

Modelo híbrido LSTM e Regressão Linear para Previsão do Mercado Financeiro

IGOR M. NASCIMENTO

O presente estudo aborda a aplicação de modelos híbridos de previsão de séries temporais no mercado financeiro, com foco no Bitcoin. Utilizamos a combinação de Redes Neurais Long Short-Term Memory (LSTM) e Regressão Linear para capturar tanto padrões não lineares de longo prazo quanto tendências lineares presentes nos dados. Foram testados diversos modelos, incluindo Regressão Linear, Lasso, Ridge e LSTM puro, além do modelo híbrido proposto. Nossos resultados mostram que a hibridização dessas abordagens oferece uma melhor performance em termos de precisão e capacidade preditiva, especialmente em cenários de alta volatilidade. A metodologia incluiu a normalização dos dados utilizando o MinMaxScaler e a divisão em treino, validação e teste, além da utilização de métricas como MSE, MAE, RMSE, R^2 e MAPE para avaliação dos modelos. O estudo destaca a eficiência da combinação de técnicas tradicionais com aprendizado profundo para previsão de preços em mercados financeiros.

ACM Reference Format:

Igor M. Nascimento. 2024. Modelo híbrido LSTM e Regressão Linear para Previsão do Mercado Financeiro. 1, 1 (September 2024), 16 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

ABSTRACT

This study addresses the application of hybrid time series forecasting models in the financial market, focusing on Bitcoin. We employed a combination of Long Short-Term Memory (LSTM) Neural Networks and Linear Regression to capture both long-term nonlinear patterns and linear trends in the data. Several models were tested, including Linear Regression, Lasso, Ridge, and pure LSTM, in addition to the proposed hybrid model. Our results show that the hybridization of these approaches offers better performance in terms of accuracy and predictive capacity, especially in highly volatile scenarios. The methodology included data normalization using MinMaxScaler and splitting into training, validation, and test sets. Evaluation metrics such as MSE, MAE, RMSE, R^2 , and MAPE were used to assess the models. The study highlights the efficiency of combining traditional techniques with deep learning for price forecasting in financial markets.

1 INTRODUÇÃO

A previsão de preços de ações desempenha um papel crucial no mercado financeiro, oferecendo insights valiosos para investidores, gestores de ativos e analistas de mercado. A volatilidade dos mercados financeiros torna a previsão de preços de ações um desafio significativo, impulsionando a busca por métodos preditivos mais eficientes. Ao longo dos anos, diversas técnicas de análise de séries temporais têm sido aplicadas para enfrentar esse desafio, com destaque para métodos tradicionais como o ARIMA (Auto-Regressive Integrated Moving Average), amplamente utilizado para dados lineares e estáticos.

Author's address: Igor M. Nascimento.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Association for Computing Machinery.

Manuscript submitted to ACM

Manuscript submitted to ACM

1

Com o rápido avanço do aprendizado de máquina, surgiram abordagens mais robustas e capazes de lidar com a não linearidade e a alta dimensionalidade dos dados financeiros. Modelos como Redes Neurais Recorrentes (RNN) e sua variante mais sofisticada, Long Short-Term Memory (LSTM), tornaram-se populares. Esses modelos são eficazes na captura de padrões temporais complexos devido à sua capacidade de armazenar e utilizar informações passadas ao longo do tempo. No entanto, a simplicidade e a eficácia de métodos como a Regressão Linear ainda são atrativas, especialmente para casos em que os dados apresentam padrões mais simples e lineares.

Neste trabalho, comparamos a Regressão Linear e o LSTM na tarefa de prever preços de ações. Além disso, exploramos a hibridização dessas duas abordagens, combinando os resultados de ambos os modelos para avaliar se a combinação pode melhorar a precisão das previsões em cenários onde as características dos dados variam entre comportamentos lineares e não lineares. Este estudo visa fornecer uma análise abrangente sobre a eficácia desses métodos isoladamente e em conjunto, contribuindo para a literatura de previsões financeiras com uma abordagem híbrida.

2 OBJETIVO

Este estudo tem como objetivo investigar e comparar a precisão preditiva de dois modelos amplamente utilizados em previsão de séries temporais: Regressão Linear e LSTM (Long Short-Term Memory). Utilizando dados de séries temporais de preços de ações, analisaremos os cenários em que cada abordagem se destaca, considerando suas características intrínsecas.

Nosso foco principal é avaliar a eficácia da Regressão Linear, uma técnica tradicional, em comparação com o LSTM, um modelo de aprendizado profundo projetado para lidar com padrões não lineares e complexos. Além disso, propomos a hibridização desses dois métodos, buscando combinar suas forças. Da mesma forma, avaliamos a combinação dessas técnicas, com o objetivo de capturar tanto padrões lineares quanto não lineares em séries temporais.

Assim, o estudo pretende determinar se essa abordagem híbrida pode oferecer melhorias significativas na previsão de preços, comparada ao uso isolado de cada método, e explorar as situações em que a regressão linear pode se mostrar competitiva com técnicas mais avançadas.

3 TRABALHOS RELACIONADOS

A previsão de preços de ações tem sido amplamente explorada na literatura, utilizando diferentes modelos e técnicas, que vão desde abordagens tradicionais até o uso de redes neurais profundas e métodos híbridos. Selvin et al. (2017) investigaram a aplicação de Redes Neurais Recorrentes (RNN), Long Short-Term Memory (LSTM) e Redes Neurais Convolucionais (CNN) para a previsão de preços de ações utilizando um modelo de janela deslizante. O estudo demonstrou que os modelos profundos são eficazes na identificação de padrões latentes em dados temporais, particularmente em cenários voláteis, como o mercado financeiro [7]. Esses resultados destacam a capacidade das redes profundas de lidar com a complexidade das séries temporais, identificando padrões que métodos tradicionais, como ARIMA, falham em capturar [7].

Patidar et al. (2023) compararam a eficácia entre o LSTM e a Regressão Linear na previsão de preços de ações. Embora a Regressão Linear ainda possa apresentar resultados aceitáveis em determinados cenários de padrões lineares, especialmente em mercados menos voláteis, o estudo mostra que o LSTM tem uma vantagem clara na captura de dependências temporais e padrões não lineares, o que é crucial em mercados de alta volatilidade [6]. Esse estudo também é relevante para a proposta de um modelo híbrido, como o que estamos desenvolvendo, pois explora a sinergia entre modelos simples e complexos para diferentes condições de mercado.

Panigrahi e Mantri (2015) propuseram um modelo baseado em árvore de decisão, utilizando dados textuais para prever o comportamento do mercado de ações. Eles demonstraram que a integração de dados textuais com dados numéricos pode aumentar a precisão das previsões, especialmente quando combinada com algoritmos de transformação de dados [5]. Isso é particularmente relevante no contexto de modelos híbridos que podem combinar diferentes fontes de informação para obter previsões mais robustas.

Além disso, Karim et al. (2021) investigaram a Regressão Linear e a Regressão com Árvore de Decisão, comparando seus desempenhos na previsão do mercado de ações. Eles mostraram que, apesar da simplicidade, a Regressão Linear ainda oferece precisão significativa em grandes volumes de dados. A combinação de tais abordagens tradicionais com redes neurais mais avançadas pode levar a um ganho substancial de desempenho [3].

Em um estudo sobre a análise do mercado de ações, Crisostomo et al. (2023) utilizaram Regressão Linear para prever o fechamento de preços de ações. Eles argumentaram que, em cenários com padrões de dados mais previsíveis, a simplicidade da Regressão Linear pode ser suficiente para fornecer previsões precisas. No entanto, sua eficiência diminui quando há uma alta complexidade e não-linearidade nos dados [1].

Ghosh et al. (2021) destacaram o uso de modelos híbridos que combinam LSTM e Random Forests para prever o movimento direcional dos preços de ações. O estudo é relevante para a hibridização dos modelos, pois demonstra que a combinação de abordagens que capturam diferentes aspectos do mercado pode levar a previsões mais precisas [2].

Por fim, Li et al. (2024) propuseram o modelo FinReport, que utiliza análise de fatores de notícias para prever ganhos de ações. O estudo enfatiza a importância de integrar dados financeiros tradicionais com análise de sentimento derivada de fontes de notícias, demonstrando que tal combinação pode melhorar substancialmente a previsão de preços de ações [4]. Esse conceito de combinação de diferentes fontes de dados é diretamente aplicável à nossa proposta de modelo híbrido.

Com base nesses estudos, o presente trabalho visa expandir o estado da arte ao investigar a hibridização de modelos de Regressão Linear e LSTM. Ao combinar a capacidade da Regressão Linear de capturar padrões lineares de curto prazo com a habilidade do LSTM de modelar dependências temporais complexas, espera-se melhorar a precisão das previsões no mercado financeiro. Além disso, a exploração de dados textuais, como a análise de sentimento, pode ser uma extensão promissora para futuras iterações deste modelo híbrido.

4 METODOLOGIA

Neste estudo, utilizamos dados históricos de preços de Bitcoin fornecidos pela API da Binance. Os dados foram coletados em intervalos de uma hora, abrangendo o período de 1º de janeiro a 31 de dezembro dos anos de 2021, 2022 e 2023, e de 1º de janeiro a 14 de setembro de 2024. A base de dados contém os preços de fechamento por hora, oferecendo um cenário robusto para experimentos com modelos preditivos de séries temporais.

4.1 Preparação dos Dados

Os dados foram normalizados utilizando a técnica de *MinMaxScaler* para garantir que as diferentes escalas dos valores não impactassem o desempenho dos modelos. O *fit* do *MinMaxScaler* foi realizado apenas no conjunto de dados de treino, garantindo que os dados de validação e teste fossem normalizados com os mesmos parâmetros do treino, evitando qualquer tipo de vazamento de dados.

Além disso, ao incorporar os indicadores técnicos como SMA, EMA, RSI, e outros, identificamos a presença de valores *NaN* em algumas colunas, resultantes do cálculo desses indicadores em janelas de tempo. Em vez de remover esses

registros, optamos por substituir os valores *NaN* pelo valor mais próximo subsequente (*forward fill*). Essa abordagem garantiu que não houvesse perda de informações e que as séries temporais se mantivessem consistentes para os modelos.

O conjunto de dados foi dividido da seguinte maneira:

- **70% para treino:** utilizado para o ajuste dos modelos.
- **15% para validação:** utilizado para o ajuste dos hiperparâmetros e controle de *overfitting* com o uso de *Early Stopping*.
- **15% para teste:** utilizado para avaliação final da performance preditiva dos modelos.

A normalização foi aplicada separadamente para as variáveis de entrada (preços e indicadores técnicos) e saída (valores preditivos), mantendo a consistência durante o treinamento e a avaliação.

4.2 Escolha dos Hiperparâmetros

A escolha dos hiperparâmetros foi baseada em análises prévias, como correlações entre diferentes janelas de tempo e os preços de fechamento, bem como a correlação entre os indicadores financeiros e o preço de fechamento.

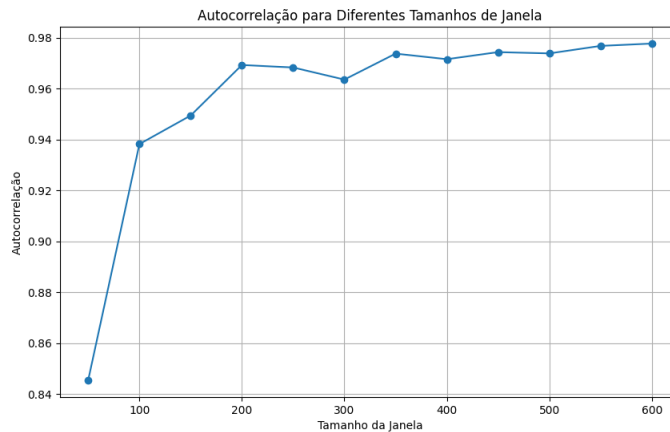


Fig. 1. Autocorrelação para diferentes tamanhos de janela

Na Figura 1, podemos observar que os valores de autocorrelação atingem o pico para janelas em torno de 350 e 400 horas, indicando que essas janelas capturam bem as dependências temporais nos dados. Assim, optamos por testar os modelos com janelas de 350 e 600 horas, a fim de avaliar o impacto de janelas maiores no desempenho preditivo.

Além disso, foi realizada uma análise de correlação entre diversos indicadores técnicos e o preço de fechamento do Bitcoin, conforme mostrado na Figura 2.

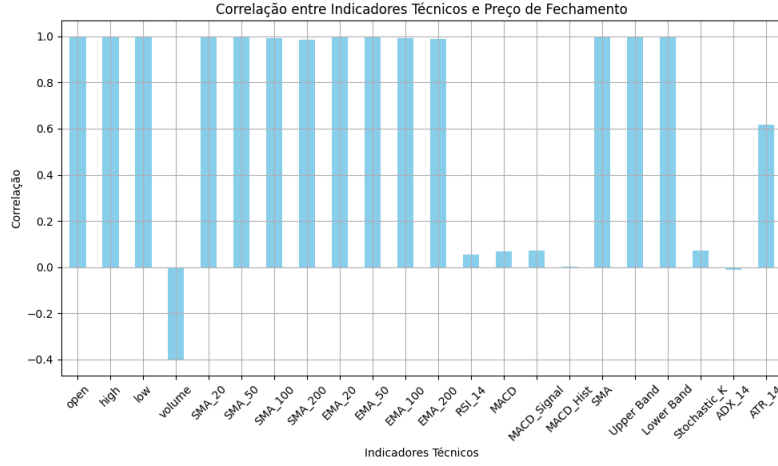


Fig. 2. Correlação entre indicadores técnicos e preço de fechamento

Os indicadores com maior correlação, como as médias móveis (SMA e EMA), foram selecionados para serem utilizados no modelo, eliminando assim redundâncias e ruídos que poderiam comprometer a performance dos modelos.

4.3 Modelos Testados

Foram implementados quatro tipos de modelos de aprendizado de máquina para prever os preços futuros do Bitcoin:

- (1) **Regressão Linear:** Um modelo linear simples sem regularização, utilizado para prever tanto um único valor futuro quanto múltiplos valores (10 passos à frente). A regressão linear é representada pela seguinte equação:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

onde y é o valor predito, X_i são as variáveis independentes, e β_i são os coeficientes ajustados pelo modelo.

- (2) **Lasso Regression:** Um modelo de regressão linear com regularização L1, que adiciona um termo de penalização à função de perda, com o objetivo de evitar overfitting. O modelo é ajustado para minimizar:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^p |\beta_j|$$

onde $\alpha = 0.1$ foi o valor usado, e β_j são os coeficientes do modelo.

- (3) **Ridge Regression:** Semelhante ao Lasso, porém utiliza regularização L2, que penaliza grandes coeficientes de forma quadrática. O modelo é ajustado para minimizar:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^p \beta_j^2$$

onde $\alpha = 1.0$ foi o valor usado neste estudo.

- (4) **Long Short-Term Memory (LSTM):** Um tipo de rede neural recorrente projetada para capturar padrões temporais de longo prazo. Cada célula LSTM consiste de três portas (input, output, forget), onde utilizamos as

funções de ativação *sigmoid* para controlar as portas e *tanh* para gerar os estados de célula. As fórmulas de atualização são:

$$\begin{aligned} f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (\text{Forget Gate}) \\ i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (\text{Input Gate}) \\ \tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (\text{Cell State Update}) \\ C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \quad (\text{New Cell State}) \\ o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (\text{Output Gate}) \\ h_t &= o_t * \tanh(C_t) \quad (\text{New Hidden State}) \end{aligned}$$

O LSTM foi configurado com 2 camadas, *batch* = 16, unidades = 70 e *dropout* de 0.4, com janelas de 350 e 600 horas. Foram testados mais de 360 combinações de parâmetros até chegar à configuração ideal.

4.4 Hiperparâmetros Utilizados

Os modelos foram configurados com os seguintes hiperparâmetros:

- **Regressão Linear:** janelas de 350 e 600 horas.
- **Lasso Regression:** $\alpha = 0.1$, janelas de 350 e 600 horas.
- **Ridge Regression:** $\alpha = 1.0$, janelas de 350 e 600 horas.
- **LSTM:** *batch* = 16, unidades = 70, *dropout* = 0.4, janelas de 350 e 600 horas.

A escolha dos hiperparâmetros foi baseada em experimentos preliminares que visavam otimizar a performance dos modelos.

4.5 Prevenção de *Overfitting*

Para evitar o *overfitting*, aplicamos *Early Stopping* nos modelos LSTM. Esse método monitora a perda no conjunto de validação e interrompe o treinamento se não houver melhora após 5 épocas consecutivas, restaurando os pesos da melhor época.

4.6 Hibridização dos Modelos

Para capturar tanto os padrões lineares quanto os não lineares presentes nas séries temporais, realizamos uma hibridização dos modelos de Regressão Linear sem regularização (que apresentou o melhor desempenho entre os métodos lineares) e o LSTM. A combinação foi feita utilizando uma média ponderada das previsões de ambos os modelos, utilizando a seguinte fórmula:

$$y_{\text{híbrido}} = 0.5 \cdot y_{\text{LSTM}} + 0.5 \cdot y_{\text{linear}}$$

onde y_{LSTM} representa as previsões do modelo LSTM, e y_{linear} representa as previsões do modelo de Regressão Linear. A ponderação foi igual para ambos os modelos, visando explorar as capacidades dos modelos lineares de capturar padrões simples e dos LSTM de lidar com dependências temporais complexas.

4.7 Avaliação dos Modelos

Os modelos foram avaliados utilizando as seguintes métricas:

- **Mean Squared Error (MSE):** $MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$
- **Mean Absolute Error (MAE):** $MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$
- **Root Mean Squared Error (RMSE):** $RMSE = \sqrt{MSE}$
- **R-squared (R²):** $R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$
- **Mean Absolute Percentage Error (MAPE):** $MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$

Essas métricas foram calculadas tanto para as previsões de uma única hora quanto para múltiplas horas à frente (10 passos à frente).

5 RESULTADOS

Nesta seção, apresentamos uma análise detalhada dos resultados obtidos por diferentes modelos preditivos aplicados à previsão do preço do Bitcoin. Foram realizados vários experimentos para avaliar o desempenho dos seguintes modelos: Regressão Linear, Lasso, Ridge, LSTM e o Modelo Híbrido (LSTM + Regressão Linear). Testamos duas configurações principais de janela de entrada: 350 horas e 600 horas. Em cada configuração, os modelos foram treinados e avaliados com previsão de uma única saída e múltiplas saídas.

As métricas utilizadas para avaliação de desempenho foram: **Mean Squared Error (MSE)**, **Mean Absolute Error (MAE)**, **R-squared (R²)**, **Mean Absolute Percentage Error (MAPE)**, e **Root Mean Squared Error (RMSE)**. Estas métricas foram aplicadas uniformemente para garantir uma comparação precisa entre os modelos.

Além disso, também exploramos a hibridização dos modelos, combinando a simplicidade e capacidade de generalização da Regressão Linear com a capacidade do LSTM de capturar padrões não lineares e dependências temporais complexas.

A seguir, detalhamos os resultados obtidos para cada uma das configurações de janelas de tempo e discutimos as previsões para diferentes modelos testados.

5.1 Resultados para Janela de 350 Horas (Saída Única)

Os primeiros experimentos foram conduzidos utilizando uma janela de entrada de 350 horas e uma saída única. Os resultados de cada modelo testado podem ser observados na Tabela 3.

Table 1. Métricas de Desempenho para Janela de 350 horas (Saída Única)

| Modelo | MSE | MAE | R ² | MAPE | RMSE |
|-------------------------|-------------|---------|----------------|--------|---------|
| Regressão Linear | 18494.22 | 83.49 | 0.9995 | 0.0025 | 135.99 |
| Lasso Regression | 58526751.08 | 6739.49 | -0.5700 | 0.2311 | 7650.28 |
| Ridge Regression | 22816.34 | 89.17 | 0.9994 | 0.0027 | 151.05 |
| LSTM | 79300.85 | 169.24 | 0.9979 | 0.0050 | 281.60 |
| Modelo Híbrido | 33597.58 | 109.34 | 0.9991 | 0.0033 | 183.30 |

Os gráficos a seguir mostram as previsões feitas pelos modelos para a janela de 350 horas.

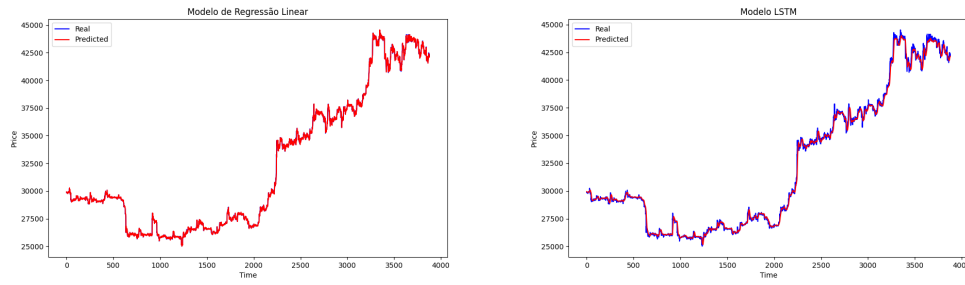


Fig. 3. Previsões da Regressão Linear (esquerda) e LSTM (direita) para a Janela de 350 horas

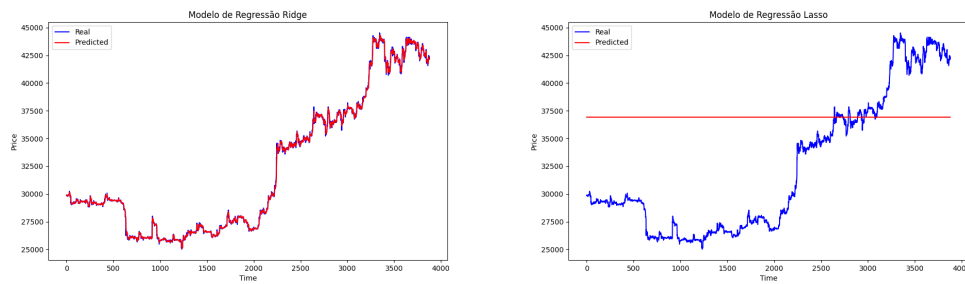


Fig. 4. Previsões da Regressão Ridge (esquerda) e Lasso (direita) para a Janela de 350 horas

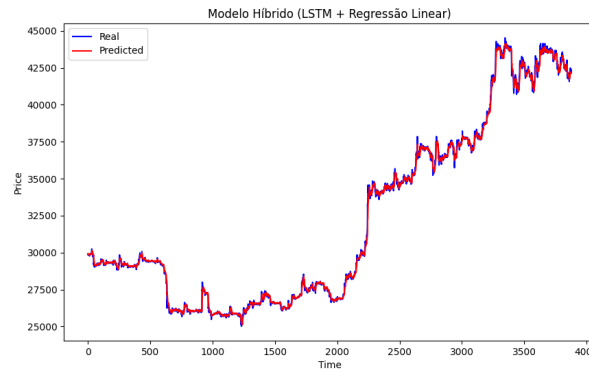


Fig. 5. Previsão do Modelo Híbrido (LSTM + Regressão Linear) para a Janela de 350 horas

5.2 Resultados para Janela de 600 Horas (Saída Única)

Para avaliar o impacto de uma janela de entrada maior, também realizamos experimentos com uma janela de 600 horas, mantendo a previsão de uma única saída. A Tabela 4 resume os resultados obtidos nessa configuração.

Table 2. Métricas de Desempenho para Janela de 600 horas (Saída Única)

| Modelo | MSE | MAE | R ² | MAPE | RMSE |
|-------------------------|-------------|---------|----------------|--------|---------|
| Regressão Linear | 18940.11 | 85.37 | 0.9995 | 0.0026 | 137.62 |
| Lasso Regression | 58361000.96 | 6724.21 | -0.5529 | 0.2305 | 7639.44 |
| Ridge Regression | 23477.09 | 91.34 | 0.9994 | 0.0027 | 153.22 |
| LSTM | 77569.22 | 194.41 | 0.9979 | 0.0059 | 278.51 |
| Modelo Híbrido | 33145.78 | 118.92 | 0.9991 | 0.0036 | 182.06 |

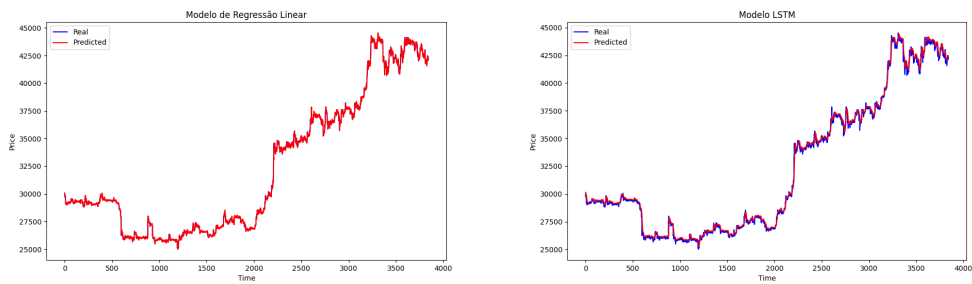


Fig. 6. Previsões da Regressão Linear (esquerda) e LSTM (direita) para a Janela de 600 horas

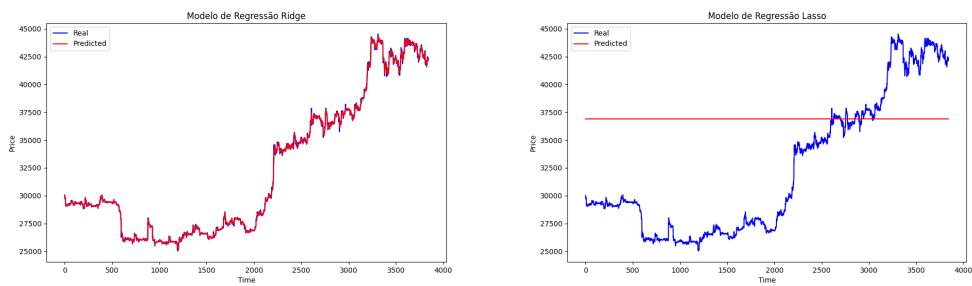


Fig. 7. Previsões da Regressão Ridge (esquerda) e Lasso (direita) para a Janela de 600 horas

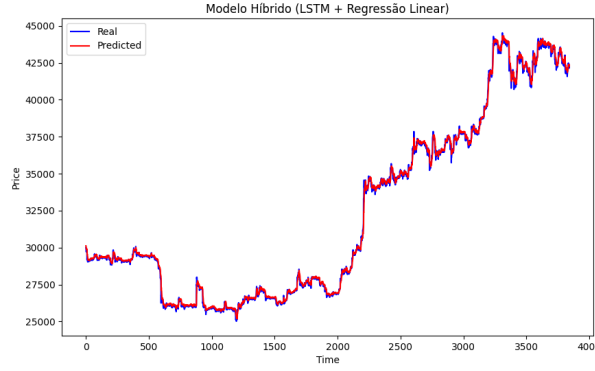


Fig. 8. Previsão do Modelo Híbrido (LSTM + Regressão Linear) para a Janela de 600 horas

5.3 Resultados para múltiplas saídas (Window = 350)

As Tabelas apresenta as métricas para a janela de 350 horas e múltiplas saídas (10 passos à frente). Os gráficos das previsões dos modelos estão ilustrados nas Figuras 9 e 10.

Table 3. Métricas para Step 1 - Comparação entre modelos com janela de 350 horas.

| Modelo | MSE | MAE | R^2 | MAPE | RMSE |
|-------------------------|-----------|--------|--------|--------|--------|
| Regressão Linear | 18474.95 | 83.39 | 0.9995 | 0.0025 | 135.92 |
| LSTM | 127752.96 | 247.07 | 0.9966 | 0.0069 | 357.43 |
| Modelo Híbrido | 45088.66 | 141.66 | 0.9988 | 0.0040 | 212.34 |

Table 4. Métricas para Step 2 - Comparação entre modelos com janela de 350 horas.

| Modelo | MSE | MAE | R^2 | MAPE | RMSE |
|-------------------------|-----------|--------|--------|--------|--------|
| Regressão Linear | 34739.02 | 115.18 | 0.9991 | 0.0035 | 186.38 |
| LSTM | 135766.22 | 252.19 | 0.9963 | 0.0072 | 368.46 |
| Modelo Híbrido | 59340.52 | 158.82 | 0.9984 | 0.0046 | 243.60 |

Table 5. Métricas para Step 3 - Comparação entre modelos com janela de 350 horas.

| Modelo | MSE | MAE | R^2 | MAPE | RMSE |
|-------------------------|-----------|--------|--------|--------|--------|
| Regressão Linear | 52691.41 | 142.97 | 0.9986 | 0.0043 | 229.55 |
| LSTM | 153321.86 | 265.92 | 0.9959 | 0.0076 | 391.56 |
| Modelo Híbrido | 76742.18 | 177.75 | 0.9979 | 0.0052 | 277.02 |

Table 6. Métricas para Step 4 - Comparação entre modelos com janela de 350 horas.

| Modelo | MSE | MAE | R^2 | MAPE | RMSE |
|-------------------------|-----------|--------|--------|--------|--------|
| Regressão Linear | 71197.00 | 165.99 | 0.9981 | 0.0050 | 266.83 |
| LSTM | 166704.63 | 279.93 | 0.9955 | 0.0081 | 408.29 |
| Modelo Híbrido | 93311.58 | 196.22 | 0.9975 | 0.0058 | 305.47 |

Table 7. Métricas para Step 5 - Comparação entre modelos com janela de 350 horas.

| Modelo | MSE | MAE | R^2 | MAPE | RMSE |
|-------------------------|-----------|--------|--------|--------|--------|
| Regressão Linear | 89377.67 | 184.63 | 0.9976 | 0.0055 | 298.96 |
| LSTM | 180882.94 | 284.66 | 0.9951 | 0.0082 | 425.30 |
| Modelo Híbrido | 109987.13 | 208.92 | 0.9970 | 0.0062 | 331.64 |

Table 8. Métricas para Step 6 - Comparação entre modelos com janela de 350 horas.

| Modelo | MSE | MAE | R^2 | MAPE | RMSE |
|-------------------------|-----------|--------|--------|--------|--------|
| Regressão Linear | 107416.34 | 202.06 | 0.9971 | 0.0061 | 327.74 |
| LSTM | 203897.62 | 324.12 | 0.9945 | 0.0096 | 451.55 |
| Modelo Híbrido | 128523.90 | 233.36 | 0.9965 | 0.0070 | 358.50 |

Table 9. Métricas para Step 7 - Comparação entre modelos com janela de 350 horas.

| Modelo | MSE | MAE | R^2 | MAPE | RMSE |
|-------------------------|-----------|--------|--------|--------|--------|
| Regressão Linear | 126122.07 | 218.73 | 0.9966 | 0.0066 | 355.14 |
| LSTM | 215630.05 | 306.98 | 0.9942 | 0.0089 | 464.36 |
| Modelo Híbrido | 145254.89 | 238.65 | 0.9961 | 0.0071 | 381.12 |

Table 10. Métricas para Step 8 - Comparação entre modelos com janela de 350 horas.

| Modelo | MSE | MAE | R^2 | MAPE | RMSE |
|-------------------------|-----------|--------|--------|--------|--------|
| Regressão Linear | 143440.37 | 232.63 | 0.9961 | 0.0070 | 378.74 |
| LSTM | 240186.25 | 326.97 | 0.9936 | 0.0095 | 490.09 |
| Modelo Híbrido | 163808.01 | 254.63 | 0.9956 | 0.0075 | 404.73 |

Table 11. Métricas para Step 9 - Comparação entre modelos com janela de 350 horas.

| Modelo | MSE | MAE | R^2 | MAPE | RMSE |
|-------------------------|-----------|--------|--------|--------|--------|
| Regressão Linear | 161882.45 | 246.89 | 0.9957 | 0.0074 | 402.35 |
| LSTM | 250687.01 | 327.84 | 0.9933 | 0.0095 | 500.69 |
| Modelo Híbrido | 180463.37 | 264.28 | 0.9952 | 0.0078 | 424.81 |

Table 12. Métricas para Step 10 - Comparação entre modelos com janela de 350 horas.

| Modelo | MSE | MAE | R ² | MAPE | RMSE |
|-------------------------|-----------|--------|----------------|--------|--------|
| Regressão Linear | 180807.98 | 261.88 | 0.9952 | 0.0079 | 425.22 |
| LSTM | 269964.63 | 341.18 | 0.9928 | 0.0099 | 519.58 |
| Modelo Híbrido | 198981.59 | 278.46 | 0.9947 | 0.0083 | 446.07 |

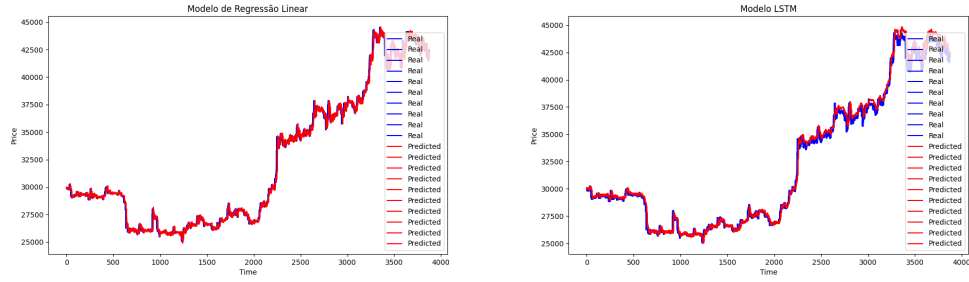


Fig. 9. Previsões de Regressão Linear (esquerda) e Modelo LSTM (direita) Janela 350 horas e múltiplas saídas

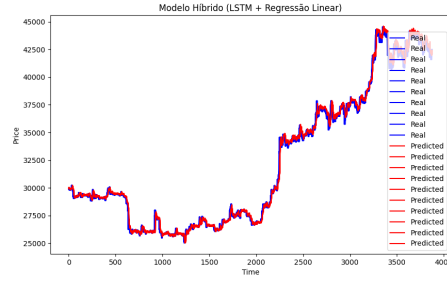


Fig. 10. Modelo Híbrido (LSTM + Regressão Linear) - Janela 350 horas e múltiplas saídas.

5.4 Resultados com Várias Features de Entrada e Saída Única

Realizamos experimentos com múltiplas features de entrada, utilizando uma janela de 350 horas e uma saída única. A Tabela 13 apresenta os resultados obtidos para essa configuração.

Os gráficos a seguir mostram as previsões feitas pelos modelos para essa configuração.

Table 13. Comparação das métricas de desempenho para janela de 350 horas com entrada de 6 variáveis e (Saída Única).

| Modelo | MSE | MAE | R^2 | MAPE | RMSE |
|-------------------------|-------------|---------|---------|--------|---------|
| LSTM | 74921.56 | 197.02 | 0.9979 | 0.0058 | 273.72 |
| Regressão Linear | 21707.80 | 98.80 | 0.9994 | 0.0030 | 147.34 |
| Regressão Lasso | 58526751.08 | 6739.49 | -0.5700 | 0.2311 | 7650.28 |
| Regressão Ridge | 36141.86 | 140.84 | 0.9990 | 0.0045 | 190.11 |
| Modelo Híbrido | 32587.52 | 124.28 | 0.9991 | 0.0037 | 180.52 |

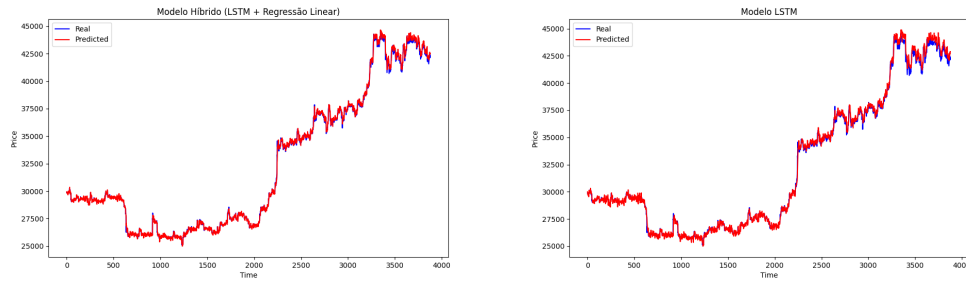


Fig. 11. Modelo Híbrido (LSTM + Regressão Linear (esquerda) e Modelo LSTM (direita) Janela 350 horas e (Saída Única)

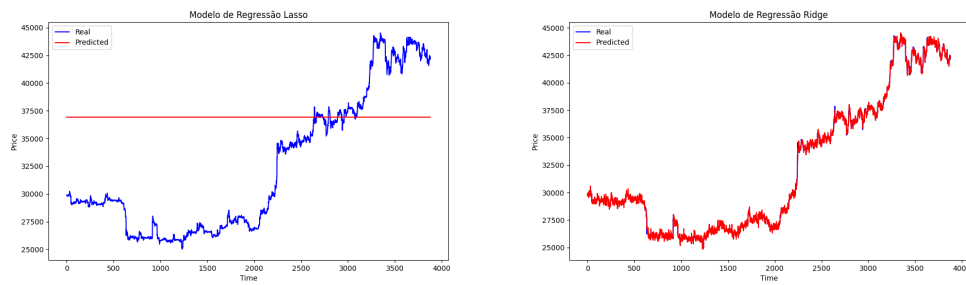


Fig. 12. Modelo Regressão Lasso (esquerda) e Regressão Ridge (direita) Janela 350 horas e (Saída Única)

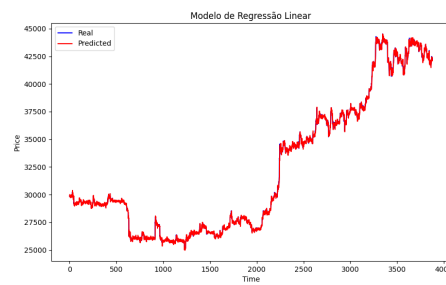


Fig. 13. Gráfico de Previsão - Modelo de Regressão Linear

5.5 Resultados com a Nova Base (Bitcoin 2024)

Por fim, realizamos testes com a nova base de dados de Bitcoin para o ano de 2024. A Tabela 14 mostra os resultados obtidos pelos modelos testados.

Table 14. Métricas para os modelos com nova base (Bitcoin 2024).

| Modelo | MSE | MAE | R ² | MAPE | RMSE |
|-------------------------|------------|---------|----------------|--------|---------|
| Regressão Linear | 125670.49 | 233.07 | 0.9981 | 0.0038 | 354.50 |
| LSTM | 2296802.20 | 1288.68 | 0.9660 | 0.0203 | 1515.52 |
| Modelo Híbrido | 668621.56 | 679.42 | 0.9901 | 0.0107 | 817.69 |

Os gráficos a seguir ilustram as previsões dos modelos com a nova base de dados.

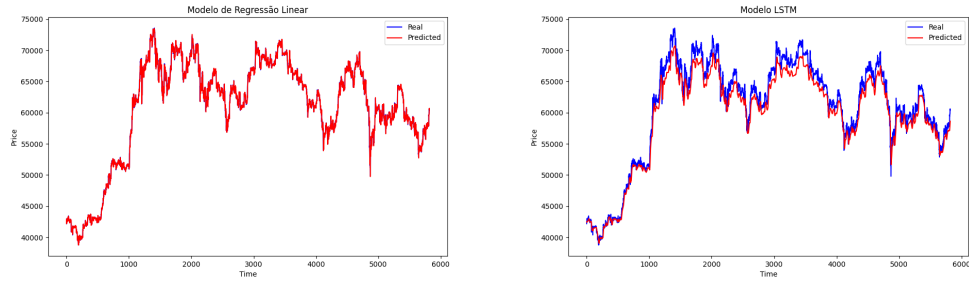


Fig. 14. Modelo de Regressão Linear (esquerda) e Modelo LSTM (direita) Janela 350 horas e (Saída Única) - Nova Base (Bitcoin 2024)

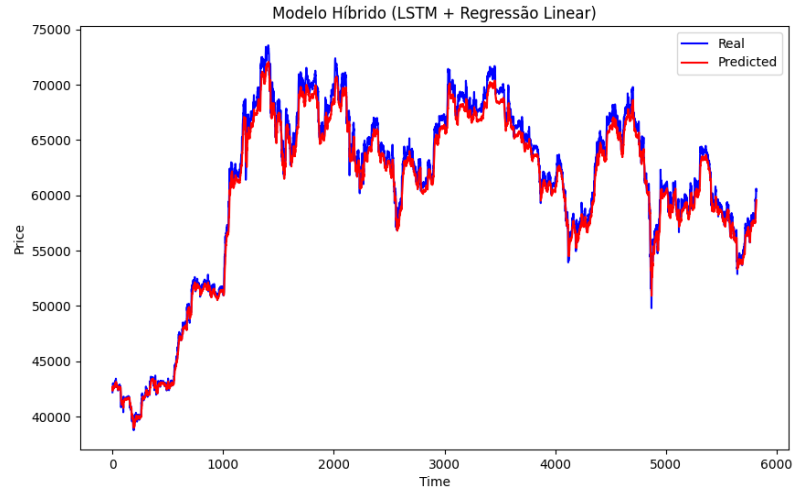


Fig. 15. Modelo Híbrido - Nova Base (Bitcoin 2024)

6 DISCUSSÃO DOS RESULTADOS

Neste estudo, buscamos comparar o desempenho de diversos modelos de previsão de séries temporais, incluindo Regressão Linear, Lasso, Ridge, LSTM, e um modelo híbrido (LSTM + Regressão Linear). Os dados utilizados foram históricos do preço de fechamento do Bitcoin entre 2021 e 2024, com uma janela de 350 horas para a previsão.

6.1 Análise Comparativa

Os resultados indicam uma clara diferença de desempenho entre os modelos testados. Como era esperado, a **Regressão Linear** apresentou boa performance em dados de natureza mais linear, obtendo um MSE de 125670.49, com uma baixa MAE (233.07) e um R^2 de 0.9981, confirmando sua capacidade de lidar com padrões mais simples. A abordagem de regressão regularizada, no entanto, apresentou resultados significativamente piores, como evidenciado pela Regressão Lasso, com um MSE de 58526751.08 e um R^2 negativo (-0.5700), sugerindo que o modelo não conseguiu captar a complexidade dos dados.

Por outro lado, o modelo **LSTM**, projetado para capturar padrões mais complexos e não lineares ao longo do tempo, mostrou-se eficaz, porém com alguns desafios. Embora tenha um desempenho inferior à Regressão Linear nos dados de 2024, com um MSE de 2296802.20 e um R^2 de 0.9660, ele ainda se destaca pela capacidade de identificar padrões de longo prazo, algo que é evidente na sua estrutura de previsão. Essa alta capacidade de previsão, no entanto, foi comprometida pelo alto erro absoluto médio (MAE de 1288.68) e pela dispersão dos resultados, refletida no RMSE de 1515.52.

6.2 Modelo Híbrido

A hibridização dos modelos de **Regressão Linear e LSTM** provou ser uma abordagem promissora. Com um MSE de 668621.56 e um R^2 de 0.9901, o modelo híbrido conseguiu combinar a capacidade do LSTM de capturar as dependências temporais de longo prazo com a simplicidade e eficácia da Regressão Linear para padrões mais lineares. Este modelo apresentou o melhor equilíbrio entre complexidade e precisão, com uma redução substancial nos erros absolutos (MAE de 679.42) em comparação ao LSTM puro, mantendo uma menor dispersão (RMSE de 817.69).

A análise gráfica reforça os achados quantitativos. Observa-se que o modelo híbrido captura com maior precisão os picos e quedas repentinas do preço do Bitcoin, quando comparado ao LSTM puro e à Regressão Linear. Este comportamento sugere que a combinação de abordagens lineares e não lineares fornece uma visão mais robusta do comportamento dos dados, especialmente em cenários com alta volatilidade.

7 CONCLUSÃO

Com base nos experimentos realizados, concluímos que a abordagem híbrida (LSTM + Regressão Linear) oferece uma solução eficiente para a previsão de séries temporais no mercado financeiro, especificamente para dados voláteis como os preços do Bitcoin. Embora a Regressão Linear continue a ser uma escolha eficiente para dados com padrões mais simples, sua capacidade é limitada em cenários mais complexos. O modelo LSTM, apesar de ser eficaz na captura de padrões temporais não lineares, sofre com a elevada variabilidade nos erros.

A hibridização das duas abordagens, como demonstrado pelos resultados, oferece o melhor dos dois mundos: a precisão da Regressão Linear em padrões lineares, combinada com a capacidade do LSTM de capturar complexidades temporais. Futuros estudos podem explorar a incorporação de dados textuais, como análise de sentimento de notícias financeiras, para complementar as previsões quantitativas dos modelos híbridos.

REFERENCES

- [1] Anna Sheila Crisostomo, Carl Christian Chaguile, and Reggie Gustilo. Stock market prediction using linear regression. In *2023 IEEE 15th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, pages 1–4, 2023.
- [2] Pushpendu Ghosh, Ariel Neufeld, and Jajati Keshari Sahoo. Forecasting directional movements of stock prices for intraday trading using lstm and random forests, 2021.
- [3] Rezaul Karim, Md Khorshed Alam, and Md Rezaul Hossain. Stock market analysis using linear regression and decision tree regression. In *2021 1st International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, pages 1–6, 2021.
- [4] Xiangyu Li, Xinjie Shen, Yawen Zeng, Xiaofen Xing, and Jin Xu. Finreport: Explainable stock earnings forecasting via news factor analyzing model, 2024.
- [5] S.S. Panigrahi and J.K. Mantri. A text based decision tree model for stock market forecasting. In *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*, pages 405–411, 2015.
- [6] Govind Patidar, Animesh Kumbhakar, Harshit Mahabale, and Juned A Siddiqui. Stock market analysis using long short-term memory(lstm) and linear regression machine learning model. In *2023 IEEE 7th Conference on Information and Communication Technology (CICT)*, pages 1–6, 2023.
- [7] Sreelekshmy Selvin, R Vinayakumar, E. A Gopalakrishnan, Vijay Krishna Menon, and K. P. Soman. Stock price prediction using lstm, rnn and cnn-sliding window model. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1643–1647, 2017.