Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Aline Ferrarini Carassai

Aplicação de LSTM para previsão do preço do etanol hidratado no Brasil

Curitiba
2022

Aline Ferrarini Carassai

Aplicação de LSTM para previsão do preço do etanol hidratado no Brasil

Monografia apresentada ao Programa de Especialização em Data Science e Big Data da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Fernando de Pol Mayer

Curitiba

2022

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

Aplicação de LSTM para previsão do preço do etanol hidratado no Brasil

Aline Ferrarini Carassai[1]
Fernando de Pol Mayer[2]

Resumo

A volatilidade do preço do etanol hidratado praticado nos postos de combustíveis do Brasil, aliada à falta de clareza a respeito das variáveis que influenciam nesta variabilidade, resultam no desestímulo do consumo deste biocombustível por parte dos consumidores e no enfraquecimento da indústria nacional de etanol, que possui grande espaço para crescimento. Por este motivo, este artigo objetivou aplicar uma arquitetura de Redes Neurais Recorrentes denominada Long Short-Term Memory (LSTM) para realizar a previsão do preço do etanol hidratado e orientar consumidores e produtores desde combustível. Foram coletadas séries históricas do preço do etanol e de variáveis correlacionadas e a LSTM foi aplicada utilizando dois métodos. Em um deles, somente a série histórica do preço do etanol foi utilizada como variável de entrada e, em outro, as séries de todas as variáveis correlacionadas (bem como a do preço do etanol) foram utilizadas. A partir dos resultados obtidos, observou-se menor erro percentual médio para a LSTM em que foram utilizadas todas as variáveis como entrada, e menores erros absolutos para o método utilizando somente os preços do etanol.
Palavras-chave: LSTM, séries históricas, etanol, redes neurais.

Abstract

The volatility of the hydrous ethanol price observed in the gas stations around Brazil, together with the lack of clarity regarding the variables that influence in the variability of these prices, result in the discouragement of this biofuel usage by consumers, as well as in the weakening of ethanol national industry. For this reason, this article aimed to apply a Recurrent Neural Network architecture called Long Short-Term Memory (LSTM) to achieve the forecast of hydrous ethanol price and provide guidance to consumers and producers of this fuel. Time series of the hydrous ethanol price, as well as of other correlated variables, were collected and the LSTM was applied using two methods. For the first method, only the ethanol price time series was used as an input variable and, for the second one, the time series of all variables (including the one for ethanol price) were used. Based on the obtained results, we could observe a smaller average percentage error for the method where all the variables were used as input for the LSTM, and smaller absolute errors for the method where only the ethanol price time series was used.
***Keywrods**: LSTM, time series, ethanol, neural networks.*

1 Introdução

De acordo com dados da ANP (Agência Nacional do Petróleo, Gás Natural e Biocombustíveis) de 2018, o Brasil se consolida como segundo maior produtor de etanol do mundo. Em 2021, a produção de etanol hidratado compreendeu 38,1% de todo o etanol produzido no país[1]. Com ações como o RenovaBio, promovido pelo Ministério de Minas e Energia (MME), que pretende expandir a participação de biocombustíveis na matriz energética do país, esta produção só tende a aumentar nos próximos anos.

Ainda assim, na prática, o que os consumidores presenciam nos postos de combustíveis são preços pouco competitivos para o etanol hidratado frente ao preço da gasolina e com volatilidade comparável à dos combustíveis fósseis. Tais condições acabam desestimulando o consumo de biocombustíveis como o etanol hidratado e fortalecendo a indústria internacional de petróleo em detrimento da nacional de biocombustíveis.

O objetivo do presente estudo é, então, obter um modelo capaz de predizer com acuracidade o preço médio do etanol hidratado nos postos de combustíveis brasileiros, almejando orientar produtores de etanol e consumidores em suas estratégias. Neste processo, foram comparados e analisados os resultados das predições obtidas pela aplicação da técnica *Long Short-Term Memory* (LSTM).

2 Materiais e Métodos

A base de dados para modelagem foi obtida através do site do Cepea (Centro de Estudos Avançados em Economia Aplicada)[2] e consiste em uma série temporal com observações diárias do preço médio do etanol hidratado praticado nos postos de combustíveis do Brasil entre Fevereiro de 2010 e Outubro de 2021 (4261 *timesteps*).

A partir da leitura de análises setoriais, também foram colhidas séries temporais de variáveis que teriam pos-

[1]Aluno do programa de Especialização em Data Science & Big Data, alinecarassai@gmail.com.
[3]Professor do Departamento de Estatística - DEST/UFPR.

sível correlação com o preço do etanol hidratado. Tais séries foram obtidas nos sites do Cepea (preço médio do açúcar), ANP (preço médio da gasolina) e Investing (cotação USD/BRL).

As unidades de medida para cada uma das variáveis estão descritas abaixo:

- Preço do etanol hidratado: R$/l
- Preço do açúcar: R$/50kg
- Preço da gasolina: R$/l
- Cotação: USD/BRL

Para uma melhor visualização da relação entre as variáveis obtidas, como também sua tendência e sazonalidade, foi plotado o gráfico abaixo com seis meses de observações.

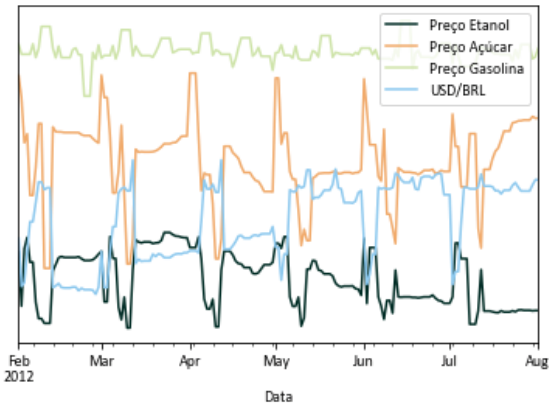

Figura 1: Visualização das variáveis.

A partir da análise do gráfico, já é possível notar a interdependência entre as variáveis, indicando que, de fato, podem possuir boa capacidade de ajuste do modelo de regressão. Picos no preço do açúcar, por exemplo, quase sempre estão acompanhados de picos no preço do etanol. Este comportamento ilustra bem situações em que o açúcar está valorizado e os produtores optam por destinar a cana para a produção deste produto ao invés do etanol, gerando uma queda na oferta de etanol e aumentando seu preço por consequência.

Através da matriz de correlação, foi possível constatar que as variáveis colhidas apresentavam, de fato, correlação acima de 0,5 com a variável dependente.

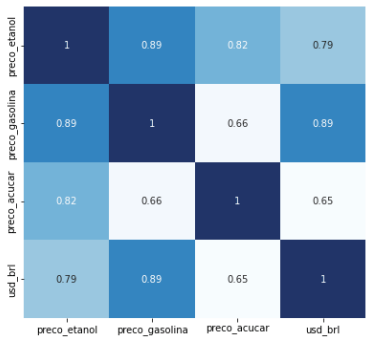

Figura 2: Matriz de Correlação das variáveis.

Com o objetivo de analisar a correlação entre as observações das séries temporais, também foi plotada a Função de Autocorrelação para cada uma das variáveis[3].

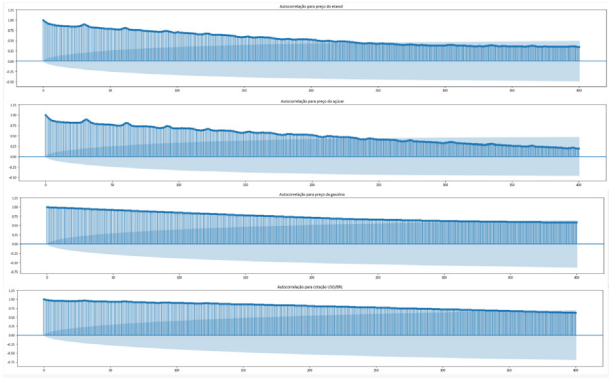

Figura 3: Função de Autocorrelação das variáveis.

A partir dos gráficos acima, é possível notar que, para o preço do etanol e do açúcar, a partir de 250 observações a correlação com a primeira observação deixa de ser significativa. Para o preço da gasolina e cotação do dólar, no entanto, esta correlação permanece significativa por cerca de 400 observações.

2.1 Preparo dos dados

Para algumas das variáveis, as séries temporais não continham as observações de preços para os finais de semana. Assim, foi utilizado o método *forward fill* para preencher os *gaps* dessas informações faltantes com o preço observado imediatamente anterior a estes intervalos.

Além disso, as variáveis do conjunto de dados são apresenadas em escalas diferentes, o que poderia enviesar os modelos adotados. Dessa forma, os dados foram normalizados utilizando a técnica *MinMaxScaler*, redimensionando-os para um intervalo entre 0 e 1.

2.2 Modelo empregado

Com objetivo de avaliar se as variáveis correlacionadas teriam impacto positivo na predição do preço do etanol, optou-se pela aplicação da LSTM através de dois métodos:

- Primeiro método: Para o primeiro método, apenas a série histórica do preço do etanol foi utilizada como variável de entrada para a LSTM. O objetivo foi treinar o modelo com o comportamento histórico do próprio preço do etanol.
- Segundo método: Para o segundo método, foram utilizadas como variáveis de entrada para a LSTM as séries históricas do preço do etanol, do preço da gasolina, preço do açúcar e a cotação USD/BRL. O objetivo foi avaliar se, treinando o modelo com o comportamento de múltiplas variáveis correlacionadas com o preço do etanol, teríamos maior precisão nas predições.

Para a aplicação de ambos os métodos, a base de dados foi dividida em 80% para utilização no treino do modelo, e 20% para teste. O modelo foi, então, ajustado utilizando a base de treino e os valores para o preço do etanol foram preditos utilizando a base de teste. Por fim, os valores preditos foram comparados aos reais da base de teste.

2.2.1 LSTM

A LSTM é uma arquitetura de rede neural recorrente (RNN) que busca atribuir o devido peso a observações passadas de uma série temporal. Por sua capacidade de "lembrar" de sequências de dados passados, ela é adequada para prever séries temporais independentemente do intervalo fornecido.

Ao contrário de RNNs tradicionais, no entanto, a LSTM possui a capacidade de armazenar informações de longo prazo e, a cada recorrência, "avaliar"se determinada informação anterior pode ser esquecida ou deve ser passada adiante. Assim, a LSTM constitui uma boa técnica para aplicação em séries temporais com tendência de longo prazo.

Para a implementação da LSTM, utilizou-se a biblioteca *keras* para criar um modelo sequencial e adicionar camadas. A função de ativação utilizada foi a ReLU (para produzir resultados no intervalo [0,∞[). O modelo foi, então, compilado utilizando a função perda como sendo o Erro Médio Quadrático (MSE).

Após a definição e compilação do modelo, o mesmo foi treinado utilizando a base de treino correspondente a 80% da base de dados. Um dos parâmetros passados para o modelo foi que o mesmo deveria ser treinado 200 vezes (200 épocas), ou até que a função perda não tivesse redução significativa. Foram utilizadas duas camadas escondidas com 100 e 50 neurônios, respectivamente. Para o treinamento do modelo, foi estabelecida uma "janela deslizante"de 5 observações. Ou seja, a cada conjunto de 5 observações consecutivas, o modelo era treinado para prever a observação seguinte (sexta observação). O teste do modelo foi realizado utilizando a base com 20% dos dados restantes, e foi plotado o gráfico de observações reais vs. observações preditas para cada um dos métodos aplicados.

2.3 Métricas de desempenho

Para analisar o desempenho de cada um dos métodos empregados, também foi calculado o Erro Médio Absoluto (MAE), Erro Médio Percentual Absoluto (MAPE) e Raiz do Erro Quadrático Médio (RMSE). Cada uma destas medidas de erro são calculadas conforme as fórmulas abaixo.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t|$$

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

O MAE, neste caso, constitui uma medida de erro de fácil interpretação, por apresentar seu valor de saída em escala igual à da série de preços do etanol. Entretanto, o mesmo não considera a representatividade deste erro frente ao valor real em termos percentuais, podendo afetar a interpretação quando a variável observada assume valores com grande amplitude ao longo da série temporal.

Para evitar este problema, o MAPE constitui uma medida de erro apropriada em variáveis com grande amplitude por representar a média do erro percentual para cada observação.

O RMSE, por sua vez, também é apresentado em mesma escala da série de preços do etanol, entretanto caracteriza-se pela maior penalização de erros muito grandes entre valores observados e preditos.

3 Resultados e Discussões

Para observação da precisão do modelo obtido por cada um dos métodos, plotou-se o gráfico de observações reais versus preditas nas bases de treino e teste.

As figuras 4 e 5 ilustram os resultados de cada um dos modelos na base de treino:

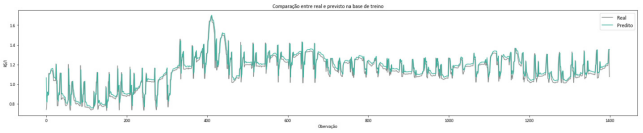

Figura 4: Valores reais vs. preditos para o método 1 (uma variável de entrada).

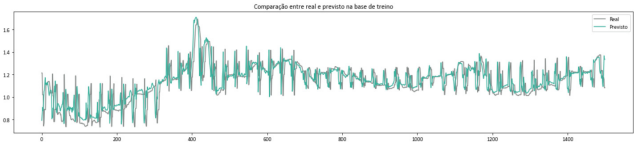

Figura 5: Valores reais vs. preditos para o método 2 (múltiplas variáveis de entrada).

As medidas de erro para cada um dos modelos na base de treino foram:

Tabela 1: Medidas de erro na base de treino

Erro	Método 1	Método 2
MAE	0.0527	0.0992
RMSE	0.0926	0.1570
MAPE	25.4694%	7.4082%

A partir dos resultados acima, é possível observar que o método 1, aplicando a LSTM apenas com a série de preços do etanol como variável de entrada, apresentou bom ajuste e boas medidas de erro absoluto. O MAPE, entretanto, apresentou melhores resultados para o método 2 na base de treino (mostrando que, percentualmente, o método 2 apresentou valores preditos mais próximos dos reais, na média).

Na base de teste, por sua vez, os resultados para os dois métodos foram plotados a seguir:

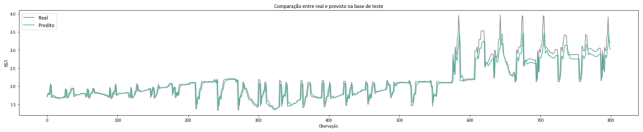

Figura 6: Valores reais vs. preditos para o método 1 na base de teste.

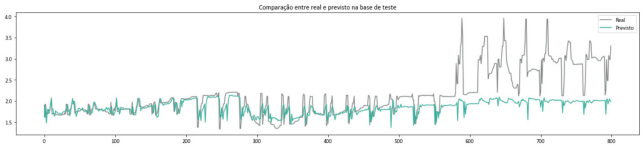

Figura 7: Valores reais vs. preditos para o método 2 na base de teste.

As medidas de erro obtidas na base de teste para cada um dos métodos foram:

Tabela 2: Medidas de erro na base de teste

Erro	Método 1	Método 2
MAE	0.1209	0.3897
RMSE	0.2092	0.6011
MAPE	26.5561%	14.6172%

Como esperado, os resultados para ambos os modelos comparando valores reais vs. preditos na base de teste apresentaram menor precisão do que os resultados apresentados na base de treino. Assim como na base de treino, entretanto, o método 1 apresentou melhores medidas de erro absoluto, mas maior MAPE.

O modelo ajustado pelo método 2, com as 4 séries históricas como variáveis de entrada, apresentou um descolamento dos valores preditos vs. reais nas últimas 300 observações da base de teste. Ainda assim, o erro percentual médio para este método permaneceu em 14,62%.

Analisando o mercado de etanol hidratado para o periodo referente às últimas 300 observações da base de teste, observou-se que o periodo de descolamento coincidiu com o início da pandemia de Covid-19 no Brasil. Neste periodo, houve queda brusca na demanda por combustíveis no mundo, o que levou o preço do barril do petróleo a atingir valores mínimos. O etanol, como combustível concorrente da gasolina, teve que ter seu preço reduzido, como consequência.

Por outro lado, no mesmo periodo, houve forte desvalorização do real, elevando a cotação USD/BRL. Neste cenário, os produtores optaram por aumentar a produção de açúcar refinado para exportação em detrimento do etanol[4].

4 Conclusões

O presente artigo objetivou utilizar a técnica *Long Short-Term Memory* (LSTM) para realizar predições do preço do etanol hidratado revendido nos postos de combustíveis do país, e estabelecer uma comparação entre métodos utilizados para aplicação da técnica.

A partir dos resultados obtidos, observou-se que com a utilização apenas da própria série histórica dos preços do etanol hidratado como variável de entrada para a LSTM, foram atingidas melhores medidas de erro absoluto. Entretanto, este método apresentou resultados elevados para o Erro Médio Percentual Absoluto, indicando que, possivelmente, o modelo obteve maiores erros em suas predições nos periodos em que o valor do etanol hidratado era menor.

O método utilizando mais de uma série histórica (de variáveis correlacionadas com o preço do etanol hidratado), por sua vez, apresentou medidas de erro absoluto elevadas, mas melhores resultados para o erro médio percentual. Mesmo com o descolamento das predições nas últimas observações da base de teste, coincidentes com o periodo da pandemia de covid-19 no Brasil, este método errou em média 14,62% em relação aos valores reais na base de teste.

Agradecimentos

Deixo, ao final deste artigo, meu agradecimento ao Professor Orientador Fernando Mayer por sua paciência, auxílio e importantes contribuições durante o desenvolvimento deste projeto.

Agradeço, também, a minha família, por todo o suporte e incentivo à continuidade e finalização deste trabalho.

Referências

[1] Painel dos combustíveis. https://www.gov.br/anp/pt-br/centrais-de-conteudo/paineis-dinamicos-da-anp/paineis-e-mapa-dinamicos-de-produtores-de-combustiveis-e-painel-dinamico-de-produtores-de-etanol. Accessed: 2022-06-01.

[2] Séries históricas. https://www.cepea.esalq.usp.br/br/consultas-ao-banco-de-dados-do-site.aspx. Accessed: 2022-06-01.

[3] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer New York, 2014.

[4] Produção e mercado do etanol. https://www.bnb.gov.br/etene/caderno-setorial. Accessed: 2022-06-01.