II Workshop de Matemática, Estatística e Computação Aplicadas à Indústria

Aprendizado de Máquina aplicado à predição do preço do arroz utilizando dados climatológicos e econômicos

Lucas Valle Mielke[1]
ICMC - USP
Paulino Ribeiro Villas Boas[2]
Embrapa Instrumentação
ICMC - USP

1 Introdução

O arroz é um dos cereais mais importantes do mundo, com produção mundial superior a 780 milhões de toneladas e consumo por metade da população mundial. A produção brasileira é estimada em 12 milhões de toneladas, o que coloca o Brasil como o único país não asiático a estar entre os 10 maiores produtores, e 70% da produção se concentra no Rio Grande do Sul [3].

Como em todo cultivo agrícola, o custo de produção do arroz é dependente de insumos, como área plantada, fertilizantes, defensivos, diesel, entre outros, e é afetada por fatores climáticos, tais como precipitação, insolação e temperatura. A oscilação no preço dos insumos e variações climáticas podem, portanto, impactar a produção, a oferta e, consequentemente, o preço do arroz, causando problemas econômicos e sociais importantes, como redução do poder de compra dos consumidores ou redução da renda no campo. Desta forma, possibilitar melhor previsão da variação de preço do arroz pode trazer benefícios sociais e econômicos ao permitir melhor planejamento dos envolvidos.

Diversos pesquisadores já usaram modelos computacionais de aprendizagem de máquina distintos para tentar prever o preço de commodities. Geralmente esses estudos usam o histórico de preços disponíveis no mercado financeiro além de outros dados de interesse que possam justificar a oscilação do preço [2]. O objetivo desse estudo foi testar diferentes modelos computacionais clássicos de aprendizagem de máquina, amplamente utilizados em estudos de previsão de preço, para prever o preço do arroz no Rio Grande do Sul usando o histórico do preço desta commoditie, assim como no histórico de dados climáticos da região, como temperatura e precipitação, e de insumos necessários, como área de plantio, produção, estoque inicial, e custo de adubos, calcário e diesel.

[1]lucas.mielke@usp.br
[2]paulino.villas-boas@embrapa.br

2 Materiais e Métodos

Base de dados

Para elaboração do modelo proposto, foram coletados preços de arroz, para uso como variável preditiva, e dados de insumos agrícolas e de clima, para uso como variáveis explicativas. A variável preditiva foi obtida a partir de pesquisa de preço à vista do arroz no Rio Grande do Sul na base de dados do Centro de Estudos Avançados em Economia Aplicada (CEPEA) da Universidade de São Paulo (USP).

Para variáveis explicativas de insumos agrícolas, obtiveram-se custos médios de fertilizantes e de calcário, e dados de estoque e de consumo nacional da commoditie na base de dados da Companhia Nacional de Abastecimento (CONAB) [1]. Coletou-se o preço do Petróleo Brent na API da bolsa de NASDAQ (National Association of Securities Dealers Automated Quotations, em inglês) e preço futuro e volume negociado do arroz na bolsa de Commodities de Chicago (CBOT, Chicago Board of Trade em Inglês). Obtiveram-se a produção do arroz e sua área de cultivo para a região estudada utilizando a Pesquisa Agrícola Municipal (PAM) realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Para representar as variáveis explicativas de clima, foram obtidos os dados de temperatura e de precipitação de estações meteorológicas da região, coletados pelo Instituto Nacional de Meteorologia (INMET). Também foi coletado a cotação do dólar comercial do Banco Central (BACEN) para conversão dos dados de Combustível e de Arroz Futuro de dólar para reais. Todos os dados foram coletados para o período entre Junho de 2010 e Maio de 2022, totalizando 12 anos.

Ajuste e análise dos dados

Os dados de Petróleo e Preço Futuro de Arroz foram convertidos de Dólar para Reais utilizando o dólar médio do período. Os dados temperatura e chuva da região foram obtidos para várias estações meteorológicas do estado e interpolados para uma região central do mesmo, pelo método do inverso da distância. Todos os dados foram estruturados em uma base única com frequência semanal para todas as variáveis exceto Área, Consumo, Produção e Estoque, que estavam em frequência anual; e Preço de Adubos e Calcário, que estavam em frequência mensal. Em seguida todas variáveis foram convertidos para diferença absoluta com o período anterior (diferenciação), e foram calculadas variáveis extras para trazer médias móveis, variações contra períodos anteriores e identificação de outliers. A Tabela 1 resume todos os dados e tratamentos.

Distribuição dos dados e Aprendizagem de Máquina

Neste trabalho, usamos o teste de Shapiro-Wilk para avaliar se a distribuição dos preços do arroz é normal. Neste teste, a hipótese nula é que a população possui distribuição normal, e a alternativa, o contrário e o valor-p inferior a 0,05.

A situação do estudo pode ser expressa na forma de equação onde temos uma variável Y de interesse que pode ser explicada por vetores X, chamados de variáveis explicativas ou independentes da forma que ocorre com um modelo linear clássico. Os 12 anos de amostras foram divididos em bases de 4 anos cada e analisados separadamente usando os 85% dos dados iniciais para treino e 15% dos dados finais para teste.

Os modelos testados foram Regressão Linear Múltipla, Regressão de Cume (Ridge Regression, em inglês), Regressão Árvore de Decisão, Extreme Gradiente Boosting e finalmente SARI-

Variável	Tipo	Fonte	Frequência Original	Frequência ajustada	Ajustes
Preço Arroz	Dependente	CEPEA	Diária	Semanal	
Temperatura	Independente	INEP	Horária	Semanal	Interpolação
Chuva	Independente	INEP	Horária	Semanal	Interpolação
Preço Petróleo	Independente	NASDAQ	Diária	Semanal	Ajuste para Reais
FX	Independente	BACEN	Diária	Semanal	
Produção	Independente	IBGE	Anual	Anual	
Área Produtora	Independente	IBGE	Anual	Anual	
Estoque Inicial	Independente	CONAB	Anual	Anual	
Consumo Nacional	Independente	CONAB	Anual	Anual	
Preço Arroz Futuri	Independente	CBOT	Diária	Semanal	
Volume Negociado Arroz Futuro	Independente	CBOT	Diária	Semanal	
Preço Adubos	Independente	CONAB	Mensal	Mensal	
Preço Calcário	Independente	CONAB	Mensal	Mensal	
Média Movel de Cada Variável (4, 8 e 12 semanas)	Independente	Calculada	Semanal	Semanal	Cálculo
Valor anterior de cada variável (1 a 6 semanas)	Independente	Calculada	Semanal	Semanal	Cálculo

Tabela 1: Descrição das variáveis utilizadas.

MAX (Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors, em inglês). Os 4 primeiros modelos foram treinados e testados 2 vezes para cada período, para avaliar todas as variáveis independentes num primeiro momento e, em seguida, avaliar apenas com as melhores variáveis após Eliminação Recursiva de Atributos (RFE, Recursive Feature Elimination em inglês).

A comparação de predição dos modelos descritos foi feita pelas medições dos Erro Absoluto Médio (MAE - Mean Absolute Error, em inglês), Erro Quadrático Médio (MSE - Mean Squared Error, em inglês), Raiz quadrada do erro-médio (RMSE - Root Mean Square Error, em inglês) e Coeficiente de Determinação R^2.

3 Resultados e discussão

O Teste de Shapiro-Wilk obteve valor-p inferior a 0,05 indicando que a variável a ser predita não segue uma distribuição normal. Um teste adicional ajustou a distribuição dessa variável dependente para a distribuição Cauchy com localização 0.0343 e escala 0.2614.

Ao se comparar todos os modelos em todos os períodos, verificou-se as Menores Raízes Quadradas dos Erros Médios (RMSE) para os modelos de Ridge Regressor, com e sem seleção de variáveis, e SARIMAX conforme resultados apresentados na Tabela 2.

RMSE	Período			
	05/2022-06/2018	05/2018-06/2014	05/2014-06/2010	06/2010-05/2022
Ridge Regressor RFE	0,83	0,19	0,24	1,15
Ridge Regressor	3,26	1,20	1,32	1,42
Sarimax	3,42	1,20	1,70	1,50
Regressão	3,32	1,25	1,74	1,52
Extreme Gradient Boosting RFE	1,00	0,19	0,23	1,53
Decision Tree Regressor	1,21	0,24	0,27	1,54
Decision Tree Regressor	1,21	0,24	0,27	1,54
Extreme Gradient Boosting	1,00	0,19	0,22	1,60
Regressão RFE	1,41	0,36	0,43	2,05

Tabela 2: RMSE dos modelos por período analisado.

Dessa forma, verifica-se que esses dois modelos descreveram de forma mais razoável a variável

preditiva em função das variáveis explicativas, o que indica que esses modelos desenvolvidos são promissores. Entretanto, notamos que as variáveis explicativas não captaram toda a variabilidade da variável preditiva, necessitando, portanto, incluir outras variáveis explicativas no modelo. Faltou ainda realizar uma análise de resíduos para encontrar possíveis observações aberrantes. Pretendemos também realizar uma validação cruzada para avaliar a capacidade preditiva do modelo gerado e comparar os resultados com os de outros modelos de aprendizado de máquina, tais como, máquina de suporte de vetores e redes neurais.

4 Conclusões

Neste trabalho, concluímos que a variação do preço do arroz foi melhor ajustada com os modelos Ridge Regressor e SARIMAX para as variáveis independentes obtidas. Entretanto, é necessário explorar outras variáveis e realizar análise de resíduos a fim de melhorar a capacidade de descrever a variável preditiva 'preço'.

Referências

[1] Companhia Nacional de Abastecimento 2021. *Série de Custos de Arroz*. Disponível em: https://www.conab.gov.br/info-agro/custos-de-producao/planilhas-de-custo-de-producao/itemlist/category/791-arroz.

[2] L. S. Nunes. Previsão de indicadores diários de preços no mercado futuro de commodities agrícolas utilizando aprendizagem de máquina. 2020. 155 f. Dissertação (Mestrado em Estatística Aplicada e Biometria) - Universidade Federal de Alfenas, Alfenas, MG, 2020. Disponível em: https://bdtd.unifal-mg.edu.br:8443/handle/tede/1762

[3] F. S. Silva et al. Cultivo Arroz - Estatística de Produção. Embrapa. 2021. Disponível em: https://www.embrapa.br/estatistica-de-producao.