

Aplicação de Algoritmos de *Machine Learning* na previsão de cotação do *Bitcoin*

Rodrigo de Souza Oliveira¹, Álvaro Viebrantz¹

¹ MBA em Big Data – Fatec – Serviço Nacional de Aprendizagem Industrial (SENAI)
Av. XV de Novembro, 303, Porto - CEP: 78020-300 - Cuiabá-MT

rddsouzaoliveira@gmail.com, emaildoalvaro@gmail.com

Abstract. *Resumo em inglês aqui.*

Resumo. *Com a evolução da tecnologia e a criação de diversas ferramentas de análise de dados, um assunto que vem se tornando muito relevante é a aplicação dessas técnicas avançadas de Machine Learning para entender o comportamento dos ativos e subsidiar o mercado financeiro de informações para tomada de decisão. Este trabalho teve como objetivo aplicar uma estrutura de rede neural artificial LSTM, com a função de analisar e entender o comportamento do bitcoin para projetar seu valor futuro e utilizá-lo como ferramenta para auxiliar nas estratégias de investimentos. Para isso, foram coletados dados sobre a cotação do bitcoin, através da API de uma corretora, em uma série histórica criada desde dezembro de 2017. Os resultados evidenciam que o modelo foi capaz de entender e projetar as oscilações do preço do bitcoin, porém superestimava o seu valor. Este trabalho abriu caminho para uma gama de possibilidades de futuros trabalhos para melhor entender as arquiteturas de modelos disponíveis que consigam generalizar o comportamento do bitcoin ao longo do tempo.*

1. Introdução

Investimento é a aplicação de algum recurso, com a expectativa de algum ganho futuro. A incerteza do ganho, ou um possível prejuízo, caracteriza o risco do investimento. No contexto de aplicação financeira, pode-se caracterizar o investimento como a aplicação de dinheiro que não gere custos, tenha expectativa de lucro futuro e que não exija esforços relevantes¹.

Diversos estudos, já consolidados, trazem que a diversificação do portfólio de investimentos se demonstra como uma eficiente forma para redução dos riscos, conforme trazido em [Oda et al. 1998]. A diversificação de investimentos é uma técnica que visa a diluição dos riscos e a maximização dos ganhos [Pactual 2017].

Aliado à sua popularização nos últimos anos, o Bitcoin se tornou uma alternativa para os investidores comporem suas carteiras de investimentos e despertou-se a necessidade de melhor entender seu comportamento. Porém, a oscilação e a incerteza do futuro do Bitcoin, gera muita insegurança, principalmente entre os investidores mais conservadores [Oliveira 2020].

¹Detalhes em Blog Rico. O Que é Investimento e Por Que Poupança é Ruim. Recuperado em 3 de outubro de 2020, de <https://blog.rico.com.vc/o-que-e-investimento>

Como a tecnologia e as novas ferramentas de análise de dados, podem contribuir para transformar o bitcoin como uma alternativa de investimento, minimizando-se os riscos das operações?

A evolução da tecnologia proporcionou o surgimento de novas ferramentas de análise de dados. Estas ferramentas podem contribuir para transformar o bitcoin como uma alternativa de investimento, pois, através do entendimento de seu comportamento, possibilitam a minimização dos riscos das operações com a moeda.

O Objetivo deste trabalho foi implementar e testar modelos capazes de analisar dados históricos das cotações do Bitcoin, para identificar o comportamento do preço do Bitcoin e projetar as cotações futuras.

Os modelos desenvolvidos tiveram como ideia, possibilitar que o investidor tenha uma ferramenta técnica para auxiliar no processo de tomada de decisões sobre as estratégias que investimentos que serão adotadas.

Justifica-se o desenvolvimento deste trabalho, pela importância que o Bitcoin conquistou no mercado financeiro nos últimos anos, pelo desafio na aplicação de técnicas extremamente avançadas para análise de dados com a finalidade de prever o valor futuro do ativo, que é um dos grandes desafios dos cientistas de dados.

A estrutura deste trabalho está organizada da seguinte forma: A Metodologia (Capítulo 2), onde será abordado os principais passos para coleta de dados, análise exploratória e o desenvolvimento do modelo, A Revisão de Literatura (Capítulo 3), onde serão apresentados os conceitos acerca do Bitcoin, Machine Learning e sobre as Redes Neurais Artificiais, A Apresentação da Pesquisa (Capítulo 4) e Discussão de Resultados (Capítulo 5), onde abordará sobre os resultados obtidos em cada etapa do desenvolvimento do modelo e o resultado da previsão realizada e por último, as Considerações Finais (Capítulo 6), onde se conclui sobre as aplicações do modelo e recomendações de futuros estudos.

2. Metodologia

A finalidade dessa pesquisa tem natureza aplicada e abordagem qualitativa e quantitativa para realizar estudo de caráter descritivo.

Foi realizado um levantamento bibliográfico sobre os temas abordados neste artigo, através da coleta de informações em livros, revistas, sites e outros artigos.

Para a coleta e análise dos dados utilizou-se um Macbook Pro 2018 com memória RAM de 16GB, processador Intel i7 de 2.2Ghz, vídeo Radeon Pro 555x de 4Gb e um Flash Storage de 251Gb, além de sistema operacional MacOS Catalina, com acesso à internet.

Utilizou-se como linguagem de programação o R [R Core Team 2020] para todas as etapas de análises apresentadas neste artigo e o Latex [Goossens et al. 1993] para escrita. Como banco de dados, utilizou-se o Microsoft SQL Server, hospedado na nuvem Azure.

2.1. Coleta dos dados

A coleta dos dados se iniciou em dezembro de 2017, através da API REST da corretora de cyptomoedas Bitfinex. Estruturou-se uma rotina em linguagem R, dentro de uma máquina

virtual no ambiente Google Cloud Plataform (GCP), que acessava a API e armazenava os dados em um banco de dados SQL Server, num período de 10 em 10 segundos.

A rotina, consultava dois grupos de informações, fornecidos pela API: ”*Ticker*” e ”*Orderbook*”. O primeiro grupo, trazia informações sobre a cotação no momento da consulta e trazia os seguintes campos:

- Valor atual de venda
- Valor atual de compra
- Último valor negociado
- Menor valor negociado das últimas 24h
- Maior valor negociado nas últimas 24h
- Volume total negociado nas últimas 24h

O segundo grupo, trazia o ”livro de ordens”, que era a listagem de todas as ordens de compra e venda na corretora, vigentes no momento da consulta. Essa consulta retornava milhares de linhas e para reduzir o espaço de armazenamento, optou-se por resumir as informações antes de salvar no banco de dados. Os dados resumidos traziam as seguintes informações tanto das ordens de compra, quanto das ordens de venda:

- Mínimo
- 1o Quartil
- Mediana
- 3o Quartil
- Máximo

2.2. Análise exploratória

Com os dados já estruturados e armazenados num banco de dados, o próximo passo foi realizar uma análise exploratória no conjunto de dados.

Como o tempo de resposta da API pode variar, analisou-se os intervalos entre cada consulta efetivamente realizada. Após essa análise, foram realizadas algumas tratativas nos dados para se chegar ao *dataset* final.

2.3. Modelos

Neste trabalho foram utilizados os modelos RandomForest e *Long-Short Term Memory* (LSTM), considerando como base de treinamento os dados compreendidos entre novembro de 2017 até dezembro de 2019 e para validação os dados entre janeiro e março de 2020.

O modelo LSTM form treinado utilizando o pacote Keras [Chollet 2015], que cria uma camada para execução do ambiente Tensorflow [Abadi et al. 2015].

3. Revisão de Literatura

Neste tópico serão apresentados os principais conceitos abordados no estudo, contextualizando o Bitcoin, os tipos de aprendizados de máquina e os principais pontos que envolvem as redes neurais artificiais.

3.1. Bitcoin

O Bitcoin consiste em uma rede estruturada ponto-a-ponto, constituída por diversos computadores que assumem o mesmo papel de validadores das transações que são realizadas entre eles [Nakamoto 2019].

Todas as transações que são realizadas pela rede do Bitcoin, são criptografadas, dificultando sua rastreabilidade.

De acordo com [Shawn 2017], o Bitcoin foi a primeira moeda digital a resolver o problema do gasto duplo, utilizando uma solução descentralizada para armazenamento dos registros das transações.

Outra característica da moeda é que não existe nenhum governo ou instituição que regulamenta suas transações e seu preço. Esses fatores ajudaram a contribuir para a popularização do Bitcoin nos últimos anos e consequentemente a sua valorização, saindo de centavos de dólares em 2009 para cerca de 7 mil em 2018, conforme imagem abaixo.



Figura 1. Preço Histórico do Bitcoin

Fonte: <https://www.coindesk.com/price/bitcoin>

Atualmente, estima-se mais de 375 milhões de dólares sejam comercializados em bitcoin em apenas um dia.

3.2. Machine Learning

Segundo [Kelleher et al. 2020], Machine Learning pode ser definido como a automação de processos que extraem padrões dos dados. Compreende um conjunto de técnicas computacionais que têm a característica de "aprender" com os dados.

De acordo com [Tanaka 2018], existem três tipos de algoritmos de *Machine Learning*:

- **Supervisionado:** Quando se diz ao algoritmo o que é cada entrada (rótulo) e ele aprende quais são as características que influenciam a entrada a ser o que ela é.
- **Não supervisionado:** Quando não se diz ao algoritmo o que é cada entrada, ou seja, os dados não são rotulados. O algoritmo classifica as entradas conforme suas características semelhantes.
- **Por reforço:** Define-se um sistema de recompensas e punições aos possíveis resultados para que o algoritmo possa ponderar as escolhas a serem feitas.

3.3. Redes Neurais Artificiais

Redes Neurais Artificiais (RNAs), São modelos matemáticos, compostos por unidades de processamentos simples, que calculam determinadas funções matemáticas. Uma rede neural artificial é um modelo de computação inspirado na forma como a estrutura do cérebro dos mamíferos processa informação [De Souto et al. 2003].

As unidades de processamento, também chamadas de nós ou neurônios, são dispostas em uma ou várias camadas e interligadas por conexões. As conexões estão ligadas a pesos, que ponderam os valores recebidos por cada neurônio.

O conhecimento sobre o problema em consideração está guardado dentro dos exemplos que têm que estar obrigatoriamente disponíveis. O algoritmo de aprendizagem generaliza esses dados e memoriza o conhecimento dentro dos parâmetros adaptáveis da rede, os pesos [Raubert 2005].

A camada que recebe os dados é chamada de camada de entrada, a camada de saída é responsável por traduzir o resultado da RNA e quaisquer outras são denominadas de camadas ocultas, ou intermediárias, vide Figura 2 abaixo. Em cada uma das camadas é aplicada uma função de ativação. Durante a fase de treinamento, a RNA "aprende" ajustando-se os pesos [Bishop 1996].

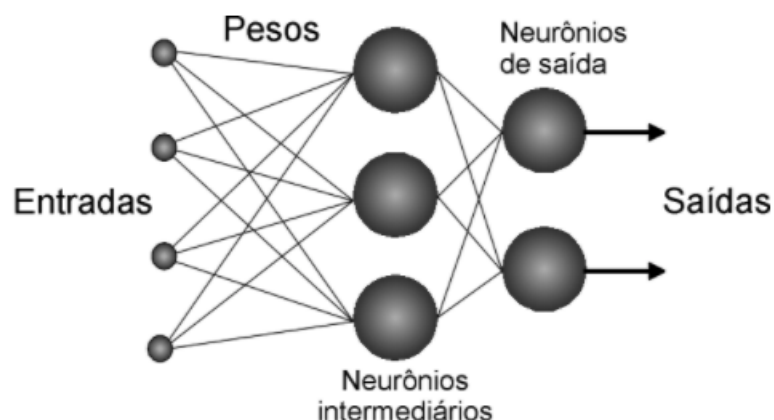


Figura 2. Exemplo de uma Rede Neural Artificial de 2 camadas com 4 entradas e 2 saídas

Fonte: <https://cerebromente.org.br/n05/tecnologia/rna.htm>,
consulta em 02 de outubro de 2020

3.3.1. Redes Neurais Recorrentes e Long Short-Term Memory

A Rede Neural Recorrente (RNN), é uma arquitetura de rede neural especializada em processar dados sequenciais. RNNs também são capazes de lidar com sequências de tamanhos variados. Porém, existe uma limitação nesse tipo de rede, quando se trata em aprender dependências após vários estágios de processamento.

Diversas variações das RNNs foram criadas com o objetivo de amenizar este problema, sendo uma delas as células de Long Short Term Memory (LSTM). Tais células possuem “memória”, ou seja, conseguem guardar informações por longos períodos.

3.4. Random Forest

Colocar random forest ou regressão?

4. Discussão dos resultados

Na tabela 1 é possível verificar algumas estatísticas acerca do conjunto de dados.

Tabela 1. Sumarização dos dados

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
AIL	3278	6384	7797	7917	9442	19891
AQ1	3561	6890	8217	8456	9877	20542
AMD	3706	7838	8889	9357	11202	22125
AQ3	3874	8442	10762	11031	12926	25195
ASL	4219	9371	13501	14550	19359	36100
BIL	0	1783	4111	4511	7300	10475
BQ1	1001	4350	6108	6110	7863	13026
BMD	2822	5448	6979	6917	8528	16156
BQ3	3067	5977	7412	7437	9097	18124
BSL	3278	6384	7797	7916	9441	19890
BAMOUNT	789.5	3785.3	9342.7	141543.1	129697.3	835193.2
AAMOUNT	923.5	4961.4	7571.0	7961.8	10788.1	20640.8
Bid	3278	6384	7797	7916	9441	19890
Ask	3278	6384	7797	7918	9442	800000
Last	3278	6384	7797	7916	9442	19891
Low	0	6252	7502	7644	9183	18734
High	0	6519	8049	8169	9725	19891
Volume	0	7102	13983	23613	31641	213851
datetime	2017-11-22 22	2018-06-25 11	2019-02-11 10	2019-01-31 06	2019-09-06 16	2020-09-29 02

Fonte: Dados do estudo

5. Considerações finais

The subsection titles must be in boldface, 12pt, flush left.

6. Referências

Referências

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

- Bishop, C. M. (1996). Neural networks: a pattern recognition perspective.
- Chollet, F. (2015). keras. <https://github.com/fchollet/keras>.
- De Souto, M., Lorena, A., Delbem, A., and de Carvalho, A. (2003). Técnicas de aprendizado de máquina para problemas de biologia molecular. *Sociedade Brasileira de Computação*, 1(2).
- Goossens, M., Mittelbach, F., and Samarin, A. (1993). *The LaTeX Companion*. Addison-Wesley, Reading, Massachusetts.
- Kelleher, J. D., Mac Namee, B., and D'arcy, A. (2020). *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT press.
- Nakamoto, S. (2019). Bitcoin: A peer-to-peer electronic cash system. Technical report, Manubot.
- Oda, A. L., Senger, M. C. M., and Chára, A. N. (1998). Um estudo sobre diversificação na bolsa de valores de são paulo. *XXII Encontro da Associação Nacional de Pós-Graduação em Administração (ENANAP)*.
- Oliveira, J. J. (2020). *Bitcoin é investimento de risco, mas dá para ganhar sem cair em golpes*. Acesso em 28 de setembro de 2020.
- Pactual, B. (2017). *Diversificação de investimentos: o que é e quais as suas vantagens*. Acesso em 28 de setembro de 2020.
- R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rauber, T. W. (2005). Redes neurais artificiais. *Universidade Federal do Espírito Santo*, page 29.
- Shawn, R. (2017). *Blockchain Key Terms, Explained*. Acesso em 28 de setembro de 2020.
- Tanaka, M. (2018). *3 tipos de aprendizado caracterizam o Machine Learning*. Acesso em 28 de setembro de 2020.