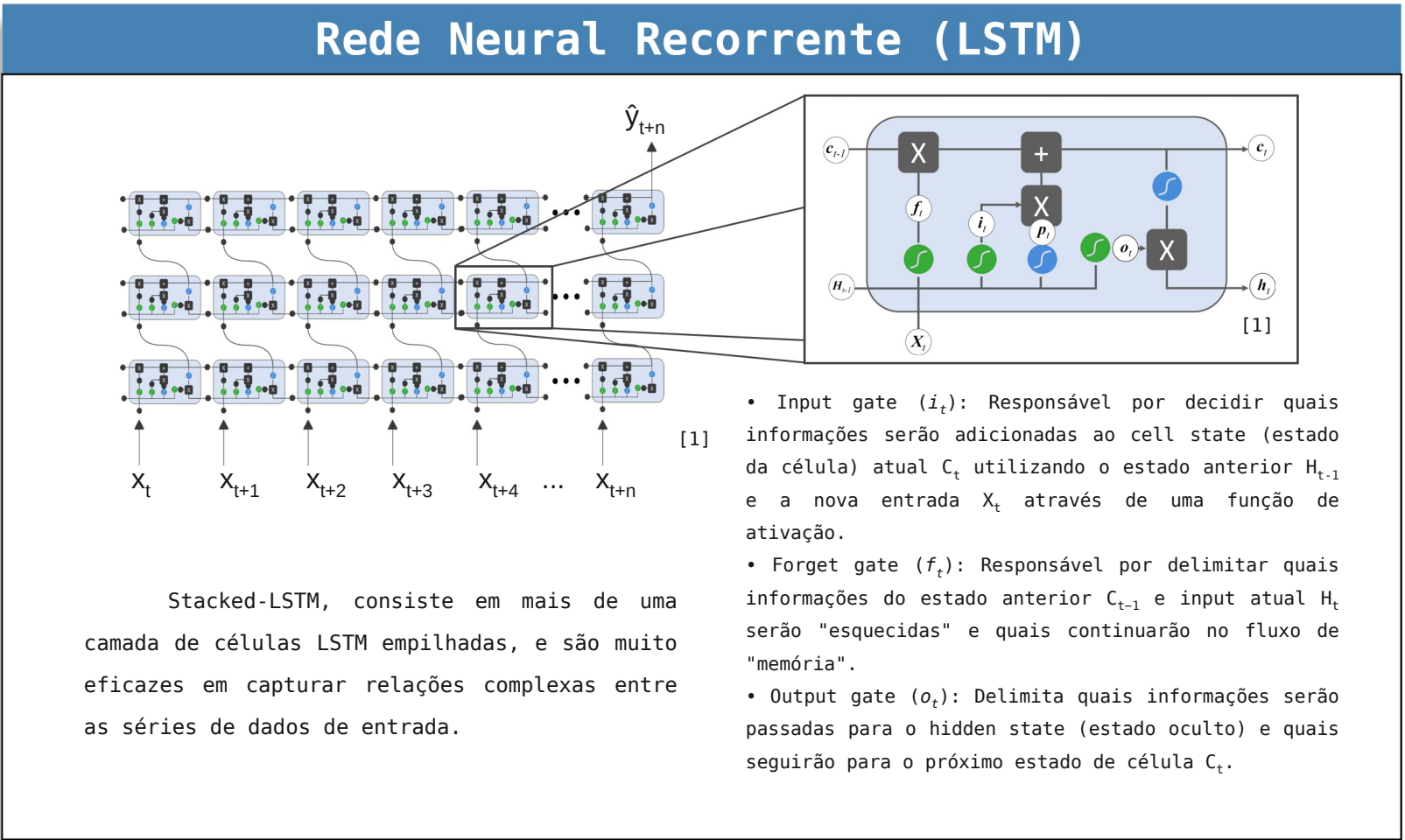
O objetivo deste trabalho é aprimorar modelos neurais recorrentes, como as redes Long-Short Term Memory (LSTM), no contexto da previsão de série temporal de consumo de cimento por unidade federativa brasileira gerando resultados de predição assertivos para o consumo referente ao intervalo de um ano.

Metodologia



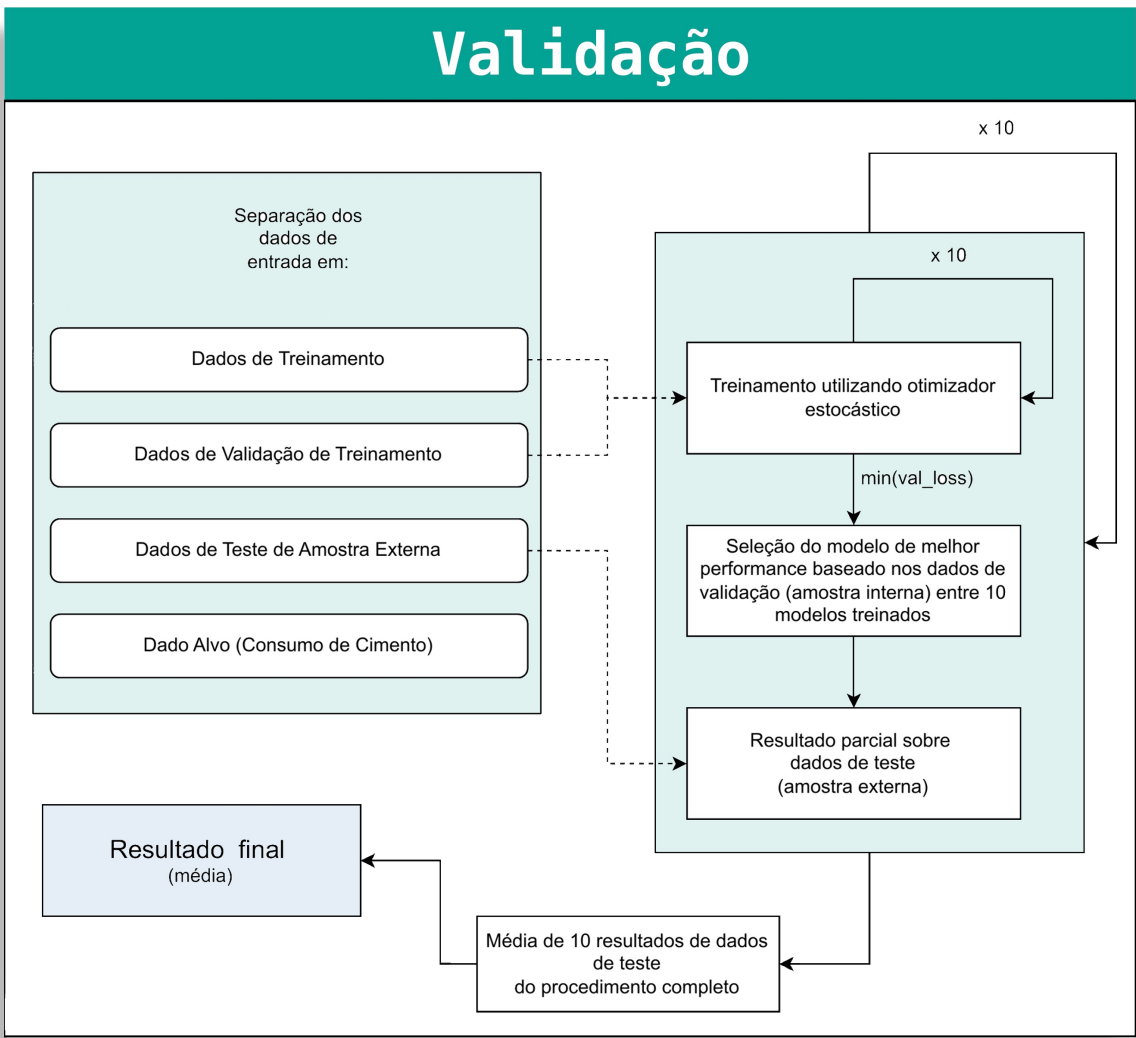
Dados de Entrada		
Socioeconômicos	Compilação de dados de indicadores sociais, indicadores de crescimento e de políticas econômicas. E.g.:	IDH, PIB, IPCA, IGP-DI etc.
Meteorológicos	Conjunto de informações usuais de condições temporais atmosféricas	Precipitação, Pressão Atmosférica, Umidade Relativa do Ar etc.
Construção	Dados e indicadores relativos a construção civil	Produção de Cimento, Valor do Cimento, Custo do M² de Área Construída etc.

Pré-processamento de Dados		
Imputação	Imputação sobre janelas ou pontos de dados ausentes	Valor de fora do domínio, Interpolação Lineal ou Polinomial
Média Móvel	Suaviza os dados reduzindo o impacto de ruídos ao passo que cada valor é uma média de seus vizinhos	Média móvel simples e centrada
Transformações	Aprimora a eficiência na captura de relações pelo modelo	Padronização e Normalização



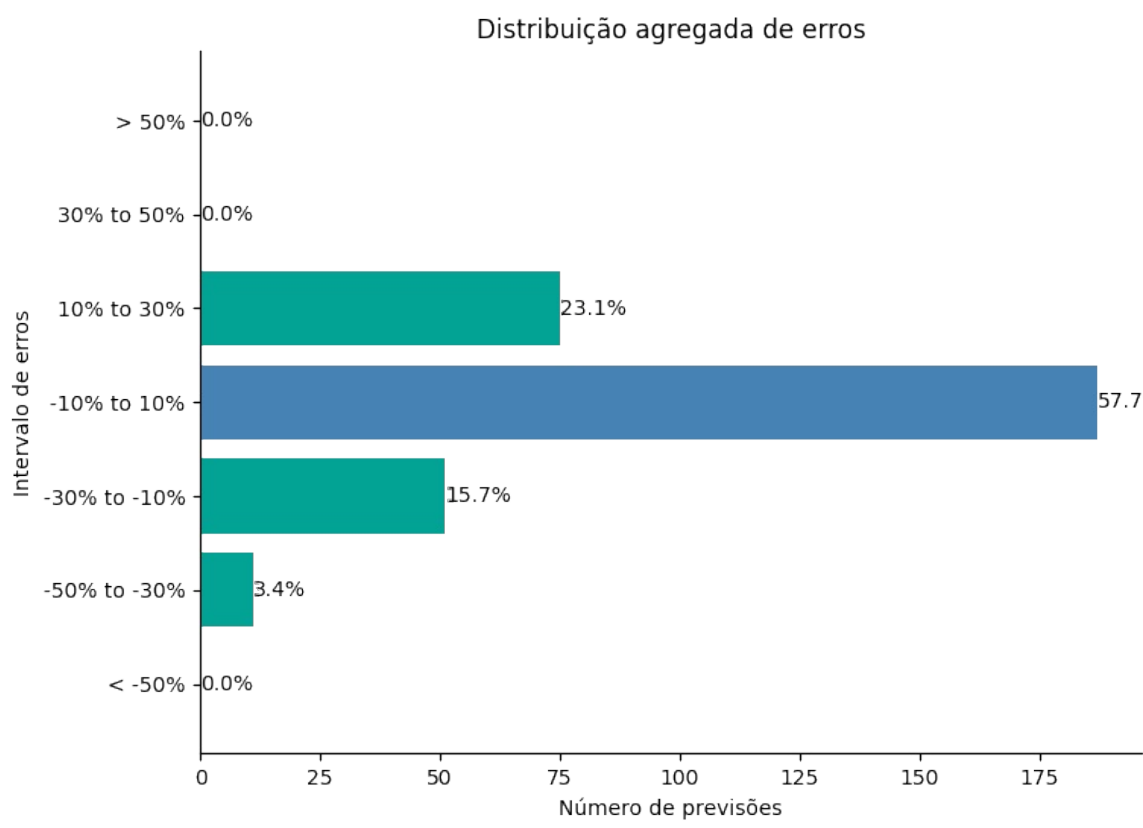
Explicabilidade

Valores SHAP atuam como um conjunto de indicadores $\phi(i)$ que representam a contribuição de cada tipo de dado i (feature) no resultado gerado pelo modelo de previsão.

$$\phi(i) = \sum_{z' \subseteq N} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus \{i\})]$$


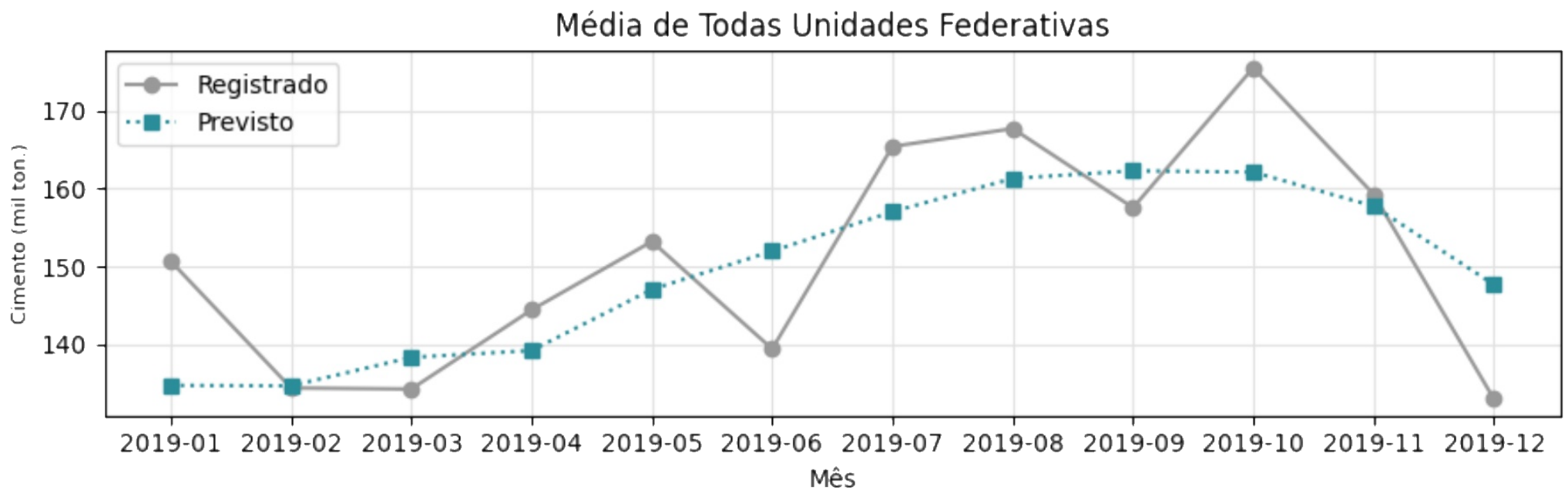
Experimentos e Resultados

Resultados		RMSE (mil ton.)	MAE (mil ton.)	MAPE %	Desv.Padrão %
Fase 1 -	Dados iniciais brutos	21,01	17,65	15,61	9,78
Fase 2 -	Dados iniciais com interpolação e média móvel	20,44	16,58	13,42	10,74
Fase 3 -	Novos tipos de dados inseridos	15,71	12,9	9,88	7,32

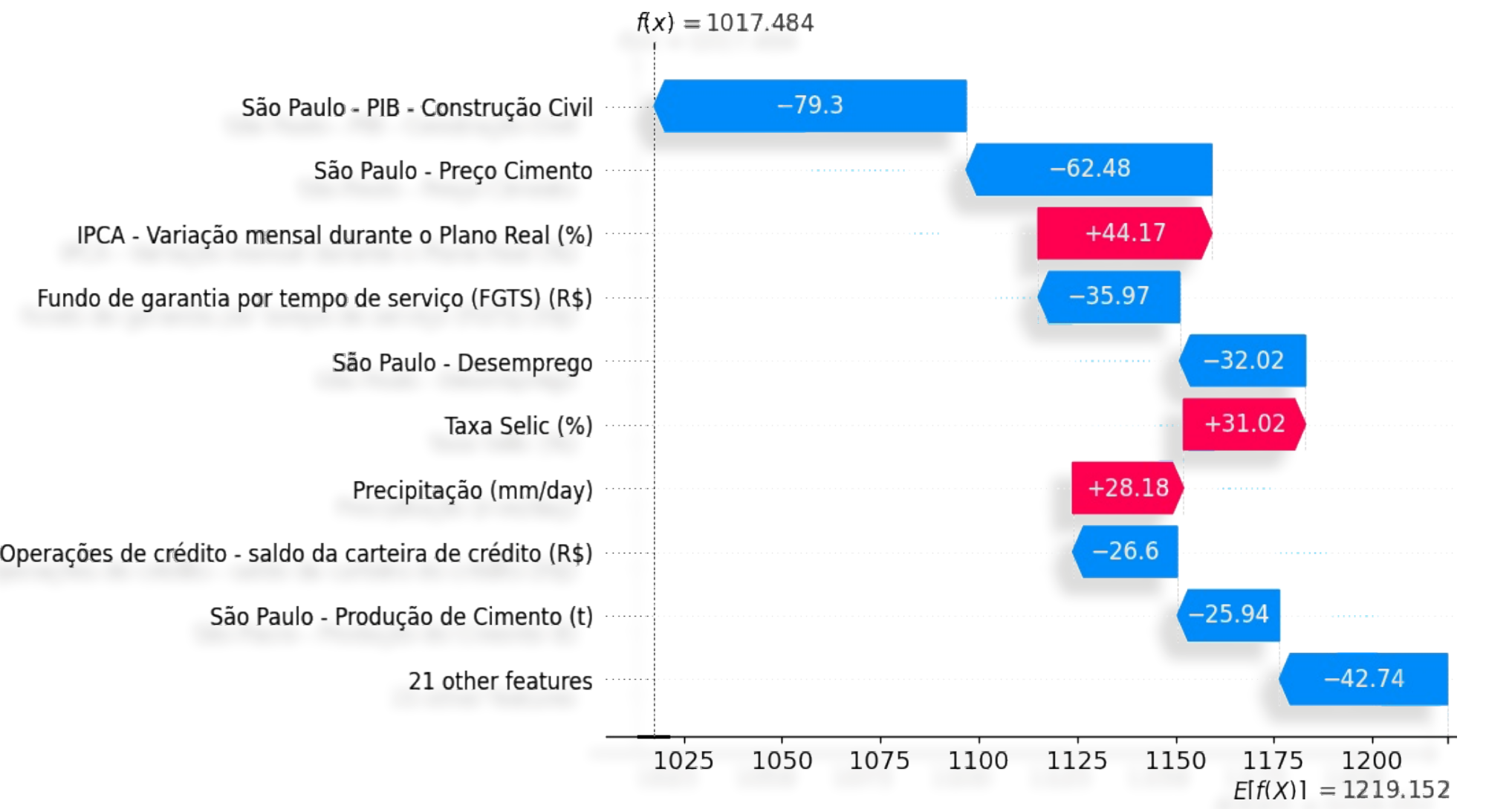


Os resultados para previsão do consumo de cimento para o ano de 2019, usando dados de 2017 e anteriores, mostram tendência de aprimoramento pelas métricas estatísticas utilizadas para avaliar a precisão dos modelos, ao passo em que avançamos nas etapas de pré-processamentos adicionais e inclusão de novos dados.

As séries temporais previstas conseguiram capturar as tendências semestrais do consumo de cimento.



Interpretabilidade por Valores SHAP



A análise de interpretabilidade por meio dos valores SHAP confirmou a significativa influência das variáveis socioeconômicas relacionadas à construção civil e ao poder financeiro da população e país no domínio do problema.

Temos, como principais variáveis, para as unidades federativas estudadas, as referentes a construção, PIB ligado a construção civil, custo do metro quadrado de área construída, assim como aquelas de esfera mais econômica, como FGTS, taxa SELIC e, em segundo plano, NFSP. Além disso, variáveis ligadas ao poder de compra da população, como operações de crédito, inadimplência e IPCA, também se destacam.

Conclusão

Ao longo das três fases de experimentos progressivos, ampliamos os horizontes dos resultados, incorporando técnicas de pré-processamento e uma maior variedade de tipos de dados. As melhorias estatísticas observadas nas diferentes etapas de pré-processamento e adição de dados pertinentes corroboram a importância e os benefícios de estratégias cuidadosas na manipulação de dados para maior eficácia dos modelos de previsão.

Referências

- Abu-Mostafa et al. **Learning from Data**, AMLBook.com.
- Charu C. Aggarwal. **Neural networks and deep learning**. Springer, 2018.
- Chris Chatfield. **Time-series forecasting**. CRC press, 2000.
- Ian Goodfellow et al. **Deep Learning**. 22ª ed. The MIT Press, 2017.
- Scott M. Lundberg e Su-In Lee. **"A unified approach to interpreting model predictions"**. Advances in neural information processing systems v.30, 2017.70
- Ricardo P. Masini et al. **"Machine Learning advances for time series forecasting"**. Journal of economic surveys 37.1 (2023), pp. 76–1110