Café e tecnologia: Seleção de métodos para projeção do preço da saca de café

Ramon Ferreira Cruz[1]*; Daniel Alvares Firmino[2]

[1] Especialista em Finanças Corporativas, Bacharel em Administração de Empresas. Rua Henrique de Carvalho, 199 – Bota Fogo; 37190-00 Três Pontas, Minas Gerais, Brasil
[2] Mestre em Economia, Bacharel em Ciências Econômicas. Av. Pádua Dias, 11 – Agronomia; 13418-900 Piracicaba, São Paulo, Brasil
*autor correspondente: ramonfer.cruz@gmail.com

Café e tecnologia: Seleção de métodos para projeção do preço da saca de café

Resumo

O Brasil é o principal produtor de café no mundo, e a sua cadeia produtiva gera forte impacto na economia nacional, porém o produto tem sofrido grandes variações de preço nos últimos anos, o que nem sempre é benéfico para o produtor rural. Estas variações geram incertezas no mercado cafeicultor. Com base neste cenário, foi realizado um estudo de três variáveis: dólar, inflação e preço do diesel, comprovando o impacto destes fatores no preço do café, por meio de correlação e teste de causalidade. Por fim, foi efetuada a criação de modelos que possibilitem a previsão de valores futuros utilizando as variáveis exógenas analisadas com o objetivo de criar ferramentas que auxiliem a tomada de decisão por parte dos agentes atuantes no mercado produtor de café. Foi testado a utilização de redes neurais em comparação a abordagem de estatísticas tradicionais, o ARIMA e modelos de aprendizado de máquina, random forest e xgboost. Por meio da validação cruzada foi possível concluir que as redes neuras apresentaram bons resultados, porém o modelo ARIMA se mostrou mais estável e constante entregando os melhores resultados.

Palavras-chave: Series Temporais; ARIMA; Validação Cruzada; Aprendizado de Máquina, redes neurais.

Introdução

O Brasil é o maior produtor de café do mundo responsável por 1/3 da produção mundial, o produto além de exportado é amplamente consumido pela população local, sendo o segundo maior consumidor de café no mundo (Nogueira e Neves, 2015).

O plantio do grão acontece em 15 estados brasileiros, por mais de 287 mil produtores, sendo Minas Gerais o maior produtor correspondendo a 50% da produção nacional (Nogueira e Neves, 2015).

A cadeia produtiva do café em 2017 foi responsável por 30,7 bilhões de reais do PIB brasileiro, sendo deste valor 34,8% referente ao café em grãos, 10,5% ao café beneficiado, 47,1% ao comercio e serviços e 7,6% aos insumos. O produto possui importante papel na geração de empregos no país, em 2017 abrangendo 694.507 postos de trabalho com uma remuneração média anual de R$ 17.954, muito superior à média nacional que era de 2.973 no ano analisado (Sesso et al, 2021).

O café faz parte de cesta básica nacional e seu preço influencia diversos indicadores de inflação, como IPCA, IGP-M e seu seus derivados. Estes indicadores são utilizados para atualização do preço de tarifas públicas, contratos de aluguel, remuneração de ativos, entre outros.

Apesar de sua grande importância e impacto socioeconômico, uma das principais fraquezas da cafeicultura brasileira é entender o mercado interno e externo de café, estas fraquezas podem ser divididas em: produção, marketing, barreiras tarifárias, logística, governança, taxas cambiais e outros. Existem diversos problemas que agravam estas fraquezas, porém o principal destaque fica para volatilidade de preços do produto, causados

pela carência de gestão de risco dos agentes de café (nacionais e internacionais) (Nogueira e Neves, 2015).

Nos últimos anos a cotação do café tipo arábica sofreu expressiva valorização, saindo de R$ 493,03 em janeiro de 2020 e chegando a 1.112,9 em dezembro de 2022, uma variação de 105% no período, contra uma variação de 6% em um período igual anterior de janeiro de 2017 a dezembro 2019.

O preço do café pode ser influenciado por fatores como a demanda, estoque, variações climáticas, ciclos de produção, especulações em bolsas de valores e mercados de café. Entender essas e outras variações permite ao produtor criar estratégias de mitigação de risco, seja através de estratégias hedge na compra ou venda de café, contratos de longo prazo, contratação de seguros, entre outros (Nogueira e Neves, 2015).

O aumento de preços do agronegócio nem sempre beneficia o produtor rural, ou seja, a cotação do café pode ser alterada por fatores que não ensejam em ganho de capital, fatores como custo de produção (mão de obra, agroquímicos, maquinário, matéria prima e etc.), volumes de perdas, clima, pragas, doenças, mercado externo, cambio, preço do diesel e no longo prazo a produtividade. Esses fatores são somados a expectativa de inflação futura, ou seja, quanto maior a expectativa no aumento dos preços no longo prazo, maior a pressão por ajustes no preço observado (CEPEA 2020).

Dito isso, este trabalho se dedicou em entender quais fatores influenciam a cotação da saca de café, descrever o comportamento da série e por fim, propor diferentes modelos de natureza supervisionada para previsão de valores futuros, assim criando ferramentas de apoio a tomada de decisões e criação de estratégias por parte dos agentes atuantes neste mercado.

Material e Métodos

Estrutura da pesquisa

O presenta trabalho foi desenvolvido seguindo as seguintes etapas, coleta da série temporal da variável principal, coleta das series temporais variáveis exógenas, analise da série temporais e suas principais características, avaliação da correção e causalidade entre as variáveis em ralação a variável alvo, transformações nos dados de acordo com as necessidades de cada estimador utilizados, otimização dos parâmetros dos modelos, avaliação dos resultados por meio do método de validação cruzada e pôr fim, a escolha do melhor modelo.

Fonte de dados

Os dados a serem utilizado neste trabalho, correspondem ao preço em reais por saca de 60 kg líquido com as seguintes características: bica corrida, tipo 6, bebida dura para melhor. Os dados foram coletados no site do Centro de Estudos Avançados em Economia Aplicada [CEPEA] mantido pela Escola Superior de Agricultura Luiz de Queiroz [ESALQ] da Universidade de São Paulo [USP]. As variáveis exógenas escolhidas foram as séries históricas da cotação mensal do dólar, índice de preço ao consumidor amplo [IPCA] acumulado em 12 meses [IPCA 12M] e o preço médio mensal do diesel.

Segundo o CEPEA (2020) a inflação no Brasil é medida pelas variações IPCA, estas variações, influenciam na expectativa futura, deste modo, afetando os preços atuais dos produtos agrícolas. Dentre os diferentes cálculos de variações do IPCA, foi escolhido o acumulado em 12 meses devido a sua maior correlação de Pearson com a cotação do café arábica, sendo de 0,50, em comparação com as demais métricas e suas correlações, IPCA mensal 0,14, IPCA acumulado em 3 meses 0,21, IPCA acumulado em 6 meses 0,33 e o IPCA do ano com 0,20. As séries históricas foram coletadas no site do Instituto Brasileiro de Geografia e Estatística [IBGE].

Ainda segundo o CEPEA (2020), alteração no preço do óleo diesel, combustível utilizado no transporte de diversas mercadorias agrícolas e nos maquinários de sua produção, podem ocasionar em variações de preços destes produtos, incluindo o café. Para representar esta variável, foi calculado o valor médio mensal do preço do diesel com base nos dados históricos de comercialização do diesel em diferentes postos no país. Estes dados foram disponibilizados pela Agência Nacional do Petróleo, Gás Natural e Biocombustíveis [ANP] no site governamental.

A série histórica da variável dólar mensal foi coletado no site do CEPEA. Para o CEPEA (2020) a moeda que é referência no comercio internacional, afeta os custos da produção do agronegócio, sendo utilizada diretamente na comercialização de maquinários, defensivos, fertilizantes, entre outros ou alterando diretamente o preço do produto agrícola. No caso deste estudo, o café, devido a sua indexação com o mercado internacional.

Series Temporais

O registro temporal das cotações do preço da saca de café é um exemplo de série temporal. Os objetivos em se analisar series temporais, podem ser: realizar previsões de valores futuros da série, descrever o seu comportamento, procurar periodicidades relevantes e/ou entender os mecanismos geradores da série (Morettin e Toloi 2019).

Séries temporais são um conjunto de observações ordenadas ao longo do tempo, podendo ser descrita em um modelo de decomposição que descreve Z(t), como a soma de três componentes não observáveis (Morettin e Toloi 2019), ou em quatro componentes (Cesar e Rossi 2014).

$$Z_t = T_t + S_t + C_T + a_t \quad (1)$$

Em que:

T = tendência

S = sazonalidade

C = Ciclos

a = Ruído

Tendência (T) de forma mais simples, significa a flutuação da série em torno de uma reta com inclinação positiva ou negativa. (Morettin e Toloi 2019).

Sazonalidade é um componente de curto prazo associado a variações provocadas pelas épocas do ano (meses quentes ou frios, épocas de festas etc.), afetando as séries temporais de produtos como refrigerantes, roupas, ar-condicionado, brinquedos, transportes, alimentos, entre outros (Cesar e Rossi 2014).

Resíduo é o componente que não se pode explicar, variável aleatória também designada por “ruído” (Cesar e Rossi 2014).

Ciclo é o componente de médio prazo associado a variações de conjuntura econômica, com períodos de expansão e de recessão (Cesar e Rossi 2014).

Em geral a sazonalidade St e a tendência Tt estão bastante relacionadas, existindo forte influência da tendência sobre o componente sazonal, portanto, não podemos isolar um componente sem tentar isolar outro.

Além de seus componentes, as series temporais podem apresentar as seguintes características, pode se discretas ou continuas, estacionarias ou não, univariada ou multivariada, unidimensional ou multidimensional e estocástica ou determinística.

Para analisar uma série temporal é necessário amostra-la, seja de forma continua onde os eventos são registrados no momento de seu acontecimento sem uma regularidade de tempo, ou de forma discreta em intervalos de tempos igual. Uma série discreta pode ser obtida a partir de uma série continua, ou agregando valores em intervalos de tempos iguais (Morettin e Toloi 2019).

No geral, o parâmetro t é utilizado se referindo ao tempo constituindo-se de um vetor onde p = 1, sendo considerado unidimensional, no entanto este poderá se referir a outros parâmetros físicos, como espaço e volume, sendo por exemplo tempo, latitude e longitude, neste caso p = 3, sendo multidimensional (Morettin e Toloi 2019).

Um processo estocástico, refere-se a um conjunto de variáveis aleatórias, em que existe uma suposição de estarem contidas em um mesmo espaço de probabilidades, portando quando analisamos uma série temporal, por exemplo, de dados econômicos, estamos avaliando uma trajetória dentre as diferentes possíveis. Em algumas situações é possível realizar experimentos, como na Oceanografia, podemos simular as diferentes trajetórias possíveis (Morettin e Toloi 2019).

Além destes fatores, uma das suposições que se faz a respeito de uma série temporal na maioria dos procedimentos estatísticos é a que ela é estacionária, ou seja, ela se desenvolve no tempo aleatoriamente ao redor de uma média constante. Entretanto em sua maioria as séries temporais encontradas no mundo real é não estacionaria, assim sendo necessário a realização de transformações. A principal delas é a de diferenças sucessivas. (Morettin e Toloi 2019).

Em virtude disso, para a definir se as séries utilizadas são ou não estacionárias, seguiu-se o recomendado por Morettin e Toloi (2019), a respeito da realização do teste estatístico de raiz unitária de Dickey-Fuller aumentado.

Teste de Causalidade de Granger

Morettin e Toloi (2020) apontam, que definir a causalidade entre variáveis é um dos grandes problemas em um estudo. Buscando solucionar esse problema seguiu-se o proposto por Morettin e Toloi (2020) e o teste de causalidade de Granger foi aplicado no intuito de entender se uma variável X causa a variável Y, Definindo a causalidade em termos de previsibilidade, de forma a determinar se a inclusão dos valores passados de uma variável X torna a previsão de uma variável Y atual mais eficiente.

Aprendizado de máquina

Aprendizado de máquina é um seguimento da inteligência artificial. Dentro de aprendizado da máquina existem três abordagens: aprendizado supervisionado, não supervisionado e por reforço. O aprendizado supervisionado, consiste na ideia da existência de um supervisor externo, onde os dados processados possuem um rótulo indicando qual é a resposta esperada para este conjunto de observações. O aprendizado não supervisionado, se caracteriza pela ausência deste supervisor externo e rótulos, o que normalmente é atribuído a tarefas de busca de padrões, agrupamento dos dados e ou regras de associação. O aprendizado por reforço, busca apresentar uma recompensa para determinada ação

positiva em detrimento de punição para uma ação negativa, um exemplo deste tipo de aprendizado é ensinar um robô a menor trajetória entre dois pontos (Faceli et al 2021).

A utilização de métodos de aprendizado de máquina para previsão de séries temporais é um tema relativamente novo, mas que tem se mostrado promissor, no entanto, como estes modelos não foram criados para esta finalidade, um importante passo é a geração de características que tragam essa noção temporal para o modelo (Nielson 2021).

Dito isso foram implementados os métodos de aprendizado supervisionado random forest e gradiente boosted, redes neurais com uma abordagem de aprendizado profundo e modelo ARIMA representando a estatística tradicional. Abaixo segue um aprofundamento em cada um deles.

Random Forest

Random Forest é uma técnica que combina a predição de diversas árvores de decisão, onde é utilizando porções aleatórias das variáveis da amostra de treinamento de forma independente, porém com a mesma distribuição para treinar as diferentes árvores. A capacidade de generalização resultante, dependerá da qualidade individual de cada árvore utilizada no modelo e a correlação entre elas. Random forest são especialmente indicadas em casos que existe muitas variáveis de entrada, nesta situação uma única árvore de decisão produziria um resultado insignificativamente melhor do que uma escolha aleatória de classes (Breiman 2001). A utilização desta técnica reduz a tendência a super adequação aos dados provenientes das árvores de decisão (HARRISON, 2020).

Gradient boosted

Esta é uma forma de combinar diversos modelos em sequência, de forma que os modelos posteriores corrijam os erros dos modelos anteriores, assim como a realização da ponderação de dados desajustados em modelo posteriores. O modelo de gradiente a ser implementado será o xgboost, neste método, o primeiro estimador realiza uma previsão, e os demais tentam prever o resíduo entre valor previsto e o valor real, realizando esse processo em sequência o modelo busca minimizar os resíduos totais (Nielson 2021).

Modelo ARIMA

Também conhecido como abordagem de Box e Jenkins (1976), o autoregressive integrated moving average [ARIMA] tem o objetivo de ajustar modelos autorregressivos combinados a médias moveis para estimar uma determinada série temporal (Morettin 2018).

A construção de um modelo ARIMA pode ser dívida em etapas cíclicas: consideração de uma classe de modelos para análise, identificação do modelo a partir da avaliação das autocorrelações e autocorrelação parcial, estimação dos parâmetros e o diagnóstico do modelo criado. Caso o modelo não seja adequado o processo deverá ser repetido (Morettin 2018).

O processo de identificação passará pela definição dos parâmetros p, d e q, onde d corresponde ao número de diferenciações necessárias para torna a séries um processo estacionário. p será a ordem do modelo autorregressivo [AR], q é a ordem do modelo de média móvel [MA] (Morettin 2018).

Modelos do tipo AR partem do pressuposto que os valores passados ajudarão a predizer os valores futuros. O modelo mais simples, o AR de ordem p = 1, pode ser representado pela seguinte equação:

$$y_t = b_0 + b_1 \times y_{t-1} + e_t \quad (2)$$

Onde o valor de y no tempo t será dado por uma constante b_0 , somado ao valor de y no tempo anterior y_{t-1} , multiplicado por uma constante b_1 , somado a um termo de erro e. O termo de erro deverá possuir uma variância constante e média igual a 0, portanto, o valor atual no processor AR o y_t, será obtido através dos valores mais recentes de p , assim existindo um processor $\mathrm{AR(p)}$, representando os coeficientes autoregressivos por ϕ , temos a seguinte equação:

$$y_t = \phi_0 + \phi_1 \times y_{t-1} + \phi_2 \times y_{t-2} + \cdots + \phi_p \times y_{t-p} + e_t \quad (3)$$

Um modelo MA pressupõe que um valor y em um instante t é uma função de termos de erro do passado recente, sendo independente dos demais valores. Expresso como:

$$y_t = \mu + e_t + \theta_1 \times e_{t-1} + \theta_2 \times e_{t-2} + \cdots + \theta_q \times e_{t-q} \quad (4)$$

Unificando os dois conceitos obtemos o modelo ARMA, que poderá ser representado na equação 4. Obteremos o modelo ARIMA incluindo as diferenciações necessárias para tornar a serie estacionaria.

Os modelos ARIMA são úteis em descrever o comportamento de séries sociais e econômicas, ondes os erros encontrados influenciam na evolução do processo e são auto correlacionados (Morettin e Toloi 2019).

Morettin e Toloi (2019) sugerem que a comparação entre diferentes modelos da família ARIMA, pode ser realizada utilizando o critério de Informação de Akaike [AIC], critério de

informação bayesiano [BIC] e/ou logaritmo da verossimilhança. Nielsen (2021) recomenda a utilização da análise gráfica da autocorrelação e autocorrelação correlação parcial em comparação com AIC para identificar a ordem de um modelo da família ARIMA.

Redes Neurais Artificiais

Redes neurais artificiais [RNA] tem o objetivo de tentar simular o funcionamento do celebro humano, como um sistema de processamento de informações, tendo os neurônios como sua unidade de processamento individual realizando a interligação das células computacionais (HAYKIN 2021).

As redes neurais foram criadas pelo neurofisiologista Warren McCulluch e pelo matemático Walter Pitts em 1943, o trabalho apresentado por eles criou uma simplificação computacional de um neurônio biológico (GERON 2019).

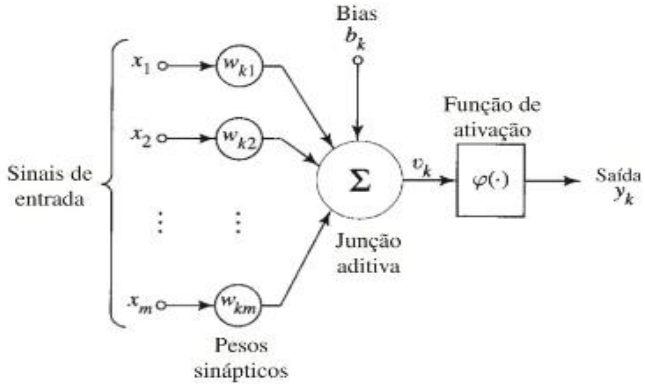

Figura 1: Modelo não-linear de um neurônio
Fonte: HAYKIN 2001

O neurônio é o cerne da rede neural. A Figura 1 apresenta o modelo de um único neurônio, ele possui um conjunto de sinapses ou elos de conexão, cada um caracterizado por um peso ou força própria. Onde um sinal de entrada x na entrada j é conectado ao neurônio k e multiplicado pelo peso sináptico w. Em seguida será realizado a soma dos sinais de entradas, ponderados pelas respectivas sinapses do neurônio. Após este processo é aplicada uma função de ativação cujo objetivo é restringir a amplitude do neurônio, o modelo incluirá o bias b, que tem o efeito de aumentar ou diminuir a entrada liquida da função de ativação. (HAYKIN 2001).

A rede neural proposta, pode ser presentada pela equação a seguir.

$$y_k = g\left(\sum_{i=0}^{n} (y_i w_{ki}) + b_k\right) \quad (5)$$

Em que:

g = Função de ativação

y_i = Saída calculada pelo neurônio i

w_{ki} = Peso sináptico entre o neurônio i e o neurônio k

b_k = bias

Neste trabalho foi utilização a implementação de uma rede neural recorrente com arquitetura long short-term memory [LSTM]. Redes neurais recorrentes são um método abrangente onde os parâmetros são aplicados repetidamente, independentemente dos valores das entradas mudaram ao passar do tempo, já a arquitetura LSTM são recomendadas para resolver os problemas de fuga e explosão de gradiente (NIELSEN 2021).

Validação Cruzada

Bergmeira et. al (2017) aponta que a validação cruzada é uma técnica já consolidada para problemas de regressão e classificação, cujo o objetivo é avaliar a generalização de um modelo, assim evitando a super adequação ao conjunto de dados, provenientes aos parâmetros do modelo. O autor defende que esse método deve ser utilizado para a avaliação de modelos de séries temporais, seja eles, auto regressivos ou de aprendizado de máquina. A abordagem proposta, conforme Figura 2, consiste em dividir os dados em 5 conjuntos de treino e teste levando em consideração o fator temporal da série, calcular as métricas de erro e avalia-las em conjunto.

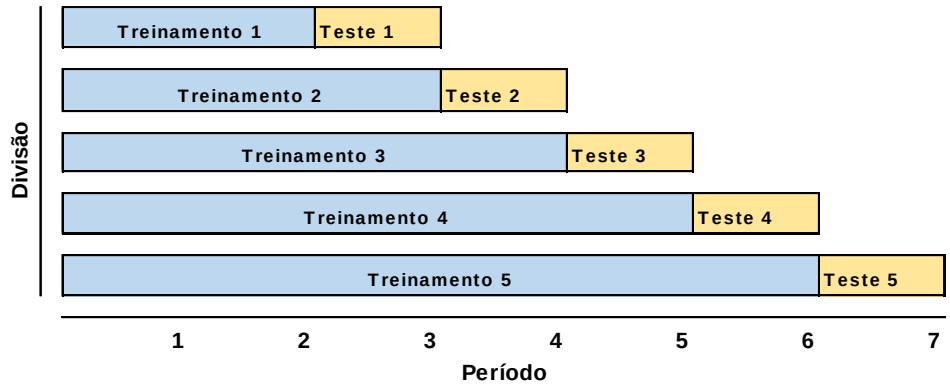

Figura 2: Divisão do treino e teste da validação cruzada

Fonte: Dados originais da pesquisa

Métricas de avaliação

Para comparação final dos modelos foram utilizadas as métricas de erro de regressão, mean squared error [MSE], mean absolute percentual error [MAPE], mean absolute error [MAE], root mean score error [RMSE] e mean squared logarithmic error [MSLE].

O MSE, consiste em calcular o erro quadrático médio. O primeiro passo é realizar o cálculo da diferença ao quadrado entre os valores esperados e os valores estimados de cada amostra, com o resultado desta operação é realizado o cálculo do valor médio. Esta abordagem busca penalizar erros maiores em detrimento de menores erro.

$$MSE(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2 \quad (6)$$

Em que:

y = valor esperado

$\hat{y}$ = valor estimado

n = tamanho da amostra

A métrica MAE, consiste em calcular o erro médio absoluto. Nesta operação obtemos o modulo da diferença entre os valores esperados e os valores estimados das amostras e por fim, é calculado a média dos resultados obtidos.

$$MAE(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} |y - \hat{y}| \quad (7)$$

Em que:

y = valor esperado

$\hat{y}$ = valor estimado

n = tamanho da amostra

O MSLE, busca apresentar o erro logaritmo quadrático médio. A métrica é obtida calculando o quadrado da diferença entre o logaritmo natural na base dez do valor esperado e o logaritmo natural na base dez do valor previsto, por fim é obtido o valor médio desta operação. Esta métrica é indicada quando existe um crescimento exponencial do valor esperado.

$$MSLE(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} (\log_e(1 + y_i) - \log_e(1 + \hat{y}_i))^2 \quad (8)$$

Em que:

y = valor esperado

$\hat{y}$ = valor estimado

n = tamanho da amostra

Já a métrica RMSE, busca calcular a raiz quadrada do erro quadrático médio, ou seja, após obtermos o MSE, calculamos a raiz quadrada desta métrica.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2} \quad (9)$$

Em que:
y = valor esperado
$\hat{y}$ = valor estimado
n = tamanho da amostra

O MAPE, demonstra o erro percentual médio absoluto. O objetivo é apresentar o erro em termos percentuais, demonstrando um erro relativo que não é afetado por mudanças de escala. É calculado o valor da diferença entre o valor esperado e o valor estimado dividido pelo valor esperado de cada amostra, por fim é obtido a média de todas as amostras, caso o valor previsto de uma determinada amostra seja 0, o divisor é substituído por um número pequeno positivo, pois não é possível realizar divisão por zero.

$$MAPE(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} \frac{|y_i - \hat{y}_i|}{\max(e, |y_i|)} \quad (10)$$

Em que:
y = valor esperado
$\hat{y}$ = valor estimado
n = tamanho da amostra
e = Valor positivo pequeno próximo a zero.

As métricas apresentadas serão avaliadas em conjunto, respeitando suas indicações e limitações, porém privilegiando a métrica MAPE devido a possibilidade de se avaliar valores em escalas diferentes.

Resultados e Discussão

Decomposição e exploração do preço do café

A série temporal a ser analisada, consiste nos valores mensais da cotação da saca de café do tipo arábica vendido à vista no Brasil de janeiro de 2010 a julho de 2023. Ao realizar uma análise descritiva da série, pode-se extrair as seguintes informações: existem 163 observações, o valor máximo alcançado foi de R$ 1.485,35 em fevereiro de 2022, o valor mínimo de R$ 247,73 em novembro de 2013, o valor médio apresentado no período foi de R$ 554,78, com desvio padrão de R$ 289,18. O produto demonstrou grande valorização no ano de 2021, sendo comercializado inicialmente com um valor de R$ 639,71 e finalizado com um valor de R$ 1.471,15.

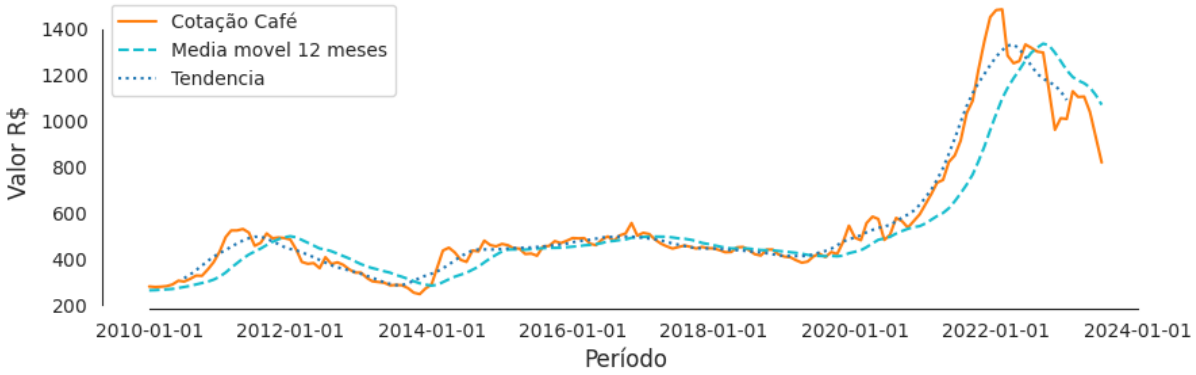

Figura 3: Tendência cotação do preço do café arábica
Fonte: Resultados originais da pesquisa

Analisando a tendência dos dados e considerando uma média móvel de 12 meses, disponíveis na Figura 7, pode-se observar algumas mudanças na inclinação da série. Primeiro uma valorização de 2010 a 2011, em seguida, uma queda no preço de outubro de 2011 a setembro de 2013, apresentando uma rápida valorização que se estabilizou em setembro de 2014. O preço continuou sofrendo oscilações sempre seguindo uma tendência, mas estabilizando em torno de R$ 450,00. A partir do ano de 2020, houve uma valorização exponencial, chegando na máxima em 2022 como já citado, após este ápice a cotação tem apresentado tendência de queda.

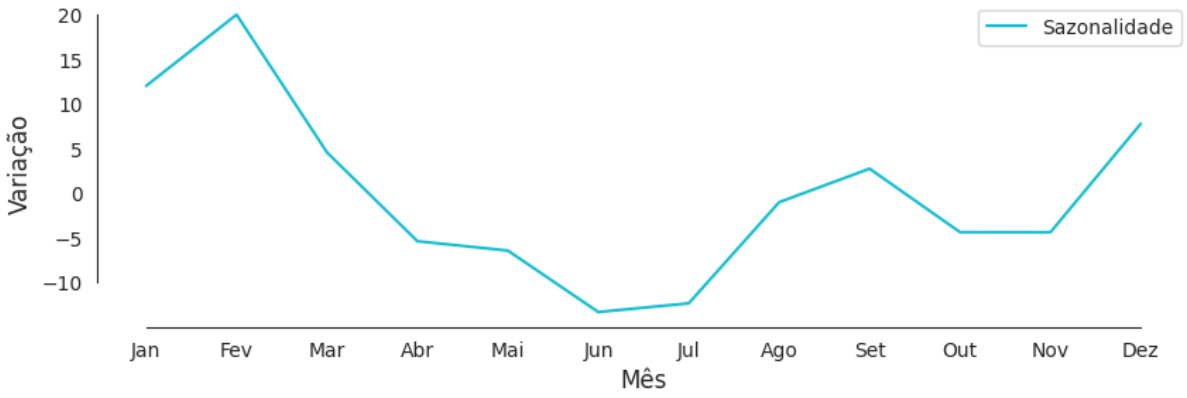

Figura 4: Sazonalidade cotação do preço do café arábica
Fonte: Resultados originais da pesquisa

Após efetuar a decomposição da sazonalidade, é observado que as variações mais significativas são: um período de alta da cotação nos meses de dezembro a março e uma queda nos preços de abril a julho, estas variações podem ser justificadas pelo período de colheita do produto, onde existe uma maior oferta dos grãos.

Após realizar o teste estatístico de Dickey Fuller, com base no valor de p de 0.867531, há evidencias para aceitarmos a hipóteses nula, portanto a série não possui comportamento estacionários, neste caso, a transformação mais comum recomendada é a de diferenças sucessivas (Morettin e Toloi 2019).

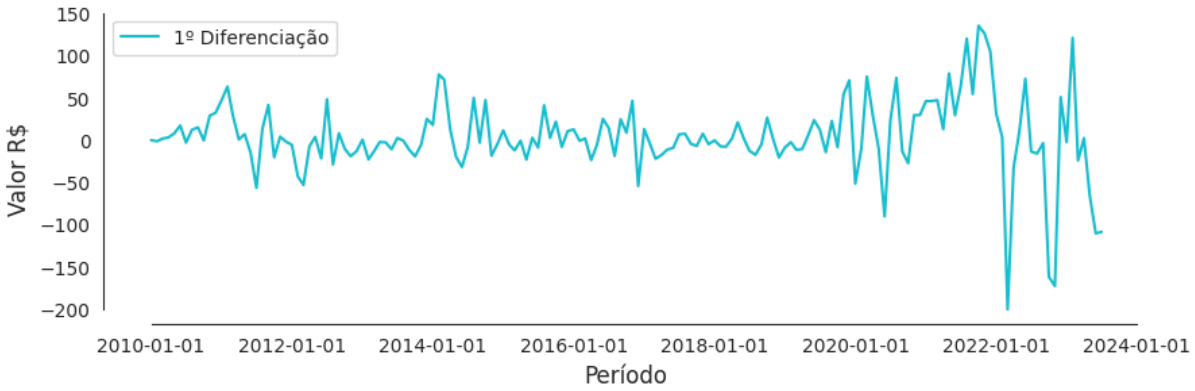

Figura 5: Serie com primeira diferenciação
Fonte: Resultados originais da pesquisa

Na Figura 5 é possível visualizar a série com a sua primeira diferenciação. Após este processo, o próximo passo é iniciar as análises gráficas das autocorrelações da série. Analisando a autocorrelação da série diferenciada, presente na Figura 6, foi identificada correlações relevantes dos dados na segunda, oitava e décima primeira janela, apresentando um aparente decaimento aproximando de zero. A análise gráfica indica que o valor do parâmetro q a ser utilizado no modelo ARIMA é 2.

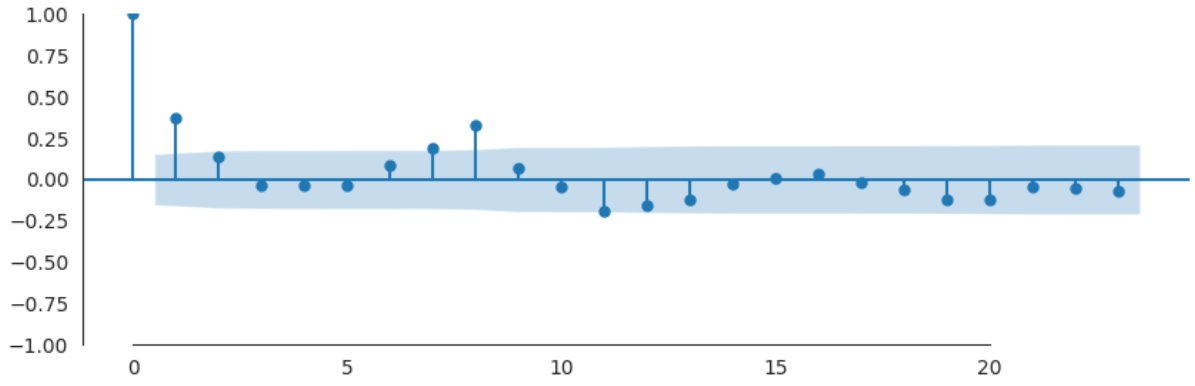

Figura 6: Auto correlação da série diferenciada
Fonte: Resultados originais da pesquisa

A Figura 7 apresenta o gráfico de autocorrelação parcial da série diferenciada, assim como na autocorrelação, observa-se correlações parciais relevantes dos dados na segunda,

oitava e décima primeira janela, tendendo a zero no logo prazo. A análise gráfica indica que o parâmetro de p do modelo ARIMA será 2.

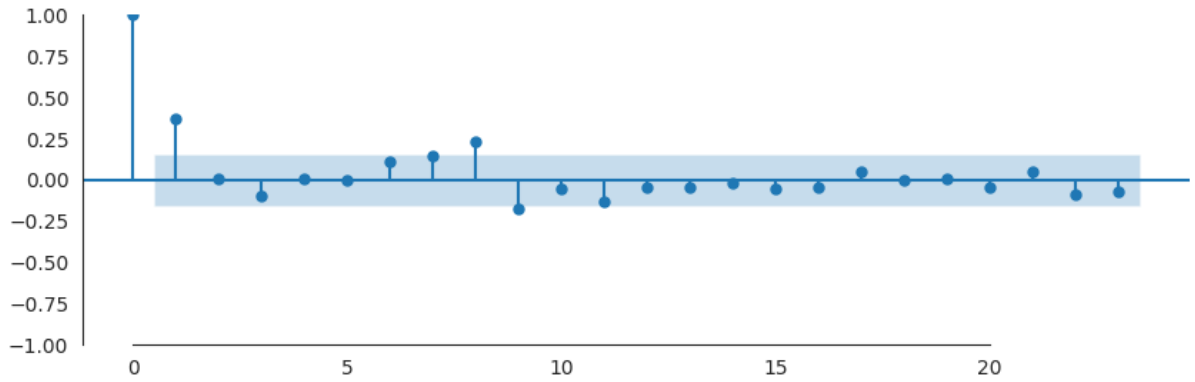

Figura 7: Auto correlação parcial da série diferenciada

Fonte: Resultados originais da pesquisa

Variáveis que afetam a cotação do café

Neste tópico foram analisadas as variáveis exógenas, diesel, dólar e IPCA 12M. Segundo referenciado anteriormente, estes itens possuem influência na cotação da saca de café tipo arábica. Identificamos uma forte correlação de Pearson com o preço do objeto de estudo, sendo diesel 0,89, dólar 0,71 e IPCA 12M 0,50.

Comparando o preço do café verso o dólar, encontramos uma correlação de Pearson significativa de 0.707098. Na Figura 8 é possível observar ambas as series e visualizar a correlação afirmada. Após dividir a cotação do café pelo dólar, com o objetivo de reduzir ou eliminar o efeito do dólar na cotação do café, devido a indexação de alguns produtos agrícolas ao dólar, é possível visualizar uma série de cotação do café muito mais estável ao longo do período, apesar da constante valorização da moeda estrangeira.

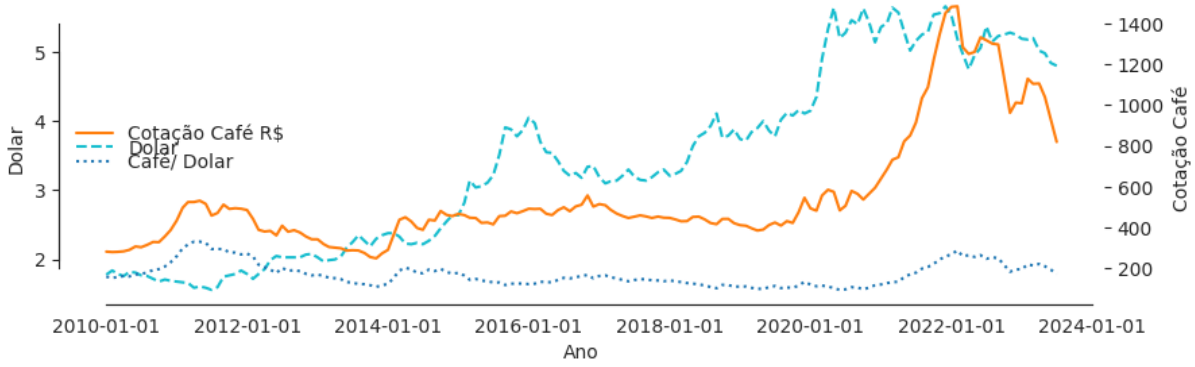

Figura 8: Comparativo, Café arábica x Dólar

Fonte: Resultados originais da pesquisa

A série temporal da cotação do dólar não é estacionária, conforme teste de Dickey-Fuller, obtivemos um valor de p de 0.828543, porém realizando a primeira diferenciação a série se torna estacionaria apresentando um valor de p de 3.534679e-14. Realizando o teste de causalidade de Granger a um nível de significância de 5%, pode-se rejeitar a hipótese nula e afirmar que a cotação do dólar ajuda a prever o valor do café com apenas um atraso (F test p=0.0255, chi2 test p=0.0228, likelihood ratio test p=0.0240).

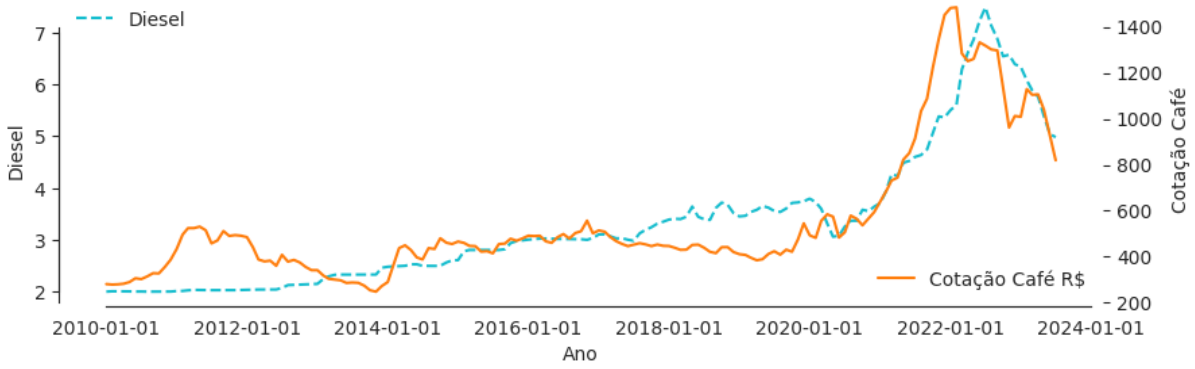

Figura 9: Comparativo café arábica e preço médio do diesel nos postos do Brasil

Fonte: Resultados originais da pesquisa

Analisando o preço do café frente ao diesel, encontramos uma correlação de Pearson significativa de 0,89. O preço médio do combustível tem evoluído ao longo do período e apresentado uma queda acentuada a partir de 2022. Como demonstrado na Figura 9, apesar de estarem em escalas diferentes ambas as séries possuem comportamento muito similares. Efetuando o teste de causalidade de Granger a um nível de significância de 5%, pode-se rejeitar a hipótese nula e afirmar que o preço do combustível ajuda a prever o valor do café com apenas 2 atraso (F test p= 0.0002, chi2 test p=0.0001, likelihood ratio test p=0.0001).

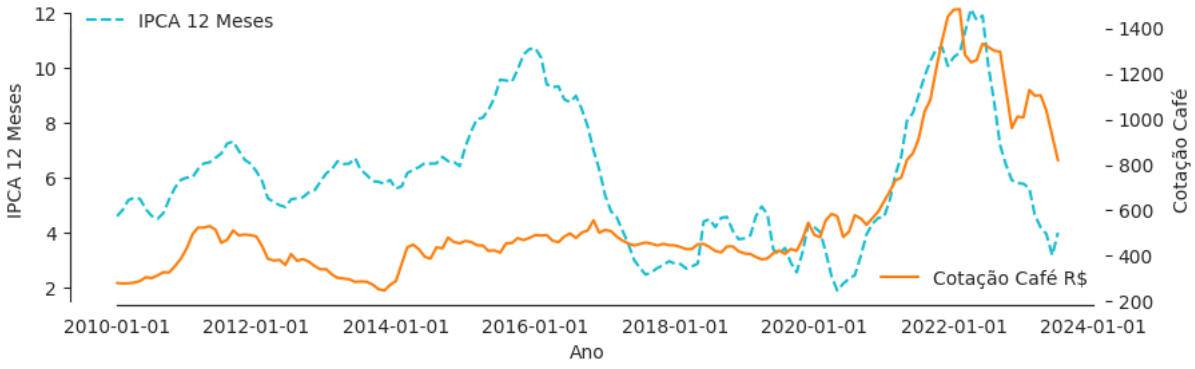

Figura 10: Comparativo café arábica e IPCA

Fonte: Resultados originais da pesquisa

Ao avaliar a Figura 10, é possível concluir que o IPCA 12M apresentava várias oscilações ao longo do período, com diversas mudanças de tendência, o que ainda sim, demonstra visualmente influência na cotação do café, existindo uma correlação de 0,50 entre as duas séries e ao realizar o teste de Granger a um nível de significância de 5%, pode-se rejeitar a hipótese nula e afirmar que o IPCA 12M ajuda a prever o valor do café com apenas 2 atrasos (F test p= 0.0182, chi2 test p=0.0143, likelihood ratio test p=0.0160).

Modelos

Foi realizado o treinamento de um modelo de redes neurais com arquitetura long short term memory [LSTM]. O modelo sequencial proposto, apresenta uma camada de entrada com 6 neurônios. A segunda camada foi implementada com 200 neurônios recorrentes, utilizando função de ativação tangente hiperbólica na camada de saída e função de ativação sigmoide no portão recorrente. A terceira camada possuem 200 neurônios recorrentes com função de ativação tangente hiperbólica e por fim, uma camada de um neurônio com a função de ativação linear. A otimização dos pesos da rede neural foi realizada através do método de otimização rmsprop e a métrica de erro mse. O treinamento foi iniciado com 1.000 épocas e o critério de parada do treinamento levou em consideração 5 épocas sem melhoria na métrica de erro com os pesos de cada neurônio sendo atualizados a cada 12 registro. A rede neural foi criada utilizando a biblioteca keras da linguagem python. Os dados utilizados no treinamento foram redimensionados para a escala de 0 e 1, porém para apresentação de gráficos e cálculos de métricas, foi utilizado a escala original. Na Figura 11, é possível avaliar o resultado da predição um passo à frente considerando os 10 meses finais do período analisado, esse foi o melhor ajuste encontrado, dentre todos os modelos.

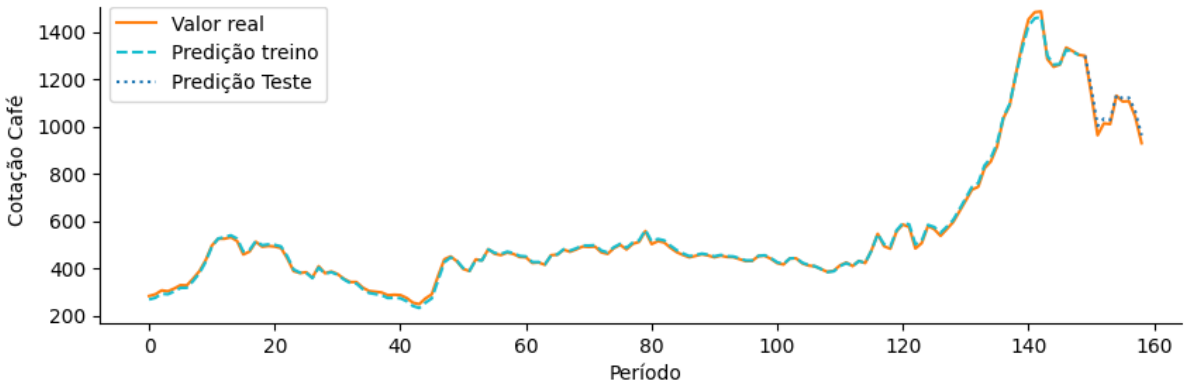

Figura 11: Previsões de teste LTSM

Fonte: Resultados originais da pesquisa

Em comparação a redes neurais, foi proposto outros modelos, com abordagem estatística tradicional ARIMA e modelos de aprendizado de máquina: random Forest e xgboost, afim de determinar o melhor resultado.

A respeito da abordagem estatística, apesar da análise gráfica apontar que o modelo adequado seria um SARIMAX com sazonalidade de 12 meses e os parâmetros de (2, 1, 2), ao realizar o teste de diversos parâmetros por meio de processo de otimização com o objetivo de minimizar a métrica AIC, o melhor conjunto de parâmetros encontrados foi de um modelo ARIMAX (1, 1, 0), obtendo as seguintes métricas de resultado: logaritmo da verossimilhança -750.410, AIC 1512.820 e BIC 1530.963. Os parâmetros se mostraram estatisticamente significantes, não existindo autocorrelação dos resíduos, conforme teste de Ljung-Box com valor de p de 0.94. É possível chegar a mesma conclusão observando o gráfico de correlação dos resíduos demonstrado na Figura 12.

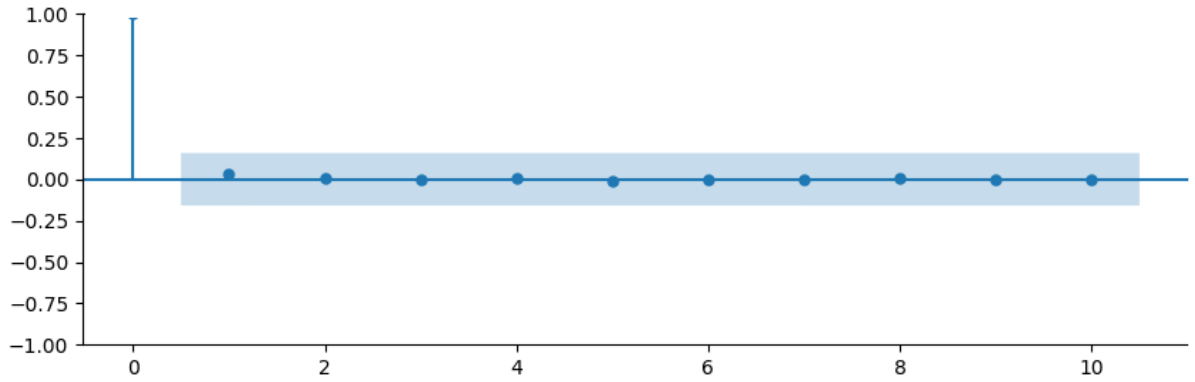

Figura 12: Autocorrelação dos resíduos modelo ARIMAX

Fonte: Resultados originais da pesquisa

Para os modelos de gradiente boosted, baseado na biblioteca xgboost, foram realizadas duas implementações, uma baseada em árvore de decisão e outra baseada em regressões lineares, ambas as técnicas foram aplicadas com os valores padrões de parâmetros sugeridos pela biblioteca contendo100 estimadores em sequência. A Random Forest foi implementada utilizando a biblioteca sklearn, sem realizar otimização de parâmetros, contendo 100 árvores de decisão. Para treinamento dos modelos de gradiente boosted e random forest, foram utilizados os 12 meses anteriores para prever o próximo passo, os dados foram transformados de modo que tenham média 0 e variância igual a 1, porém para cálculo de métricas e apresentação em gráficos foram utilizados a escala original dos valores.

Avaliação dos resultados

Para comparação dos modelos foi aplicado o método de validação cruzada, prevendo um passo à frente. Os dados foram divididos em 6 cortes, conforme apresentado na Figura 13, uma parte de treinamento e 5 partes para teste onde a parte anterior é acrescentada na base de treino, por exemplo para teste da parte 3, foram acrescidos na base de treinamento os recortes de dados 1 e 2. Deste modo podemos compreender a capacidade de generalização de cada modelo e seu comportamento nos recortes de dados.

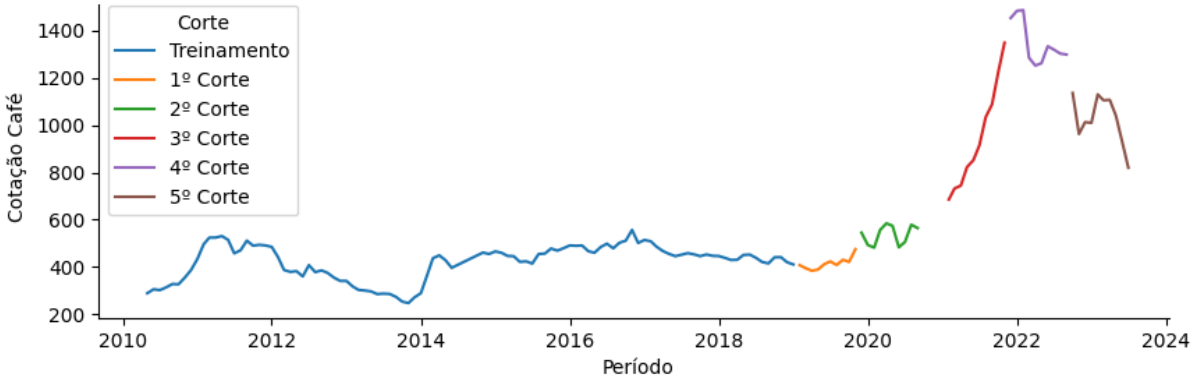

Figura 13: Divisão da validação cruzada
Fonte: Resultados originais da pesquisa

Avaliando a Tabela 2 em comparação com a Figura 13, observou-se que no primeiro recorte de dados, período mais estável utilizado para teste, temos bons resultados de todas as implementações, tendo o modelo ARIMAX o pior resultado, o xgboost linear obteve o melhor resultado, já a rede neural demonstrou um resultado mediano e similar aos demais que não foram citados.

No segundo recorte, período onde a várias oscilações, o modelo ARIMAX apresenta o melhor resultado, sendo a rede neural o segundo melhor e o pior resultado para Random Forest.

No terceiro recorte, período onde a uma forte aceleração dos preços, os modelos de aprendizado de máquina baseados em árvores apresentam péssimos ajustes, já os modelos lineares com xgboost linear e ARIMAX, se mostraram constantes, já o LSTM apresentou dificuldade na estimação dos valores.

No quarto recorte temos uma nova quebra estrutural, dando início a uma forte queda dos preços, ARIMAX apresenta seu melhor resultado e novamente o LSTM tem um resultado mediano.

No quinto recorte, período com fortes oscilações, porém com tendência a queda a rede neural demonstrou um erro muito baixo e um excelente ajuste no período, sendo a menor métrica de erro proporcional aos valores estimados encontrada em todos os recortes.

Tabela 2: Resultado dos modelos

Modelo	Rodada	MAPE	MSE	MSLE	MAE	RMSE
Random Forest	1	3,92%	453,06	0,0025	16,47	21,29
Xgboost Arvore	1	5,19%	878,41	0,005	22,25	29,64
Xgboost Linear	1	3,42%	480,35	0,0025	14,81	21,92
LSTM	1	4,90%	1.344,21	0,0058	23,79	36,66
ARIMAX	1	7,19%	2.469,28	0,0098	36,04	49,69
Random Forest	2	9,18%	3.734,73	0,0134	51,32	61,11
Xgboost Arvore	2	8,36%	3.272,95	0,0116	46,92	57,21
Xgboost Linear	2	8,79%	3.039,66	0,0112	46,82	55,13
LSTM	2	7,87%	2.336,30	0,0079	43,99	48,34
ARIMAX	2	7,02%	2.087,40	0,0069	38,84	45,69
Random Forest	3	39,04%	201.456,53	0,3165	395,32	448,84
Xgboost Arvore	3	39,85%	207.855,16	0,3322	402,83	455,91
Xgboost Linear	3	7,04%	8.115,89	0,0074	73,1	90,09
LSTM	3	13,20%	16.962,26	0,0212	119,93	130,24
ARIMAX	3	6,65%	5.215,36	0,0057	64,63	72,22
Random Forest	4	14,78%	47.969,76	0,0298	203,08	219,02
Xgboost Arvore	4	5,56%	7.444,78	0,0039	76,1	86,28
Xgboost Linear	4	11,83%	31.864,28	0,0159	155,2	178,51
LSTM	4	10,10%	24.947,48	0,0128	132,14	157,95
ARIMAX	4	3,64%	7.699,98	0,0039	47,8	87,75
Random Forest	5	19,59%	45.408,06	0,0401	188,86	213,09
Xgboost Arvore	5	14,19%	24.388,61	0,022	139,63	156,17
Xgboost Linear	5	7,25%	7.084,04	0,0068	73,3	84,17
LSTM	5	2,80%	1.232,79	0,0012	28,47	35,11
ARIMAX	5	8,20%	10.465,74	0,0092	82,53	102,3

Fonte: Resultados originais da pesquisa

Consolidando todos resultados e calculando a média das métricas, ARIMAX, se mostrou o modelo com a melhor generalização e consistência, tendo um desempenho constante e proporcional aos valores preditos. A rede neural apresentou bons resultados,

porém inconsistente em momentos de grandes quebras estruturais ou em estimar valores que estavam fora do intervalo dos dados de treinamento. O xgboost linear apresentou resultados similares ao LSTM. O ARIMAX alcançou um desempenho muito superior aos modelos baseados em árvore de decisão, e com resultados médios similares aos demais modelos, assim como descrito na tabela 3.

Tabela 3: Media métricas de erro

Modelo	MAPE	MSE	MSLE	MAE	RMSE
Random Forest	17,31%	59.804,43	0,0805	171,01	192,67
Xgboost Arvore	14,63%	48.767,98	0,075	137,55	157,04
Xgboost Linear	7,67%	10.116,85	0,0088	72,65	85,96
ARIMAX	6,54%	5.587,55	0,0071	53,97	71,53
LSTM	7,77%	9.364,61	0,0098	69,67	81,66

Fonte: Resultados originais da pesquisa

Em conclusão, a avaliação dos resultados revela insights valiosos sobre o desempenho dos modelos utilizados para a previsão de séries temporais. A aplicação do método de validação cruzada em cinco recortes distintos permitiu uma análise abrangente da capacidade de generalização de cada modelo em diferentes contextos de mercado. A constatação de que o modelo ARIMAX demonstrou a melhor generalização e consistência ao longo dos diferentes períodos é notável, destacando-se pela sua performance estável e proporcional aos valores previstos.

Considerações Finais

Neste trabalho foi possível analisar a série temporal da cotação mensal da saca de café, descrever o seu comportamento, avaliar fatores que influenciam o preço alvo e propor métodos de natureza supervisionada para estimação de valores futuros da série temporal. Os modelos propostos foram avaliados utilizando o método de validação cruzada e as variáveis exógenas estudas foram o dólar, IPCA 12M e o preço médio do diesel.

Com base nos dados coletados, análises gráficas, cálculo das correlações e teste de Granger efetuado, é possível concluir que as variáveis exógenas coletas possuem forte influência na cotação da saca de café dentro do período analisado. Porém esses três fatores ainda não são suficientes para explicar toda a variação do preço da saca de café.

Após as análises efetuadas, foi possível realizar a implementação de diferentes modelos e algoritmos de aprendizado de máquina, redes neurais e estatística clássica. Dentre

essas implementações foi possível concluir que o método tradicional conhecido como abordagem de Box e Jenkins, o ARIMA, apresentou o melhor resultado demonstrando métricas de erros constantes, mesmo em momentos de grande quebra estrutural e oscilações na cotação da saca de café.

Em pesquisas futuras proponho a investigação de variáveis ditas como relevantes, porém que não estavam disponíveis para estudo dentro do período de realização deste trabalho, informações como: os preços de insumos agrícolas, dados do clima nas áreas produtoras de café (temperatura, volume de chuvas, geada, etc.), taxa Selic, oferta de grãos por parte do mercado cafeicultor e demandas por grãos por parte dos consumidores. Quanto as implementações de modelos ainda existem espaço para aprofundar na otimização dos parâmetros ou na abordagem de outras técnicas.

Agradecimentos

Gostaria de agradecer primeiramente a Deus, a minha esposa Leticia e minhas filhas Sofia e Alice que foram muito parceiras e compreensivas em momentos que precisei estar ausente para confecção deste trabalho e durante o andamento de todo o curso, elas são a minha motivação para seguir em frente. Quero agradecer a minha amiga e colega de trabalho Catia que me indicou livros, bibliográficas e conteúdo acadêmicos sobre séries temporais, aos quais me dediquei a estudar antes do início do período de orientação, por fim, mas não menos importante, ao meu orientador que me guiou nessa trajetória e que sempre se mostrou disponível e solicito.

Referências

Breiman, Leo. 2001, Random Forests. Disponível em <https://link.springer.com/article/10.1023/a:1010933404324>. Acesso em: 08/07/2023

Centro de Estudos Avançados em economia Aplicada [CEPEA]. 2020. Agronegócio e a Inflação. Disponível em: <https://www.cepea.esalq.usp.br/upload/kceditor/files/Cepea_agro_e_inflacao_(2).pdf>. Acesso em 02 de junho de 2023.

Bergmeira, Christoph. Hyndman, Job J. Bonsoo, Koob. 2017. A Note on the Validity of Cross-Validation for Evaluating Autoregressive Time Series Prediction. Acesso em: <https://perma.cc/YS3J-6DMD>. Acesso em 10 setembro 2023

Faceli, Katti. Lorena, Ana C. Gama, João. et al. 2021. Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina. Grupo GEN. Rio de Janeiro, RJ, Brasil. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788521637509/. Acesso em: 24 mar. 2023.

Gèron, Aurélio. 2021. Mãos à Obra: Aprendizado de Máquina com Scikit-Learn, Keras & TensorFlow: Conceitos Ferramentas e Técnicas para construção de sistemas inteligentes. 2ed. Alta Books. Rio de janeiro, RJ, Brasil.

Harrison, Matt. 2020. Machine Learning: Guia de Referência Rápida. 1ed. Novatec Editora Ltda. São Paulo, SP, Brasil.

Haykin, Simon. 2001. Redes neurais princípios e prática. Londrina, Bookman Companhia Editora. Porto Alegre, RS, Brasil. Disponível em: <https://integrada.minhabiblioteca.com.br/#/books/9788577800865/>. Acesso em: 12 mar. 2023.

Morettin, Pedro A. Toloi, Clélia M. C. 2018. Análise de Séries Temporais: Modelos lineares univariados. Volume 1. Editora Blucher. São Paulo, SP, Brasil. Disponível em: <https://integrada.minhabiblioteca.com.br/#/books/9788521213529/>. Acesso em: 12 março de 2023.

Morettin, Pedro A. Toloi, Clélia M. C. 2020. Análise de séries temporais: Modelos multivariados e não lineares. Volume 2. Editora Blucher. São Paulo, SP, Brasil. Disponível em: <https://integrada.minhabiblioteca.com.br/#/books/9786555060065/>. Acesso em: 20 setembro de 2023.

Morettin, Pedro A. 2017. Econometria Financeira: Um Curso em Séries Temporais Financeiras. Editora Blucher. São Paulo, SP, Brasil. Disponível em: <https://integrada.minhabiblioteca.com.br/#/books/9788521211310/>. Acesso em: 22 março de 2023.

Neves, Cesar das. Rossi, José W. 2014. Econometria e Séries Temporais com Aplicações à Dados da Economia Brasileira. Grupo GEN. São Paulo, SP, Brasil. Disponível em:

<https://integrada.minhabiblioteca.com.br/#/books/978-85-216-2685-5/>. Acesso em: 12 março de 2023.

Nielsen, Ailen. 2021. Análise Prática de Séries Temporais: Predição com Estatística e Aprendizado de Máquina Starlin. 1ed. Alta Editora e Consultoria. Rio de Janeiro, RJ, Brasil.

Nogueira, José Guilherme A. NEVES, Marcos F. 2015. Estratégias para a Cafeicultura no Brasil. Editora Atlas. São Paulo, SP, Brasil. Disponível em: <https://integrada.minhabiblioteca.com.br/#/books/9788522497867/>. Acesso em: 19 mar. 2023.

Sesso, Patrícia Pompermayer. Filho, Umberto Antonio Sesso; PEREIRA, Luiz Filipe Protasio. 2021. Dimensionamento do agronegócio do café no Brasil. Disponível em: <https://ainfo.cnptia.embrapa.br/digital/bitstream/item/225200/1/dimensionamento-do-agronegocio-do-cafe-2021.pdf>. Acesso em: 18 maio 2023.