



UFOP

Universidade Federal
de Ouro Preto

**Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Departamento de Computação e Sistemas**

Análise do Impacto do Dólar no Preço do Café Arábica por meio de Inteligência Computacional

Luiz Felipe Pereira Anunciação

TRABALHO DE CONCLUSÃO DE CURSO

ORIENTAÇÃO:

Fernando Bernardes de Oliveira

**Outubro, 2022
João Monlevade—MG**

Luiz Felipe Pereira Anunciação

Análise do Impacto do Dólar no Preço do Café Árábica por meio de Inteligência Computacional

Orientador: Fernando Bernardes de Oliveira

Monografia apresentada ao curso de Sistemas de Informação do Instituto de Ciências Exatas e Aplicadas, da Universidade Federal de Ouro Preto, como requisito parcial para aprovação na Disciplina “Trabalho de Conclusão de Curso II”.

Universidade Federal de Ouro Preto

João Monlevade

Outubro de 2022



FOLHA DE APROVAÇÃO

Luiz Felipe Pereira Anunciação

Análise do impacto do Dólar no preço do Café Arábica por meio de Inteligência Computacional

Monografia apresentada ao Curso de Engenharia de Computação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Engenharia de Computação

Aprovada em 04 de novembro de 2022

Membros da banca

Prof. Dr. Fernando Bernardes de Oliveira - Orientador (Universidade Federal de Ouro Preto)
Profª. Drª. Gilda Aparecida de Assis - (Universidade Federal de Ouro Preto)
Profª. Drª. Helen de Cássia Sousa da Costa Lima - (Universidade Federal de Ouro Preto)

Fernando Bernardes de Oliveira, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 13/12/2022



Documento assinado eletronicamente por **Fernando Bernardes de Oliveira, PROFESSOR DE MAGISTERIO SUPERIOR**, em 13/12/2022, às 16:02, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0443048** e o código CRC **A439D729**.

Este trabalho é dedicado a minha família, pois é graças a seu esforço que hoje posso concluir meu curso

Agradecimentos

Agradeço primeiramente a Deus, pela vida e por me permitir chegar até aqui. Meu pai, Sidney Alves, pelo apoio incondicional e exemplo, minha mãe, Maria Aparecida, pela paciência, amor e orações, e minha irmã, Gabriela, por todo o carinho. Ao meu orientador, Fernando Bernardes, pela cordialidade e gentileza. Eu não seria capaz de realizar nada sem o apoio de cada um deles.

“Science is more than a body of knowledge; it is a way of thinking.”

— Carl Sagan (1934 – 1996),
in: The Demon-Haunted World: Science as a Candle in the Dark.

Resumo

A comercialização do café é de extrema importância para a economia do Brasil, desde sua colonização, pois representa grande parte das exportações do país. O fato da negociação da *commodity* ser realizada em dólar, torna importante conhecer o impacto da moeda americana nos preços dos contratos futuros café arábica, acordos de compra e venda do café. Este trabalho utiliza séries temporais e algoritmos de aprendizagem de máquina, para verificar a influência do dólar no preço do café, com base no ganho de performance dos modelos construídos utilizando a moeda americana como uma *feature*. Os resultados encontrados através dos modelos LSTM, XGboost e Prophet sugerem uma grande correlação entre o valor do dólar e o preço dos contratos de café.

Palavras-chaves : café arábica, aprendizado de máquina, séries temporais.

Abstract

The commercialization of coffee is extremely important for the economy of Brazil, since its colonization, as it represents a large part of the country's exports. The fact that the commodity is traded in dollars makes it important to know the impact of the US currency on the prices of Arabica coffee futures contracts, coffee purchase and sale agreements. This work uses time series and machine learning algorithms to verify the influence of the dollar on the price of coffee, based on the performance gain of the models built using the American currency as a feature. The results found through the LSTM, XGboost, and Prophet models suggest a strong correlation between the value of the dollar and the price of coffee contracts.

Key-words: Arabica coffee, machine learning, time series.

Lista de ilustrações

Figura 1 – Série Temporal	24
Figura 2 – Decomposição: Tendência	25
Figura 3 – Decomposição: Sazonalidade	25
Figura 4 – Decomposição: Resíduo	25
Figura 5 – Boxplot	26
Figura 6 – Preço do Café em Dólar	26
Figura 7 – Dólar Cotado em Real	27
Figura 8 – Comparação das Séries 2015 - 2018	27
Figura 9 – Gráfico de Dispersão 2015 - 2018	28
Figura 10 – Mapa de Calor 2015 - 2018	29
Figura 11 – Dados de Treino e Teste	30
Figura 12 – LSTM - Modelo 1	31
Figura 13 – LSTM - Modelo 2	32
Figura 14 – XGBoost - Modelo 1	32
Figura 15 – XGBoost - Modelo 2	33
Figura 16 – Prophet - Modelo 1	34
Figura 17 – Prophet - Modelo 2	34

Lista de abreviaturas e siglas

AI *Artificial Intelligence*

AR Auto-regressão

ARIMA *Autoregressive Integrated and Moving Average*

BM&F Bovespa Bolsa de Valores do Brasil

EDM *Educational Data Mining*

ES *Exponential Smoothing*

ELM *Extreme Learning Machines*

Feedforward Neural Network feedforward

Gradient Boosting Aumento de Gradiente

ICEA Instituto de Ciências Exatas e Aplicadas

ICEA Instituto de Ciências Exatas e Aplicadas

LSTM *Long short-term memory*

MAE Erro Médio Absoluto

MAPE Erro Percentual Absoluto Médio

MSE Erro Quadrático Médio

MLP *Multilayer Perceptron*

RNN *Recurrent Neural Networks*

XGBoost *Extreme Gradient Boosting*

Sumário

1	INTRODUÇÃO	14
1.1	Justificativa	14
1.2	Objetivos	15
1.3	Metodologia	15
1.4	Organização do trabalho	16
2	REVISÃO BIBLIOGRÁFICA	17
2.1	Séries Temporais	18
2.2	Machine Learning	19
2.3	Redes Neurais	19
2.3.0.1	LSTM	20
2.4	XGBoost	20
2.5	Prophet	21
2.6	Avaliação	22
2.7	Considerações finais	22
3	DESENVOLVIMENTO	23
3.1	Coleta de Dados	23
3.2	Limpeza e Organização dos Dados	23
3.3	Análise Exploratória	24
3.3.1	Café Arábica	24
3.3.1.1	Cotação do Dólar Americano em Real	27
3.3.2	Relação entre o Preço do Café Arábica e a Cotação Dólar	27
3.4	Considerações finais	29
4	RESULTADOS	30
4.1	Modelagem	30
4.2	LSTM	31
4.2.1	LSTM - Modelo 1	31
4.2.2	LSTM - Modelo 2	31
4.3	XGBoost	32
4.3.1	XGBoost - Modelo 1	32
4.3.2	XGBoost - Modelo 2	33
4.4	Prophet	33
4.4.1	Prophet - Modelo 1	33
4.4.2	Prophet - Modelo 2	34

4.5	Considerações Finais	35
5	CONCLUSÃO	36
5.1	Trabalhos Futuros	36
	REFERÊNCIAS	37

1 Introdução

O plantio e comercialização do café possui papel fundamental no desenvolvimento socioeconômico do mundo e principalmente no Brasil. Os primeiros grãos chegaram em solo brasileiro por volta do início do século XVIII, pelo estado do Pará. No entanto, em um curto espaço de tempo, se espalhou por todo o território nacional. Desde então, o café aparece em todo o contexto econômico do país, tendo sido por muitos anos sua principal fonte de dinheiro. (PINTO, 2022)

O Brasil é o maior produtor de café arábica (*Coffea arabica*) do mundo. Atualmente, o café ainda é de extrema importância para a economia brasileira, pois contribui com mais de 2% do valor das exportações totais, e representa mais de um terço da produção mundial. Desta forma, entender o comportamento do preço de tal produto se torna muito interessante para pessoas interessadas em negociar a *commodity*. (CAMPOS, 2005)

A negociação do café é feita sob a forma de contratos no mercado futuro na Bolsa de Valores do Brasil (BM&F Bovespa). O mercado futuro se comporta como um contrato de compra e venda a um valor e um período de validade. Desta maneira, quando o prazo de vencimento é atingido, as perdas ou os ganhos do investidor são calculados, de acordo com a cotação atual. A cotação do Café Arábica Futuro, espécie de café de maior qualidade e interesse econômico, é auferida em dólares americanos (US\$) por saca de 60 quilogramas. Cada contrato de Café Arábica Futuro representa a negociação de 100 sacas de 60 Kg. Logo, as oscilações da moeda americana estão diretamente ligadas ao comportamento do mercado de café. (B3, 2022)

Tendo em vista a importância do café para a economia brasileira, estudar e entender o comportamento de seu mercado é um grande passo para o desenvolvimento socioeconômico do Brasil. O fato da negociação do café estar atrelada à moeda americana, torna grande a necessidade de estudar e conhecer a magnitude do impacto do dólar sobre a *commodity* brasileira. Portanto, este trabalho se propõe a estudar e verificar os impactos do dólar no preço do café arábica por meio de séries temporais e modelos de aprendizagem de máquina.

1.1 Justificativa

Apesar de os dados sobre a cotação do café arábica e do dólar americano estarem disponíveis na internet, o processo de compreender seu relacionamento não é algo simples, pois depende de outros fatores, como as oscilações do mercado internacional e nacional. Além disso, uma das características do mercado de café é a alta volatilidade, o que

tonar sua análise uma tarefa complexa. Com a cotação diária dos contratos de café, é possível estudar suas variações durante uma janela de tempo, de maneira a identificar suas características por meio da utilização de séries temporais. Ademais, com o auxílio de técnicas de aprendizagem de máquina, é possível quantificar seu relacionamento com o dólar americano, pois um modelo construído para prever o preço do café tende a ter um ganho de performance quando uma variável correlata é inserida.

1.2 Objetivos

Este projeto pretende estudar o impacto do dólar americano no preço do café arábica brasileiro utilizando séries temporais e técnicas de aprendizagem de máquina. De modo geral, este trabalho possui os seguintes objetivos específicos:

- Reunir dados históricos sobre o valor do café arábica e do dólar.
- Analisar os dados reunidos.
- Definir e implementar algoritmos de aprendizagem de máquina para prever o valor do café arábica em períodos futuros.
- Analisar os resultados.

1.3 Metodologia

A metodologia seguida no desenvolvimento do projeto consiste em realizar uma revisão bibliográfica, com o intuito de reunir mais conhecimento acerca dos temas, assim como, desenvolver e experimentar modelos. As atividades a serem realizadas podem ser descritas como:

- Revisão da literatura: Estudar artigos relacionados à previsão de ações utilizando aprendizagem de máquina e o funcionamento do mercado de café arábica.
- Coleção de dados: Reunir e estudar dados históricos contendo o valor dos contratos de café e do dólar.
- Definição das técnicas de aprendizagem de máquina: definir os algoritmos para a predição do valor do contrato de café.
- Implementação: implementar as técnicas de aprendizagem de máquina definidas.
- Análise e discussão: avaliar os resultados obtidos.

1.4 Organização do trabalho

O restante deste trabalho é organizado como se segue. O Capítulo 2 apresenta alguns trabalhos correlatos, incluindo aqueles que utilizaram os mesmos algoritmos de aprendizagem. Além disso, descreve todas as técnicas utilizadas no desenvolvimento do trabalho, como os algoritmos e técnicas de avaliação utilizados. No Capítulo 3, estão descritos os dados utilizados, o pré-processamento e análises. No Capítulo 4, estão os resultados das previsões geradas pelos modelos implementados. Já o Capítulo 5 é composto pelas considerações referentes aos resultados obtidos.

2 Revisão bibliográfica

Entender e prever o comportamento do mercado de futuros é de interesse de muitas pessoas ao redor do mundo, devido ao alto volume de dinheiro que ele movimenta. Desta forma, muitas pesquisas procuram formas de prever de forma precisa seu comportamento futuro. [Fofanah \(2021\)](#) em sua pesquisa, conseguiu alcançar precisão de mais de 90% na maioria dos algoritmos de previsão, utilizando *Long short-term memory (LSTM)*, XGB e LR. Seu foco está no desenvolvimento de um aplicativo que ajude o mercado de café da Etiópia, espinha dorsal da economia do país.

Os preços das commodities agrícolas de grãos são sabidamente voláteis devido a diversos fatores que influenciam esses preços. Além disso, diferentes combinações desses fatores, como demanda, oferta e indicadores macroeconômicos, são responsáveis pela volatilidade dos preços em diferentes momentos. Big Data apresenta oportunidades para coletar e integrar conjuntos de dados de várias fontes com o objetivo de descobrir padrões úteis e extrair insights que podem ser usados para obter vantagem competitiva ou melhorar a tomada de decisões. As redes neurais apresentam oportunidades de pesquisa para treinar algoritmos para modelar padrões lineares e não lineares que possam existir em conjuntos de dados, com o objetivo de fazer previsões. Em seu trabalho, [Ayankoya, Calitz e Greyling \(2016\)](#) propõem uma abordagem de Big Data e Redes Neurais para prever preços de *commodities* de grãos na África do Sul.

[Deina et al. \(2022\)](#) apresenta uma metodologia para estimar preços de café com base no uso de *Extreme Learning Machines (ELM)*. O processo é iniciado identificando a presença de componentes não estacionários, como sazonalidade e tendência. Esses componentes são retirados se forem encontrados. Em seguida, as defasagens temporais são selecionadas com base na resposta da função de autocorrelação parcial. Como preditores, foram utilizados os seguintes modelos: *Exponential Smoothing (ES)*, Auto-regressão (AR), *Autoregressive Integrated and Moving Average (ARIMA)*, *Multilayer Perceptron (MLP)* e ELM.

No artigo de [Dai e Zhang \(2013\)](#), eles utilizaram um conjunto de dados com 3 milhões de observações. O conjunto de dados continha informações diárias sobre ações no período de Janeiro/2008 a Agosto/2013. Muitos algoritmos foram utilizados, como Regressão Logística, e SVM. O algoritmo SVM alcançou a maior precisão, 79,3%.

Em [Selvin et al. \(2017\)](#), o método proposto é uma abordagem independente de modelo. Neste trabalho, os dados não são ajustados a um modelo específico, mas é feita a identificação dinâmica latente existente nos dados usando arquiteturas de *deep learning*. Neste trabalho, são utilizados três diferentes arquiteturas de *deep learning* para a previsão

de preços de ações de empresas listadas no NSE. Uma abordagem de janela deslizante é aplicada para prever valores futuros em uma base de curto prazo. O desempenho dos modelos foi quantificado usando o erro percentual.

Em sua pesquisa, [Castro, Teixeira e Lima \(2005\)](#) analisa o impacto da desvalorização da moeda Brasileira, ocorrida em 1999, no valor de insumos e na composição dos custos na cultura do café no Brasil. Ele constata, por meio de seus resultados, que o café foi beneficiado pela desvalorização cambial, embora tenha ocorrido um aumento nos preços dos insumos.

Em seu trabalho, [Gonçalves \(2022\)](#) apresenta uma metodologia capaz de identificar o perfil dos alunos do Instituto de Ciências Exatas e Aplicadas (ICEA) em relação ao comportamento curricular com foco no sucesso acadêmico. Através da utilização de técnicas de *Artificial Intelligence* (AI) e *Educational Data Mining* (EDM)) foi possível identificar precocemente os discentes com maior probabilidade de evadir do Instituto. O problema foi abordado com o desenvolvimento de Modelos de Aprendizagem de Máquina baseados na *Random Forest*.

2.1 Séries Temporais

Uma série temporal é definida como um conjunto de observações quantitativas organizadas em ordem cronológica. Geralmente é considerado que o tempo é uma variável discreta. As séries temporais sempre foram utilizadas no campo da econometria ([KIRCHGÄSSNER; WOLTERS; HASSLER, 2012](#)). Já no início, [Tinbergen \(1939\)](#) construiu o primeiro modelo econométrico para os Estados Unidos e assim iniciou o programa de pesquisa científica de econometria empírica. Naquela época, porém, dificilmente se levava em conta que observações ordenadas cronologicamente poderiam depender umas das outras. A suposição predominante foi que, de acordo com o modelo de regressão linear clássico, os resíduos das equações estimadas são estocasticamente independentes entre si.

Isso não mudou até [Box e Jenkins \(1973\)](#) publicou um livro sobre análise de séries temporais que recebeu considerável atenção. Em primeiro lugar, introduziram modelos univariados para séries temporais que simplesmente faziam uso sistemático das informações incluídas nos valores observados das séries temporais. Isso ofereceu uma maneira fácil de prever o desenvolvimento futuro dessa variável. Tornou-se ainda mais popular quando ([GRANGER; NEWBOLD, 1975](#)) mostraram que previsões simples que consideravam apenas informações fornecidas por uma única série temporal muitas vezes superavam as previsões baseadas em grandes modelos econométricos que às vezes consistiam em muitas centenas de equações.

2.2 Machine Learning

O aprendizado de máquina aborda a questão de como construir computadores que melhoram automaticamente através da experiência. É um dos campos técnicos de crescimento mais rápido da atualidade, encontrando-se na interseção da ciência da computação e estatística, e no centro da inteligência artificial e à ciência de dados (MITCHELL et al., 1990). O progresso recente no aprendizado de máquina foi impulsionado tanto pelo desenvolvimento de novos algoritmos de aprendizagem quanto pela explosão na disponibilidade de conhecimento gratuito.

2.3 Redes Neurais

Redes Neurais Artificiais podem ser definidas como técnicas computacionais capazes de apresentar um modelo matemático inspirado na estrutura do neurônio de organismos capazes de pensar e adquirirem conhecimento com base na experiência. Uma grande rede neural pode ser composta por uma enorme quantidade de unidades de processamento (ICMC, 2003).

Um histórico resumido sobre Redes Neurais Artificiais deve começar por três das mais importantes publicações iniciais, desenvolvidas por: McCulloch e Pitts (1943), Hebb (1949), e Rosenblatt (1958). Estas publicações introduziram o primeiro modelo de redes neurais simulando “máquinas”, o modelo básico de rede de auto-organização, e o modelo Perceptron de aprendizado supervisionado, respectivamente.

Uma rede neural artificial é composta por várias unidades de processamento, cujo funcionamento é bastante simples. Essas unidades, geralmente são conectadas por canais de comunicação que estão associados a determinado peso. As unidades fazem operações apenas sobre seus dados locais, que são entradas recebidas pelas suas conexões. O comportamento inteligente de uma Rede Neural Artificial vem das interações entre as unidades de processamento da rede. (ICMC, 2003)

A operação de uma unidade de processamento, proposta por McCulloch e Pitts (1943), pode ser resumida da seguinte maneira:

- Sinais são apresentados à entrada
- Cada sinal é multiplicado por um número, ou peso, que indica a sua influência na saída da unidade
- É feita a soma ponderada dos sinais que produz um nível de atividade
- Se este nível de atividade exceder um certo limite *threshold* a unidade produz uma determinada resposta de saída.

As redes neurais mais comuns são do tipo *feedforward* ([Feedforward Neural Network](#)), nessas redes a informação percorre sempre a mesma direção, até a última camada. Existem diversas maneiras de se conectar os neurônios e modelar uma rede, dentre elas na *Recurrent Neural Networks* ([RNN](#)), as informações podem percorrer diferentes direções, desta maneira, um neurônio além de alimentar o próximo, consegue se realimentar produzindo uma categoria de memória a curto prazo. Em decorrência destas características, as RNNs podem criar modelos mais complexos e extensíveis a um grupo maior de problema ([CAVAIONI, 2014](#)).

2.3.0.1 LSTM

As *LSTM* foram apresentadas por [Hochreiter e Schmidhuber \(1997\)](#) e fazem uso do mesmo conceito das *RNN* com um aperfeiçoamento relacionado a memória. Especificamente no que se refere a dependência entre elementos distantes na rede, as RNNs tradicionais não são eficientes em capturar essa conexão mais afastada. Projetadas para evitar o problema da dependência de longo prazo, as LSTMs possuem um módulo de repetição especial. Nas RNNs padrões o módulo de repetição possui uma única camada de rede neural, já nas LSTMs o mesmo possui 4 camadas de redes neurais, as quais interagem de uma maneira particular ([GREFF et al., 2016](#)). O principal ponto do funcionamento da *LSTM* é o estado da célula, que atravessa toda a estrutura da rede, podendo sofrer alterações gerenciadas pelos chamados portões. Estes portões são compostos de uma camada de rede neural com função de ativação sigmoide e uma operação de multiplicação, o que controlará a quantidade que passará de cada componente, gerando assim, a capacidade de remover ou adicionar informações ao estado da célula *LSTM* ([CHEN, 2017](#)).

2.4 XGBoost

O Aumento de Gradiente ([Gradient Boosting](#)), é um método que passa por ciclos para adicionar modelos de forma iterativa em um conjunto. Ele começa inicializando o conjunto com um único modelo, cujas previsões podem ser bastante ingênuas. Mesmo que suas previsões sejam extremamente imprecisas, adições subsequentes ao conjunto podem solucionar esses erros. ([FOFANA, 2021](#)). Abaixo, uma descrição do funcionamento do ciclo.

- Primeiro, é utilizado o conjunto atual para gerar previsões para cada observação no conjunto de dados. Para fazer uma previsão, são adicionadas as previsões de todos os modelos no conjunto.
- Essas previsões são usadas para calcular uma função de perda (como erro quadrático médio, por exemplo).

- Em seguida, é utilizada a função de perda para ajustar um novo modelo que será adicionado ao conjunto. Os parâmetros do modelo são ajustados para que a adição desse novo modelo ao conjunto reduza a perda. (O gradiente em aumento de gradiente refere-se ao fato de que é usado gradiente descendente na função de perda para determinar os parâmetros neste novo modelo)
- Por fim, é adicionado o novo modelo ao conjunto e então o ciclo se repete.

O *Extreme Gradient Boosting* ([XGBoost](#)) é uma implementação escalável e altamente precisa de aumento de gradiente que ultrapassa os limites do poder de computação para algoritmos de árvore de decisão, sendo construído em grande parte para melhorar o desempenho e a velocidade computacional dos modelos. Com o [XGBoost](#), as árvores são construídas em paralelo, em vez de sequencialmente.

As árvores de decisão criam um modelo que prevê o resultado avaliando uma árvore de perguntas de se-então verdadeiro-falso e estimando o número mínimo de perguntas necessárias para avaliar a probabilidade de tomar uma decisão correta. As árvores de decisão podem ser usadas para classificação para prever uma categoria ou regressão para prever um valor numérico contínuo.

2.5 Prophet

O Facebook Prophet ([TAYLOR; LETHAM, 2018](#)) é um algoritmo de código aberto para gerar modelos de séries temporais que usa ideias já consolidadas. É particularmente bom na modelagem de séries temporais que têm várias sazonalidades e não enfrenta algumas das desvantagens de outros algoritmos já conhecidos. Em seu núcleo está a soma de três funções de tempo mais um termo de erro:

$$y(t) = g(t) + s(t) + h(t) + e_t \quad (2.1)$$

- crescimento: A função de crescimento modela a tendência geral dos dados.
- sazonalidade: A função de sazonalidade é simplesmente uma série de Fourier em função do tempo.
- feriados: A função de feriado permite que o Facebook Prophet ajuste a previsão quando um feriado ou evento importante pode alterar a previsão.

O Prophet permite que o analista forneça uma lista personalizada de eventos passados e futuros. Uma janela em torno desses dias é considerada separadamente e parâmetros adicionais são ajustados para modelar o efeito de feriados e eventos.

2.6 Avaliação

Após obter os resultados de uma previsão, o modelo é avaliado utilizando métricas. Os valores obtidos são comparados com os valores verdadeiro, esse processo é realizado para quantificar a eficácia dos modelos. Foram utilizadas três das principais métricas de avaliação utilizadas neste tipo de estudo.

1. Erro Médio Absoluto ([MAE](#))
2. Erro Quadrático Médio ([MSE](#))
3. Erro Percentual Absoluto Médio ([MAPE](#))

O erro médio absoluto, é calculado a partir da média dos erros absolutos, é utilizado o módulo de cada erro para evitar a subestimação, pois é menos afetado por pontos muito próximos.

$$MAE = \frac{1}{n} \sum_{t=1}^n |e_t| \quad (2.2)$$

O [MSE](#), é muito utilizado para verificar a acurácia de modelos pois dá um maior peso aos maiores erros, pois cada erro é elevado ao quadrado individualmente, e então a média desses erros é calculada.

$$MSE = \frac{1}{n} \sum_{t=1}^n e_t^2 \quad (2.3)$$

O [MAPE](#) é uma forma mais direta de exibir a qualidade, pois o erro é medido como uma porcentagem, o que facilita o entendimento.

$$MAPE = \frac{100\%}{n} \sum_{t=1}^n \left| \frac{e_t}{y_t} \right| \quad (2.4)$$

2.7 Considerações finais

Este capítulo apresentou uma descrição a cerca das técnicas utilizadas no decorrer do projeto e as fontes consultadas como referência. O conhecimento reunido nesta seção é utilizado no decorrer do projeto.

No próximo capítulo será descrita a etapa de desenvolvimento, composta pelas análises dos conjuntos de dados estudados. Além disso, descreve o ambiente e bibliotecas utilizadas na construção do código.

3 Desenvolvimento

O trabalho se desenvolve a partir da coleta e investigação dos dados utilizados, que servirão de base para o desenvolvimento dos modelos.

Todos o desenvolvimento prático deste projeto foi realizado utilizando a linguagem de programação *Python*, versão 3.7.10, através da ferramenta gratuita *Google Colaboratory*. Foram utilizadas as seguintes bibliotecas: *matplotlib* e *seaborn* para a criação dos gráficos, *statsmodels* para realização de transformações, *scikit-learn* para avaliação dos modelos, e por fim *XGBoost*, *tensorflow* e *Prophet* para a criação dos modelos.

3.1 Coleta de Dados

Para esta pesquisa, foram utilizados os seguintes conjuntos de dados: cotação do café arábica em dólar americano e cotação do dólar americano em real. Os dados contendo informações sobre os contratos futuros de café arábica foram retirados do site *Investing*. As informações têm como base a [BM&F Bovespa](#), responsável pela negociação de contratos futuros no Brasil. Os dados são correspondentes ao período de Janeiro/2002 a Março/2022. As informações sobre a cotação do dólar americano em real foram retirados do mesmo site e compreendem o mesmo período citado acima.

3.2 Limpeza e Organização dos Dados

Após rápida avaliação do conjunto de dados, foram realizadas operações para a adequação dos dados ao formato de série temporal. A série temporal criada é composta apenas pela data e pelo valor de interesse, que para este trabalho está contido na coluna Preço. Desta forma, apenas as colunas contendo o valor da *commoditie* e suas respectivas datas de observação foram preservadas. A imagem abaixo 1, demonstra a estrutura resultante. O mesmo processo foi replicado para as cotações das moedas.

Figura 1 – Série Temporal

	Preço
Data	
2022-03-18	264.15
2022-03-17	260.20
2022-03-16	261.90
2022-03-15	253.55
2022-03-14	263.25
...	...
2002-01-08	52.90
2002-01-07	52.90
2002-01-04	53.10
2002-01-03	52.50
2002-01-02	52.70

Fonte: Elaborado pelo Autor

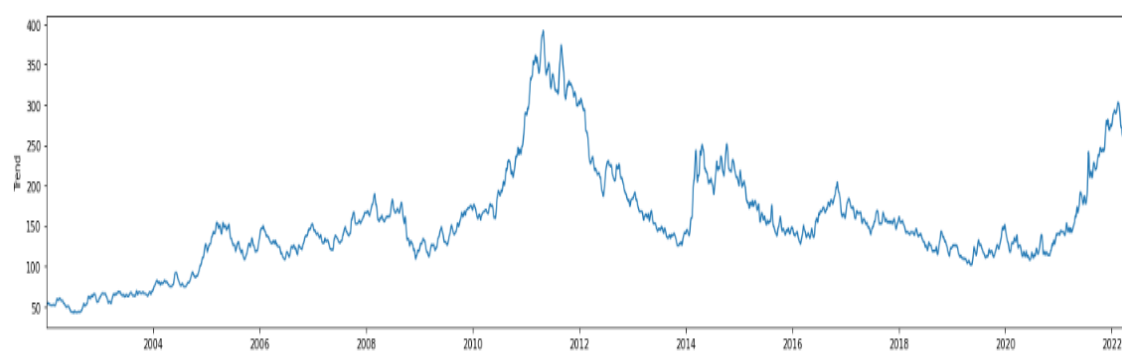
3.3 Análise Exploratória

Esta etapa é de grande importância para o desenvolvimento do projeto. A análise permite identificar as características dos conjuntos de dados, o que facilita a criação dos modelos. O estudo se concentra principalmente na cotação do café arábica, pois é o objeto de estudo principal.

3.3.1 Café Arábica

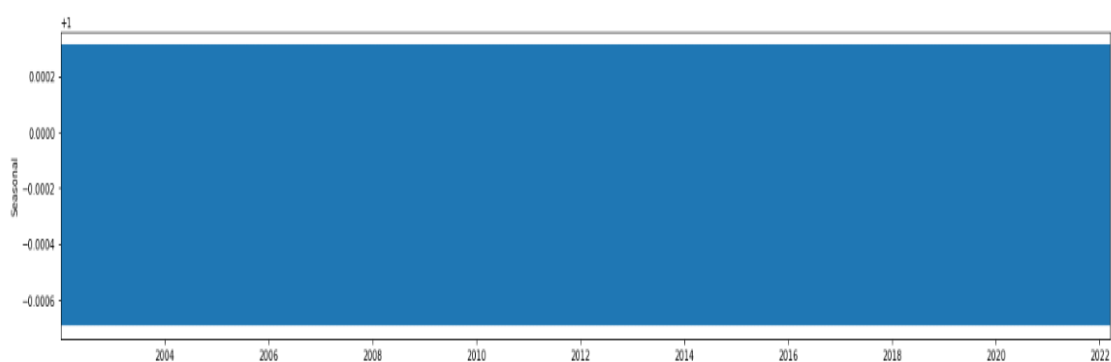
Inicialmente, foi verificada a existência de valores nulos, e os mesmos foram tratados durante processo de reamostragem, utilizado para gerar amostras para os dias que não existiam no conjunto. Em seguida, a série foi decomposta nos seguintes componentes utilizando o método multiplicativo: sazonal, o qual aponta eventos que acontecem de forma repetida durante todo o tempo de observação, tendência, que indica o comportamento da série, e por último o resíduo, que representa o que resta da série depois de ser retirar a tendência e a sazonalidade. Abaixo, as figuras 2, 3 e 4 apresentam os resultados da decomposição.

Figura 2 – Decomposição: Tendência



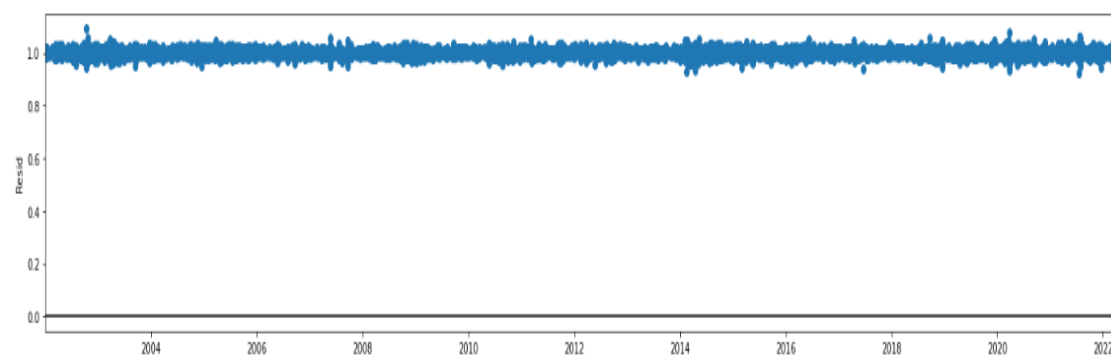
Fonte: Elaborado pelo Autor

Figura 3 – Decomposição: Sazonalidade



Fonte: Elaborado pelo Autor

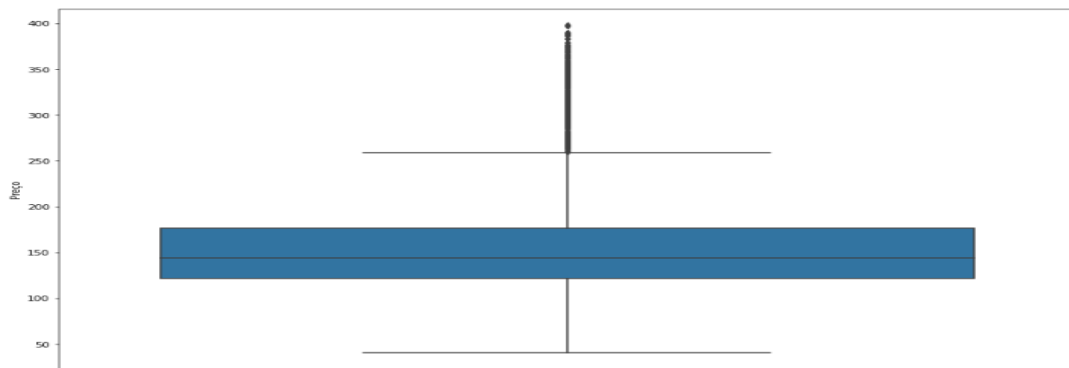
Figura 4 – Decomposição: Resíduo



Fonte: Elaborado pelo Autor

O próximo passo, foi utilizar o gráfico *boxplot* para investigar a presença de *outliers*, valores que se afastam muito das demais observações da série e podem implicar inconsistência. Como pode ser observado na figura número 5, é notável a quantidade de *outliers*. Tal fato aponta grande volatilidade da série. Mesmo após sua identificação, os pontos destoantes não foram removidos, pois fazem parte da característica do conjunto de dados, o que os torna importantes para a modelagem da série.

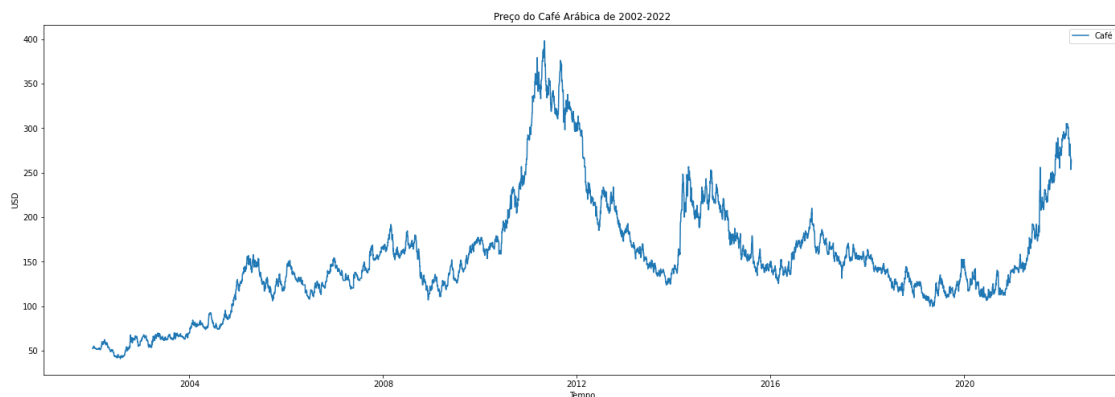
Figura 5 – Boxplot



Fonte: Elaborado pelo Autor

Ao analisar a figura 6, foi possível constatar a volatilidade do preço da *commodity*. Durante o período analisado, o café arábica chegou a alcançar um valor máximo de 398 dólares e um valor mínimo de 41 dólares. A média encontrada foi de 155,21 dólares com um desvio padrão de 64,7. O ponto mais significativo da série é com certeza o pico alcançado em 2011, pois foge muito do preço observado ao longo do tempo, assim como o período pós 2020, que pode ser explicado pela pandemia de Coronavírus.

Figura 6 – Preço do Café em Dólar



Fonte: Elaborado pelo Autor

3.3.1.1 Cotação do Dólar Americano em Real

Durante o período analisado, figura 7, o dólar apresentou um comportamento decrescente de 2001 a 2008, após este período, apresentou uma tendência crescente de forma geral até o ano de 2020, após este período o valor se tornou muito volátil. A moeda americana chegou a alcançar um valor máximo de 5,88 e um valor mínimo de 1,53 reais.

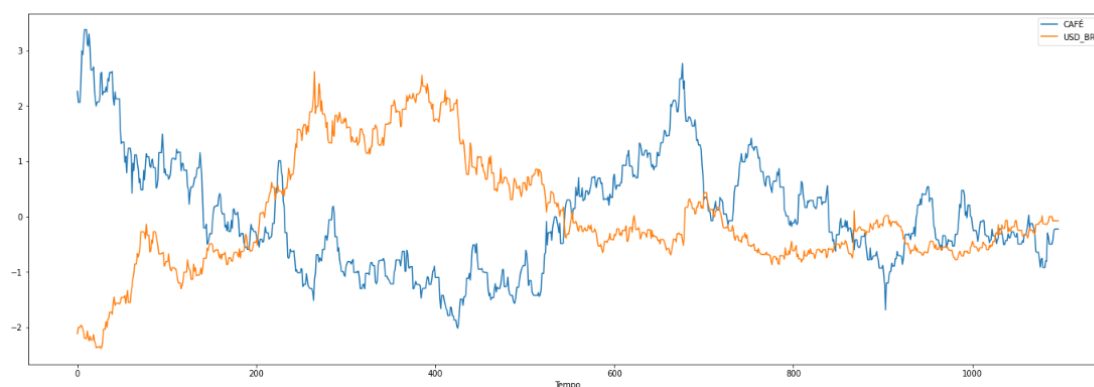
Figura 7 – Dólar Cotado em Real



Fonte: Elaborado pelo Autor

3.3.2 Relação entre o Preço do Café Arábica e a Cotação Dólar

Figura 8 – Comparação das Séries 2015 - 2018



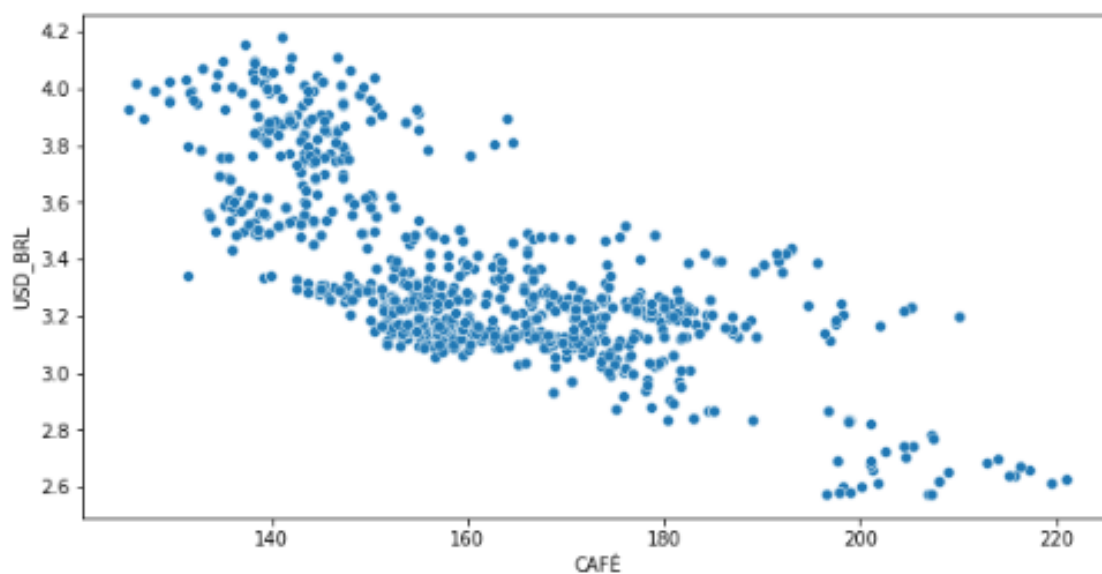
Fonte: Elaborado pelo Autor

Depois de considerações, foi decidido utilizar apenas o período de 2015 a 2018 para os demais estudos. Como é possível observar nas análises apresentadas, a cotação do café arábica está sujeita as oscilações do mercado internacional. Desta forma, foi escolhido um período sem nenhum grande acontecimento que pudesse influenciar a economia global, como a pandemia de coronavírus.

Para verificar a existência de alguma relação entre o valor do café e a moeda estudada, as colunas contendo os dados foram transformadas e dispostas na mesma escala,

por meio do método *StandardScaler*, tal transformação permite uma melhor visualização dos dados em um único visual. Figura 8. Foram criados dois gráficos de dispersão, visuais que utilizam coordenadas cartesianas para exibir o quanto uma variável é afetada por outra, os resultados podem ser vistos na Figura 9. Por fim, foram calculados os coeficientes de correlação das colunas, utilizando o método *corr*.

Figura 9 – Gráfico de Dispersão 2015 - 2018

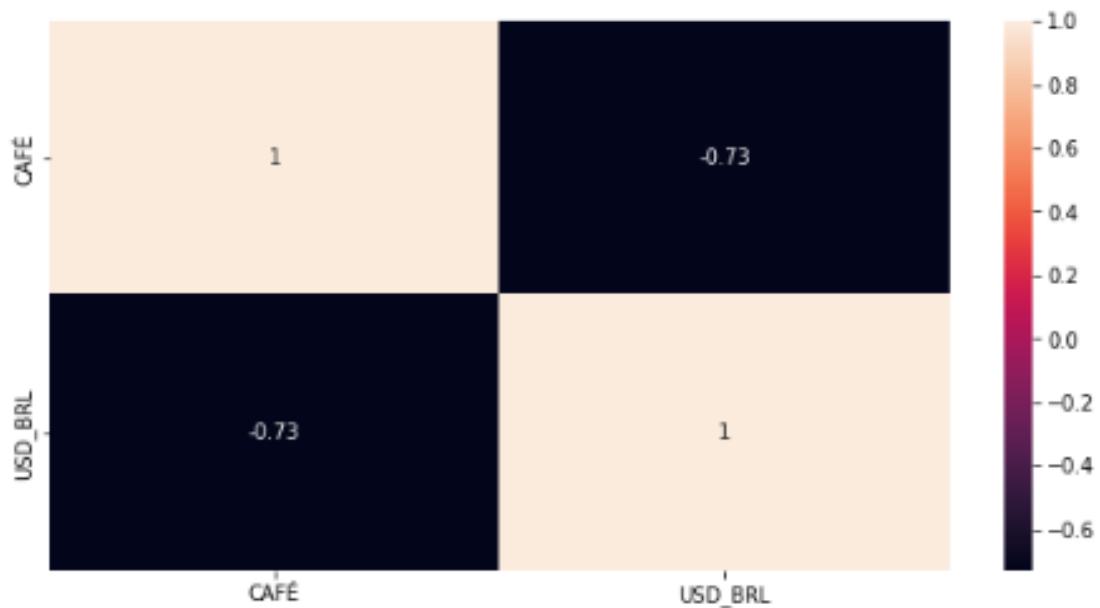


Fonte: Elaborado pelo Autor

Os coeficientes de correlação são métodos estatísticos para se medir as relações entre variáveis e o que elas representam. A correlação busca entender como uma variável se comporta em um ambiente onde outra está variando, visando identificar se existe alguma relação entre a variabilidade de ambas.

O valor do coeficiente de correlação pode alternar entre -1 e 1 e o resultado define se a correlação é negativa ou positiva. 1 significa que a correlação é perfeita positiva e -1 significa que é perfeita negativa. Se o coeficiente for 0, significa que as variáveis não possuem correlação. O visual de mapa de calor, Figura 10, foi utilizado para apresentar os coeficientes de correlação de forma gráfica.

Figura 10 – Mapa de Calor 2015 - 2018



Fonte: Elaborado pelo Autor

3.4 Considerações finais

Após a análise de todos os visuais gerados, ficou evidente uma expressiva correlação entre as variáveis. É possível observar uma correlação negativa de -0,73 entre a cotação dólar em real e o preço do café arábica, os resultados permitem supor que as séries analisadas são influenciadas por algum mecanismo comum e tendem a várias juntas, o que torna a moeda americana uma excelente candidata a *feature* para os modelos de aprendizagem de máquina.

A desvalorização do preço do café frente a alta do dólar foi um resultado inesperado. O fato da *commodity* estar atrelada ao dólar, cria a concepção de que uma queda na moeda americana seria refletida no preço dos contratos de café e provocaria o mesmo fenômeno, o que se mostrou não ser a realidade.

Este capítulo apresentou uma descrição detalhada dos conjuntos de dados utilizados no projeto. Os resultados desta análise permitiram entender o comportamento do preço do café e assim desenvolver os modelos de aprendizagem de máquina. No próximo capítulo são apresentados os modelos gerados e seus respectivos resultados.

4 Resultados

4.1 Modelagem

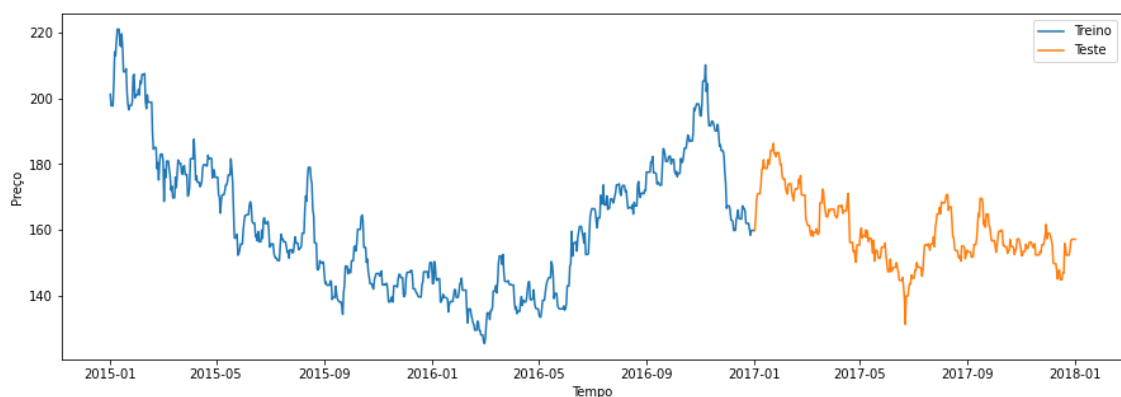
A série temporal construída com os dados do café arábica foi dividida em duas partes, um conjunto para o treino dos modelos e outro para a validação dos resultados. O conjunto de treino compreende o período de Janeiro/2015 a Janeiro/2017, e o de validação, o período de Fevereiro/2017 a Janeiro/2018.

Para todos os algoritmos de aprendizagem de máquina, foram construídos dois modelos. O primeiro, utiliza apenas *features* simples, derivadas exclusivamente da data da observação. São elas:

- Dia da semana
- Trimestre
- Mês
- Ano
- Dia do Ano
- Dia do Mês
- Dia da Semana

Já no segundo modelo, também foi adicionada como *feature* a cotação do dólar em real. O objetivo deste processo, é avaliar se estes dados podem ser úteis na previsão do preço do café, devido à sua expressiva correlação.

Figura 11 – Dados de Treino e Teste



Fonte: Elaborado pelo Autor

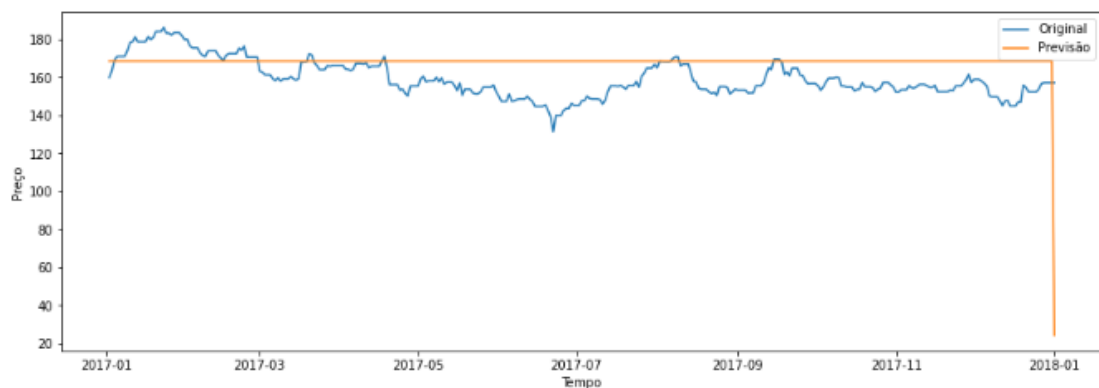
4.2 LSTM

4.2.1 LSTM - Modelo 1

O primeiro modelo [LSTM](#), representando na figura [12](#), que utiliza apenas *features* baseadas na data, demonstrou não conseguir capturar as características do conjunto de dados, pois só apresentou uma linha que pode ser interpretada como a tendência geral da série. O [MSE](#) de 225,96 mostra que os valores obtidos se distanciam bastante dos reais.

- MSE: 225,96
- MAE: 11,8
- MAPE: 7,68%

Figura 12 – LSTM - Modelo 1



Fonte: Elaborado pelo Autor

4.2.2 LSTM - Modelo 2

Já o segundo modelo, figura [13](#), apresenta uma melhora significativa em relação ao primeiro, demonstrando uma maior acurácia em capturar as características da série, que pode ser comprovada pelo valor das métricas de avaliação. O [MSE](#) demonstrou uma melhora de cerca de 70% em relação ao primeiro modelo, o que significa uma redução na distância entre os valores previstos e os originais. Já o [MAE](#) diminuiu para 10,32 e o [MAPE](#) para 6,44%. Embora a diferença entre as duas últimas métricas não sejam tão expressivas, são o suficiente para evidenciar a redução dos erros de forma absoluta.

- MSE: 159.39
- MAE: 10.32
- MAPE: 6.44%

Figura 13 – LSTM - Modelo 2



Fonte: Elaborado pelo Autor

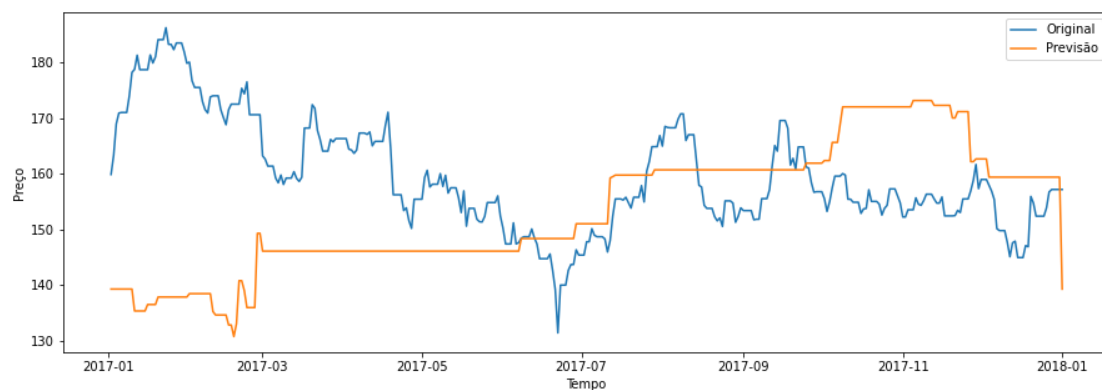
4.3 XGBoost

4.3.1 XGBoost - Modelo 1

Assim como o primeiro modelo [LSTM](#), o modelo [XGBoost](#) mais simples demonstra baixa capacidade em prever o preço do café arábica de forma precisa. O MSE de 358,22 deixa claro a falha do modelo em produzir previsões no mínimo próximas aos valores reais. Com um [MAE](#) de 14,54 e um [MAPE](#) de 8,82% o modelo performa pior que o primeiro modelo [LSTM](#). Como pode ser visto na figura 14.

- MSE: 358.22
- MAE: 14.54
- MAPE: 8.82%

Figura 14 – XGBoost - Modelo 1



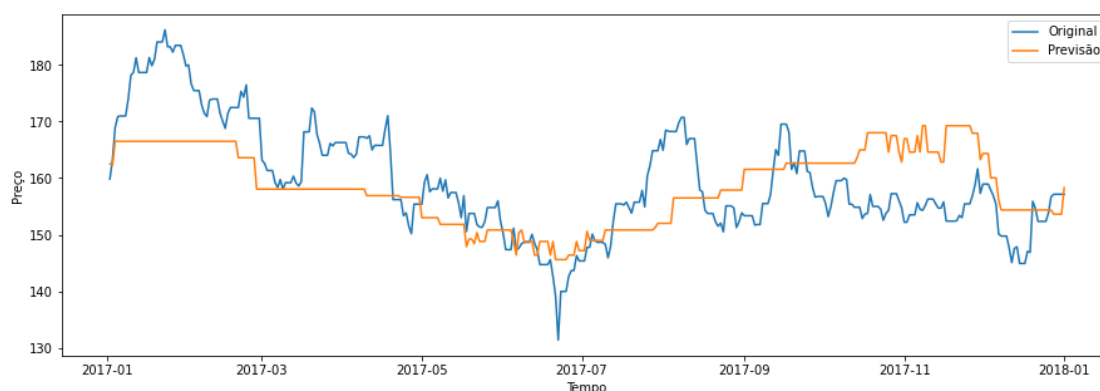
Fonte: Elaborado pelo Autor

4.3.2 XGBoost - Modelo 2

O segundo modelo, figura 15, **XGBoost**, também demonstra uma melhora significativa de performance quando a cotação do dólar em real é adicionada como *feature*. O **MSE** de 58,37 e o **MAPE** de 3,91%, redução de quase 5% no último indicador, evidencia a melhora obtida. O modelo foi o que demonstrou maior evolução com a adição da nova *feature*.

- MSE: 58.37
- MAE: 6.3
- MAPE: 3.91%

Figura 15 – XGBoost - Modelo 2



Fonte: Elaborado pelo Autor

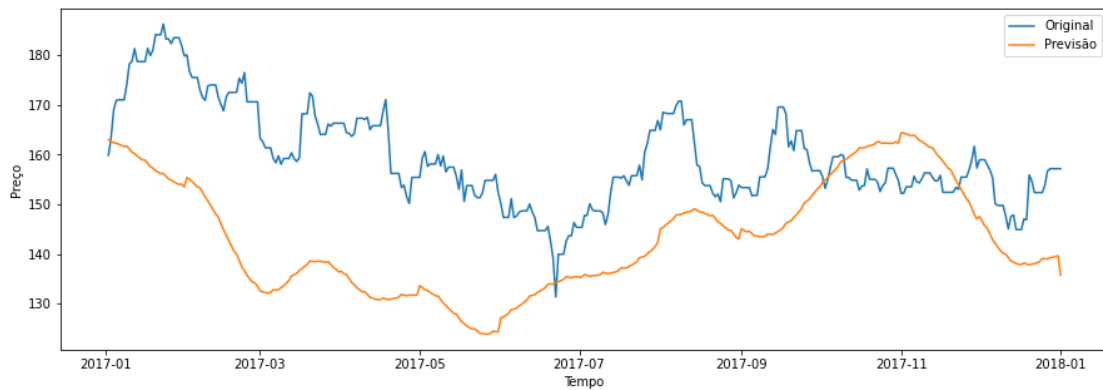
4.4 Prophet

4.4.1 Prophet - Modelo 1

O primeiro modelo *Prophet* apresentou resultados consideravelmente ruins quando comparado aos demais. O modelo falhou em capturar características mais simples da série, suas previsões se distanciaram muito do intervalo de valores corretos, o que fica comprovado pelo **MSE** de 410,95. O modelo foi o responsável pelo pior desempenho observado. O **MAPE** resultante mostra que a diferença média entre o valor previsto e o valor real é de 10,97%, um resultado que não pode ser considerado bom. Figura 16.

- MSE: 410.95
- MAE: 17,7
- MAPE: 10,97%

Figura 16 – Prophet - Modelo 1



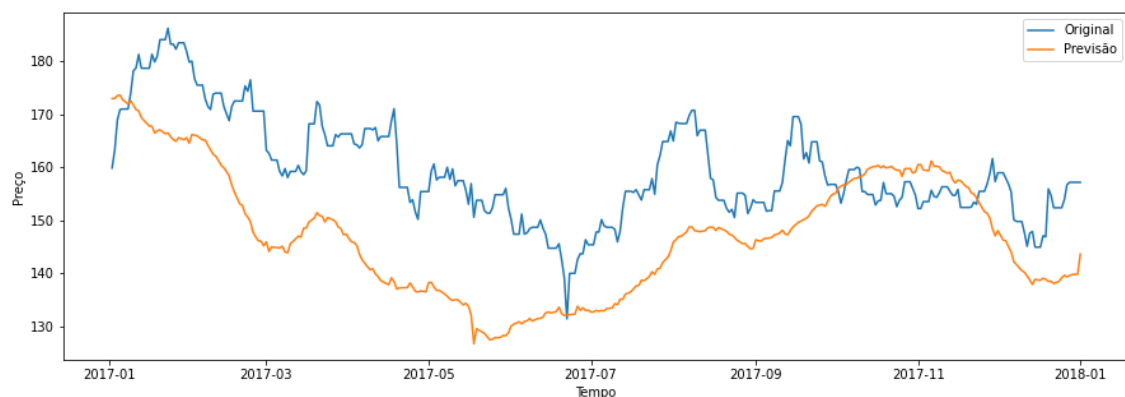
Fonte: Elaborado pelo Autor

4.4.2 Prophet - Modelo 2

Assim como todos os outros modelos, apresentou melhora significativa em sua acurácia quando a nova *feature* foi adicionada. No entanto, ainda não foi o suficiente para que o modelo conseguisse ser preciso em um nível próximo aos demais modelos criados. Além disso, foi o que menos se beneficiou da nova *feature*, pois apresentou uma diminuição de apenas 1,37% no **MAPE**, segundo pior resultado. Figura 17.

- MSE: 303.27
- MAE: 15.36
- MAPE: 9.6%

Figura 17 – Prophet - Modelo 2



Fonte: Elaborado pelo Autor

4.5 Considerações Finais

Os modelos construídos para realizar a previsão do preço do café, de forma geral não foram capazes de prever com alta acurácia o valor da *commodity*. No entanto, os resultados obtidos pelo segundo modelo [LSTM](#) e o segundo modelo [XGBoost](#), construídos utilizando o dólar como *feature*, demonstraram uma melhora significativa. O ganho de performance causado pelo dólar, deixa clara a influência da moeda americana no preço do café, uma vez que permitiu aos modelos que capturassem o comportamento da série com maior precisão.

5 Conclusão

Este trabalho procurou investigar o impacto do dólar americano no preço dos contratos de café arábica por através de algoritmos de aprendizagem de máquina. O objetivo era construir modelos utilizando o dólar como uma *feature* e depois comparar com outros modelos que não possuíam essa variável adicional. Por meio da revisão bibliográfica, foi possível estudar e selecionar os métodos a serem utilizados no projeto.

Após a análise das séries temporais, foi verificado uma forte correlação de $-0,73$ entre o preço da *commodity* e a moeda americana, o que se tornou uma prova preliminar de que o dólar poderia contribuir para a modelagem da série, uma vez que uma variável correlata à série alvo da predição tende a melhorar a acurácia dos modelos desenvolvidos. A correlação negativa chamou a atenção no primeiro momento, pois a concepção inicial era de que uma desvalorização do dólar seria refletida no preço do café, o que não foi corroborado pelos resultados obtidos.

De forma geral, todos os modelos utilizados no projeto se mostraram beneficiados pela adição do dólar. O modelo com melhor performance observado, **XGBoost**, apresentou uma evolução de cerca de 84%, já o modelo **LSTM**, embora não tenha performado tão bem quanto o **XGBoost**, conseguiu um ganho de performance de 29,5%. Mesmo o modelo com pior desempenho observado, *Prophet*, demonstrou uma melhora de 26%. A adição permitiu aos algoritmos capturar aspectos mais complexos da série, o que resultou uma previsão mais próxima da realidade.

Embora o desenvolvimento de modelos para a previsão de valores relacionados ao mercado financeiro seja uma tarefa complexa, a contribuição oferecida pela adição do dólar permite concluir que a moeda americana impacta fortemente no preço dos contratos de café arábica. O dólar e o os contratos de café, mostram um comportamento quase inversamente proporcional durante todo o período avaliando, um fato notável e inesperado.

5.1 Trabalhos Futuros

Para trabalhos futuros sugere-se a avaliação da cotação do real em dólar americano. A cotação da moeda pode ser utilizada para medir os impactos da situação econômica interna no preço do café, o que pode melhorar ainda mais a qualidade de modelos criados para a previsão do valor da *commodity*.

Referências

AYANKOYA, K.; CALITZ, A. P.; GREYLING, J. H. Real-time grain commodities price predictions in south africa: a big data and neural networks approach. *Agrekon*, Taylor & Francis, v. 55, n. 4, p. 483–508, 2016. Citado na página 17.

B3. *Café Arábica* 4/5. 2022. Disponível em: <https://www.b3.com.br/pt_br/produtos-e-servicos/negociacao/commodities/ficha-do-produto-8AE490CA6D41D4C7016D45F3CA44383E.htm>. Citado na página 14.

BOX, G. E.; JENKINS, G. M. Some comments on a paper by chatfield and prothero and on a review by kendall. *Journal of the Royal Statistical Society. Series A (General)*, JSTOR, v. 136, n. 3, p. 337–352, 1973. Citado na página 18.

CAMPOS, J. I. *A importância do café nosso de todos os dias*. 2005. Disponível em: <<https://www.embrapa.br/busca-de-noticias/-/noticia/17987068/a-importancia-do-cafe-nosso-de-todos-os-dias#:~:text=Hoje%2C%20o%20caf%C3%A9%20continua%20sendo,um%20ter%C3%A7o%20da%20produ%C3%A7%C3%A3o%20mundial.>> Citado na página 14.

CASTRO, E. R. d.; TEIXEIRA, E. C.; LIMA, J. E. d. Efeito da desvalorização cambial na oferta, no preço de insumos e na relação entre os fatores na cultura do café. *Revista de Economia e Sociologia Rural*, SciELO Brasil, v. 43, n. 3, p. 421–441, 2005. Citado na página 18.

CAVAIONI, M. *DeepLearning series: Sequence Models*. 2014. Disponível em: <<https://medium.com/machine-learning-bites/deeplearning-series-sequence-models7855babeb586>>. Citado na página 20.

CHEN, E. *Exploring LSTMs*. 2017. Disponível em: <<http://blog.echen.me/2017/05/30/exploring-lstms/>>. Citado na página 20.

DAI, Y.; ZHANG, Y. Machine learning in stock price trend forecasting. *Stanford University*, <http://cs229.stanford.edu/proj2013/DaiZhang-MachineLearningInStockPriceTrendForecasting.pdf>. *Erişim Tarihi*, v. 21, p. 2021, 2013. Citado na página 17.

DEINA, C. et al. A methodology for coffee price forecasting based on extreme learning machines. *Information Processing in Agriculture*, Elsevier, v. 9, n. 4, p. 556–565, 2022. Citado na página 17.

FOFANAH, A. J. Machine learning model approaches for price prediction in coffee market using linear regression, xgb, and lstm techniques. *International Journal of Scientific Research in Science and Technology*, n. 6, 2021. Citado 2 vezes nas páginas 17 e 20.

GONÇALVES, A. P. F. Definição de um modelo de inteligência artificial para a identificação do padrão curricular dos alunos do icea. *Biblioteca Digital de Trabalhos de Conclusão de Curso*, 2022. Disponível em: <<http://www.monografias.ufop.br/handle/35400000/4384>>. Citado na página 18.

- GRANGER, C. W.; NEWBOLD, P. *Economic Forecasting-the atheist's viewpoint*. [S.l.]: University of Nottingham, 1975. Citado na página 18.
- GREFF, K. et al. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, IEEE, v. 28, n. 10, p. 2222–2232, 2016. Citado na página 20.
- HEBB, D. O. The first stage of perception: growth of the assembly. *The Organization of Behavior*, Wiley, New York, v. 4, p. 60–78, 1949. Citado na página 19.
- HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural computation*, MIT Press, v. 9, n. 8, p. 1735–1780, 1997. Citado na página 20.
- ICMC. *Redes Neurais Artificiais*. 2003. Disponível em: <<https://sites.icmc.usp.br/andre/research/neural/#cara>>. Citado na página 19.
- KIRCHGÄSSNER, G.; WOLTERS, J.; HASSLER, U. *Introduction to modern time series analysis*. [S.l.]: Springer Science & Business Media, 2012. Citado na página 18.
- MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, Springer, v. 5, n. 4, p. 115–133, 1943. Citado na página 19.
- MITCHELL, T. et al. Machine learning. *Annual review of computer science*, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 4, n. 1, p. 417–433, 1990. Citado na página 19.
- PINTO, T. d. S. *Raízes do café no Brasil. Café no Brasil e Suas Origens*. 2022. Disponível em: <<https://brasilecola.uol.com.br/historia/o-cafe-no-brasil-suas-origens.htm>>. Citado na página 14.
- ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, American Psychological Association, v. 65, n. 6, p. 386, 1958. Citado na página 19.
- SELVIN, S. et al. Stock price prediction using lstm, rnn and cnn-sliding window model. In: IEEE. *2017 international conference on advances in computing, communications and informatics (icacci)*. [S.l.], 2017. p. 1643–1647. Citado na página 17.
- TAYLOR, S. J.; LETHAM, B. Forecasting at scale. *The American Statistician*, Taylor & Francis, v. 72, n. 1, p. 37–45, 2018. Citado na página 21.
- TINBERGEN, N. On the analysis of social organization among vertebrates, with special reference to birds. *American midland naturalist*, JSTOR, p. 210–234, 1939. Citado na página 18.