Uma Comparação De Métodos Estatísticos E De Aprendizado De Máquina Na Predição Da Cotação Do Bitcoin

A Comparison Of Statistical And Machine Learning Methods In Predicting Bitcoin Quotation

Ricardo Junior Fioravante*, Gustavo Henrique Paetzold†

RESUMO

Neste trabalho busca-se comparar sistemas computacionais capazes de predizer valores futuros do Bitcoin por meio de algoritmos estatísticos e de aprendizado de máquina com exatidão e confiabilidade. Para o desenvolvimento do projeto serão utilizados algoritmos como: auto-regressivo (em inglês, AR), médias móveis (em inglês, MA), auto-regressivo integrado de médias móveis (em inglês, ARIMA), suavização exponencial simples (em inglês, SES), Prophet e rede de memória a longo prazo (em inglês, LSTM), para efeito de comparação. A predição se dará pelo preço de fechamento, diário e de uma hora, em dólar. O banco de dados para essas variáveis serão adquiridos pelo site Kaggle. Para treinamento serão utilizados 80% dos dados e para teste 20% e por último serão feitos, uma análise de desempenho e um sistema de compra e venda para comparação dos algoritmos. Assim, espera-se beneficiar investidores ao diminuir a incerteza na hora de tomar decisões.

Palavras-chave: Criptomoeda. Redes Neurais Recorrentes. Regressão. Séries Temporais.

ABSTRACT

This work aims to compare computer systems capable of predicting the future value of Bitcoin through statistical and machine learning algorithms with accuracy and reliability. For the development of the project will be used algorithms such as: autoregressive (AR), moving averages (MA), integrated autoregressive moving averages (ARIMA), simple exponential smoothing (in English, SES), Prophet and long short-term memory neural networks (LSTM) for comparison. The prediction is displayed by the closing price, daily and one hour, in dollars. The database for these variables is acquired from the Kaggle website. For training, 80 % of the data will be used and for testing 20 %, and finally, a performance analysis and a purchase and sale system will be used to compare the algorithms. We hope our models help investors to reduce uncertainty when making decisions.

Keywords: Cryptocurrency. Recurring Neural Networks. Regression. Time Series.

1 INTRODUÇÃO

A Bitcoin é um sistema de pagamento eletrônico *Peer-to-Peer* (Ponto a ponto), proposto em 2008, pelo pseudônimo de Satoshi Nakamoto, sendo que a identidade real de quem concebeu é desconhecida. Seu sistema é baseado em criptografia, permitindo que duas partes façam uma transação sem a necessidade de uma terceira parte para a validação sendo capaz de resolver o problema do gasto duplo, em que um usuário consegue gastar as mesmas moedas digitais mais de uma vez (NAKAMOTO, 2008).

* Engenharia Da Computação; ricardofioravante@alunos.utfpr.edu.br.
† Engenharia Da Computação; ghpaetzold@utfpr.edu.br.

O desenvolvimento de sistemas automáticos é desafiante tanto para teoria econômica quanto para a computacional, uma vez que oferece a oportunidade da descoberta de novas aplicações, metodologias, análises e utilidades. Assim com a resolução do problema, implementando um modelo preditivo, espera-se beneficiar investidores ao diminuir a incerteza na hora de tomar decisões.

Este trabalho tem o objetivo de comparar a eficácia de diversos métodos estatísticos e de aprendizado de máquina na predição do valor da criptomoeda Bitcoin. Para isso alguns algoritmos como: AR, MA, ARIMA, SES, Prophet e LSTM foram desenvolvidos e comparados na tentativa de abordar tal problema. Como variável de entrada somente o preço de fechamento foi usado, pois esse é o interesse da maioria dos investidores, tanto para os modelos com periodicidade de 1 dia como de 1 hora. Já para o treino foram separados dados do período de 08/10/2015 até 15/05/2019 (80%) e para testes do período de 16/05/2019 até 09/04/2020 (20%). No final para mediar a usabilidade do modelo foram também implementados sistemas que usam os algoritmos testados para realizar a compra e venda do Bitcoin na prática, além das previsões desses modelos serem comparados com métricas.

2 MÉTODO (OU PROCEDIMENTOS OPERACIONAIS DA PESQUISA)

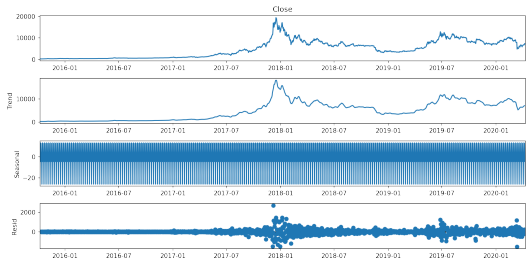
A partir de um banco de dados, das cotações diárias e de hora em hora, do Bitcoin, modelos serão treinados seguindo sua própria parametrização e posteriormente utilizados para fazerem previsões e serem analisados intrinsecamente e extrinsecamente. Os resultados serão então avaliados e comparados. A seguir serão descritos com detalhes cada uma das etapas.

2.1 COLETA DE DADOS E EXPLORAÇÃO DOS DADOS

A coleta dos dados foi feita através do site Kaggle, para a cotação do Bitcoin no período de 08/10/2015 até 09/04/2020, tanto para o período diário (1D) quanto para o de hora em hora (1H). As variáveis encontradas nele correspondem a abertura, máxima, mínima e fechamento do preço no período determinado.

A exploração dos dados consistiu na decomposição e autocorrelação da série temporal que ajudou a extrair algumas informações úteis na construção dos modelos. As Figuras 1a e 1b, mostram o que foi observado.

(a) Decomposição.



(b) Autocorrelação.

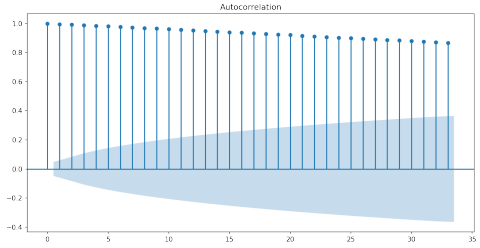


Figura 1 – Exploração dos dados.

Na Figura 1a de decomposição foram observados 3 componentes: No componente de tendencia podemos observar 3 momentos principais, um de alta que vai desde 2016 até inicio de 2018, um de baixa que vai de 2018

até inicio de 2019, e um de alta novamente que vai de 2019 até o inicio de 2020, no componente de sazonalidade, observa-se que não há sazonalidade eminente e no componente de ruido observou-se um ruido estável no inicio que vai aumentando até o fim da série, além de 2 aglomerações maiores em 2018-01 e 2019-07. Já na Figura 1b de autocorrelação, observa-se que há uma dependência até o *lag* 35, indicando que há necessidade de tornar a série estacionaria (KOTU; DESHPANDE, 2019).

2.2 PREPARAÇÃO DOS DADOS

Na preparação, os dados de fechamento (*Close*) foram separados em 80% para treino, de 08/10/2015 até 15/05/2019, e 20% para teste, de 16/05/2019 até 09/04/2020, avaliando assim os modelos com dados desconhecidos. Essa separação foi tanto para o diário, quanto para o de hora em hora. Somente o fechamento foi usado como entrada para os modelos, já que esse interessa mais aos investidores e negociadores.

Já para lidar com a necessidade de tornar a série temporal estacionaria, visto que os modelos ARIMA, MA, AR e Prophet possuem essa necessidade, ao contrario do SES em que não há necessidade de preparação de dados, foi utilizado o teste de *Augmented Dicky Fuller* (ADF), que tenta descobrir o quão forte a série é determinada pela tendencia. Para a hipótese nula (H0) do teste, temos que a série não é estacionaria, e para a hipótese alternativa (H1), temos que a série é estacionaria. Uma interpretação do valor-P (em inglês, *P-value*) é tal que se valor-P > 0.05, aceita-se a hipótese nula e no caso contrario rejeita-se.

O teste foi aplicado primeiro para os dados diários, sem nenhuma diferenciação com um $valor - P = 0.3832$, aceitando H0 e concluindo assim que a série não é estacionaria e portando precisa ser diferenciada. Para o segundo teste, foi feita 1 diferenciação e obteve-se um $valor - P = 0.0000001281$, recusando H0 e concluindo assim que a série é estacionaria e portando não precisa ser mais diferenciada. Para os dados de hora em hora ocorreu de forma análoga, com a diferença de $valor - P = 0.59$ para o primeiro teste e $valor - P = 0$ para o segundo.

Por último para facilitar o treinamento do LSTM, é importante que os dados estejam entre 0 e 1, e possuam a mesma escala. Para isso foi usado o método *MinMax* de normalização, dado pela equação a seguir:

$$z = \frac{x - min(x)}{max(x) - min(x)}, \tag{1}$$

onde z representa o valor normalizado, x representa o valor de entrada da serie e, $min(x)$ e $max(x)$ representam os valores de mínimo e máximo da série.

2.3 CRIAÇÃO DE MODELOS

A criação de modelos, corresponde ao treino e parametrização dos modelos. Para os modelos AR, MA e ARIMA, seus parâmetros p e q foram combinados variando de 0 a 5 e através do estimador AIC, a melhor combinação foi escolhida. Também para o SES o estimador escolhido foi AIC e seu parâmetro α foi analisado variando de 0.05 até 0.95, com um passo de 0.05 por interação. Já o PROPHET foi utilizado de maneira a buscar os parâmetros automaticamente pela sua biblioteca. O modelo *Naive* também foi proposto, se trata de uma aproximação simples de previsão, considerando que a previsão ira repetir o último valor, foi usado como referência na comparação de métricas.

Para o modelo de aprendizado de máquina LSTM, foram utilizados alguns parâmetros como: número de

unidades e tamanho do lote de entrada, variando seus valores entre 5, 25, 50 e 100, e número de ciclos com seus valores variando em 50 até 400. O estimador para definir a melhor combinação dos valores foi o MSE. O modelo também usou o otimizador Adam, e a função de ativação Relu. Por fim todos os modelos fizeram a previsão de 1 período a frente para todos os valores de testes e foram treinados novamente com os parâmetros definidos na fase de treino (YU et al., 2019).

2.4 ANÁLISE INTRÍNSECA E ANÁLISE EXTRÍNSECA

Após a previsão dos valores, temos a análise intrínseca, que corresponde a métricas para previsão. Nessa etapa foi aplicadas métricas para todas as previsões geradas pelos modelos, incluindo a *Naive*, para ambas as periodicidades como: MAPE, RMSE e R^2.

A análise extrínseca, que corresponde a criação de sistemas, funcionou como políticas de compra e venda da seguinte maneira: O primeiro sistema, chamado de Sistema Direcional, foi comparando o preço da previsão com o preço atual, caso seja maior é feito a compra e caso, seja menor é feito a venda;O segundo sistema, chamado de Sistema de Votação, funciona como uma combinação do sistema anterior onde uma combinação de mais de 3 votos de todos os modelos são necessários para compra ou venda; O terceiro sistema, chamado de Sistema de *Buy and Hold*, foi utilizada como referência na comparação por se tratar de algo muito simples, ele funciona comprando no início do período e vendendo no fim dele. Todas as estratégias foram simuladas considerando um saldo inicial de U$ 100 dólares e desconsiderando as taxas de operação.

3 RESULTADOS

3.1 ANÁLISE INTRÍNSECA E ANÁLISE EXTRÍNSECA

Os resultados, referentes a análise intrínseca, considerando as métricas para cada modelo e sua periodicidade, podem ser vistos nas Tabelas 1 e 2. Já para os melhores parâmetros encontrados em cada modelo para cada período temos as Tabela 3 e 4, onde a coluna melhores parâmetros representa um JSON com os melhores parâmetros encontrados.

Tabela 1 – Métricas Para Previsão - Period. 1D

Modelo	MAPE	RMSE	R²
SES	2.9498	397.6040	0.9311
NAIVE	2.9779	400.8768	0.9299
PROPHET	3.0444	405.0639	0.9284
MA	3.0547	408.3656	0.9273
AR	3.0618	409.0631	0.9270
ARIMA	3.1188	411.5988	0.9261
LSTM	3.6327	445.1507	0.9136

Tabela 2 – Métricas Para Previsão - Period. 1H

Modelo	MAPE	RMSE	R²
SES	0.5044	82.8227	0.9970
MA	0.5046	82.8333	0.9970
AR	0.5047	82.8553	0.9970
ARIMA	0.5051	82.8410	0.9970
NAIVE	0.5058	83.0016	0.9970
PROPHET	0.5062	83.0408	0.9970
LSTM	0.6333	89.3332	0.9965

Tabela 3 – Melhores Parâmetros Encontrados - Period. 1D

Modelo	Melhores parâmetros
AR	{'p': 5, 'd': 1, 'q': 0}
ARIMA	{'p': 4, 'd': 1, 'q': 5}
LSTM	{'lote': 5, 'u': 100, 'c': 400}
MA	{'p': 0, 'd': 1, 'q': 5}
NAIVE	{}
PROPHET	{'d': 1}
SES	{'alfa': 0.9}

Tabela 4 – Melhores Parâmetros Encontrados - Period. 1H

Modelo	Melhores parâmetros
AR	{'p': 4, 'd': 1, 'q': 0}
ARIMA	{'p': 4, 'd': 1, 'q': 3}
LSTM	{'lote': 100, 'u': 100, 'c': 200}
MA	{'p': 0, 'd': 1, 'q': 2}
NAIVE	{}
PROPHET	{'d': 1}
SES	{'alfa': 0.9}

Os resultados, referentes a análise extrínseca, considerando cada sistema utilizado podem ser vistos nas Tabelas 5, 6 e 7. Onde cada tabela nos mostra, como na Tabela 5 do Sistema Direcional, o desempenho de cada modelo para cada periodicidade, com seus ganhos, percas, quantidade de vezes que o modelo acertou e errou a direção prevista, porcentagem de acertos da direção prevista e balanço do saldo final (ganho - perda); Na Tabela 6 temos o segundo sistema, que consiste no sistema de votação e na Tabela 7 temos o sistema de *Buy and hold*, ambas tabelas possuem os mesmo campos da Tabela 5 com exceção do campo "Modelo".

Tabela 5 – Sistema Direcional

Modelo	Period.	Ganho	Perda	Acertos	Erros	% de acertos	Balanço
MA	1D	437.213116	-516.607889	160	169	48.63	-79.40
PROPHET	1D	463.200050	-512.388414	151	178	45.90	-49.19
AR	1D	465.067602	-492.265991	155	174	47.11	-27.19
ARIMA	1D	509.596125	-468.965202	160	169	48.63	40.63
AR	1H	2014.709900	-1971.127054	4239	3654	53.71	43.58
LSTM	1D	526.951314	-454.084891	164	165	49.85	72.87
ARIMA	1H	2033.094806	-1954.186886	4210	3683	53.34	78.90
PROPHET	1H	2038.105690	-1952.445793	4019	3874	50.92	85.66
LSTM	1H	2049.188178	-1945.988956	3862	4031	48.93	103.19
SES	1D	570.557597	-413.193854	184	145	55.93	157.36
MA	1H	2096.142757	-1894.628369	4271	3622	54.11	201.52
SES	1H	2113.400686	-1877.736219	4327	3566	54.82	235.66

Tabela 6 – Sistema de Votação

Period.	Ganho	Perda	Acertos	Erros	% de acertos	Balanço
1H	2007.756624	-1979.933676	3997	3896	50.64	27.82
1D	494.292426	-464.836645	165	164	50.15	29.46

Tabela 7 – Sistema de *Buy and hold*

Period.	Ganho	Perda	Acertos	Erros	% de acertos	Balanço
1D	423.238422	-361.356650	144	119	54.75	61.88
1H	1851.516780	-1773.085617	3737	3386	52.46	78.43

4 CONCLUSÕES

Podemos notar na Tabela 1 que todas as métricas para previsão dos modelos com periodicidade diária se ajustaram pior que o modelo *Naive*, com os menores valores de MAPE e RMSE, e com os valores do R^2 mais próximos de 1, com exceção do modelo SES que conseguiu ter todas as métricas superiores não só a ele, mas a todos os outros. Já para a Tabela 2, temos quase que o inverso, com quase todos os modelos sendo superiores ao *Naive*, com exceção do Prophet e do LSTM, indicando assim que as previsões dos modelos se ajustaram muito melhor ao período de hora em hora do que o diário. Nota-se também que em ambos os casos o SES foi o que obteve as melhores métricas em comparação aos outros modelos.

Nas tabelas 3 e 4, devem ser feitas algumas considerações para os modelos *Naive* e Prophet, onde para o primeiro não há parâmetros para serem estimados por isso o JSON vem vazio e para o segundo, o parâmetro "ordem_d"na verdade, representa o tratamento da série temporal que precisou ser diferenciada para se ajustar melhor ao modelo, e não um parâmetro do modelo como acontece com o ARIMA, MA e AR.

Para o Sistema Direcional, como visto na Tabela 5, nota-se que, o modelo SES tanto para a periodicidade diária quanto para a de hora em hora apresentou um balanço muito superior ao saldo inicial de U$ 100 dólares e apenas 3 modelos com periodicidade diária tiveram balanço negativo, MA, AR e PROPHET. Já para o Sistema de Votação, como visto na Tabela 6, observa-se que o balanço para ambas periodicidades tiveram um resultado positivo mas pouco expressivo, assim como a % de acertos que se manteve próximo a 50%. Por último no Sistema de *Buy and hold*, como visto na Tabela 7, vemos um resultado muito parecido para ambas as periodicidades.

Dentre os 3 sistemas de compra e venda, podemos observar que há 7 casos em que o balanço não conseguiu superar o Sistema de *Buy and hold*, são eles 5 casos do primeiro sistema ARIMA(1D), MA(1D), PROPHET(1D) e AR(1D e 1H), e ambos os casos do Sistema de Votação. No geral os sistemas obtiveram um resultado melhor para a periodicidade de hora em hora, e um caso do Sistema Direcional se destaca entre eles, o modelo SES.

REFERÊNCIAS

KOTU, Vijay; DESHPANDE, Bala. Time Series Forecasting. In: DATA Science. [S.l.]: Elsevier, 2019. P. 395–445. DOI: 10.1016/b978-0-12-814761-0.00012-5. Disponível em: .

NAKAMOTO, Satoshi. Bitcoin: A peer-to-peer electronic cash system, 2008.

YU, Yong et al. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. **Neural Computation**, v. 31, n. 7, p. 1235–1270, 2019. DOI: 10.1162/neco_a_01199.