CIÊNCIA DE DADOS E APRENDIZADO DE MÁQUINA PARA PREDIÇÃO EM SÉRIES TEMPORAIS FINANCEIRAS

CAIO MÁRIO HENRIQUES SILVA DA ROCHA MESQUITA

CIÊNCIA DE DADOS E APRENDIZADO DE MÁQUINA PARA PREDIÇÃO EM SÉRIES TEMPORAIS FINANCEIRAS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais - Departamento de Ciência da Computação como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Adriano César Machado Pereira

Belo Horizonte

Junho de 2019

Mesquita, Caio Mário Henriques Silva da Rocha

M583c Ciência de dados e aprendizado de máquina para predição em séries temporais financeiras / Caio Mário Henriques Silva da Rocha Mesquita. — Belo Horizonte, 2019

xxiii, 107 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais - Departamento de Ciência da Computação

Orientador: Adriano César Machado Pereira

1. Computação — Teses. 2. Ciência de dados. 3.Aprendizado do Computador. 4.Bolsa de Valores. I. Orientador. II. Título.

CDU 519.6*82 (043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

CIÊNCIA DE DADOS E APRENDIZADO DE MÁQUINA PARA PREDIÇÃO EM SÉRIES TEMPORAIS FINANCEIRAS

CAIO MÁRIO HENRIQUES SILVA DA ROCHA MESQUITA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. ADRIANO CÉSAR MACHADO PEREIRA - Orientador
Departamento de Ciência da Computação - UFMG

PROF. ARTHUR RODRIGO BOSCO DE MAGALHÃES
Departamento de Física e Matemática - CEFET-MG

PROF. CRISTIANO ARBEX VALLE
Departamento de Ciência da Computação - UFMG

PROF. PEDRO OLMO STANCIOLI VAZ DE MELO
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 19 de Junho de 2019.

Agradecimentos

Agradeço a Deus e minha família por todo apoio, incentivo e suporte incondicional ao longo de todos esses anos. A Stephany por todo carinho e companheirismo. Ao grupo de pesquisa em finanças computacionais (FICO) por todo auxílio e aprendizado.

Um agradecimento especial ao meu orientador Adriano César por toda a caminhada acadêmica que percorremos desde a graduação até hoje.

Agradeço também ao Programa de Pós Graduação em Ciência da Computação da Universidade Federal de Minas Gerais pela oportunidade. Aos professores do programa e amigos que contribuíram muito a minha formação. A CAPES pelo apoio financeiro ao longo do mestrado.

Enfim, a todos que contribuíram de alguma forma para que eu chegasse até aqui.

Muito obrigado!

Resumo

Ao longo da história surgiram diversos modelos de previsão com o objetivo de compreender o comportamento de séries de preços de ativos no mercado financeiro. O avanço do poder computacional tem facilitado a criação de novos modelos, cada vez mais complexos, que surgem com este propósito. Entretanto, mesmo com a utilização de técnicas avançadas de aprendizado de máquina utilizando um volume grande de dados históricos, tal tarefa continua sendo bastante desafiadora, permanecendo como um problema em aberto. O objetivo deste trabalho é criar estratégias automatizadas de operação no mercado, baseadas em um modelo de previsão de tendências nos preços das séries financeiras, por meio de aprendizado de máquina. É utilizada uma rede neural recorrente *Long Short Term Memory* como modelo de previsão. O trabalho também tem como objetivo demonstrar que várias das séries financeiras possuem uma correlação temporal, mesmo que pequena, o que viabiliza a construção de modelos de previsão que se baseiam em dados históricos. Para demonstrar essa correlação são analisadas as propriedades estatísticas das séries e aplicados testes de hipóteses nas mesmas. O trabalho apresenta uma metodologia robusta desde a coleta dos dados, até a simulação de operação no mercado envolvendo os custos de operação para 38 ativos da bolsa de valores brasileira. A metodologia ainda apresenta um método para criação de uma nova série mais correlacionada com valores futuros por meio de uma combinação linear das séries históricas em diferentes *lags* de tempo. Os resultados obtidos demonstram ser promissores, uma vez que os melhores modelos de predição obtiveram valores de Acurácia de até 63% e valores de retorno financeiro de até 47%. Os melhores casos obtiveram desempenhos superiores, tanto em termos de classificação quanto em termos de retorno financeiro comparados aos *baselines* de classificador aleatório, estratégia de *Buy and Hold*, taxas SELIC e CDI.

Palavras-chave: Mercado de Ações, Ciência de Dados, Séries Financeiras, Análise Estatística, Aprendizado de Máquina, Redes Neurais.

Abstract

Throughout history several forecasting models have emerged with the objective of understanding the behavior of asset price series in the financial market. The advancement of computational power has facilitated the creation of new, increasingly complex models that arise for this purpose. However, even with the use of advanced machine learning techniques using a large volume of historical data, this task remains quite challenging, remaining an open problem. The objective of this work is to create automated strategies of operation in the market, based on a forecast model of trends in the prices of financial series, through machine learning. A recurrent neural network Long Short Term Memory is used as the predictive model. The paper also aims to demonstrate that several of the financial series have a temporal correlation, even if small, which allows the construction of forecasting models that are based on historical data. In order to demonstrate this correlation, the statistical properties of the series are analyzed and hypothesis tests are applied to them. The work presents a robust methodology from the data collection to the simulation of operation in the market involving the operating costs for 38 assets of the Brazilian stock exchange. The methodology further presents a method for creating a more correlated attribute with future values by means of a linear combination of the historical series in different time lags. The results obtained are promising since the best forecasting models obtained Accuracy values of up to 63% and financial return values of up to 47%. The best cases outperformed both in terms of prediction and in terms of financial return compared to baselines techniques as random classifier, Buy and Hold strategy, SELIC and CDI rates.

Keywords: Stock Market, Data Science, Financial Series, Statistical Analysis, Machine Learning, Neural Networks.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

O mercado financeiro está intrinsecamente relacionado com a economia de um país. O mesmo é responsável por movimentar uma quantidade enorme de dinheiro diariamente. Muitos economistas, investidores, pesquisadores e acadêmicos estudam o mercado a fim de compreender seu comportamento na tentativa de conseguir uma previsão razoável dos ativos financeiros. Este problema têm sido estudado há mais de um século, e até hoje não foi noticiado nenhum modelo de previsão eficiente capaz de prever acontecimentos significativos, como por exemplo, crises financeiras. É improvável que algum modelo irá alcançar este resultado, uma vez que ao se tornar conhecido e utilizado por um grande número de pessoas, o mesmo deixará de ser eficaz. O mercado financeiro é dinâmico e a própria atuação por parte dos investidores acaba interferindo no seu comportamento. Entender como as ações deste mercado se comportam e variam é um desafio.

Com a tendência de informatização dos processos, hoje é possível para um investidor operar instantaneamente no mercado. O avanço do poder computacional e das tecnologias de armazenamento de dados tornou viável a análise histórica dos dados de ações. Qualquer investidor consegue planejar uma estratégia de investimento tomando como base dados históricos. Naturalmente surgem modelos computacionais cada vez mais complexos buscando extrair padrões nos dados. Muitos modelos de aprendizado de máquina têm sido elaborados recentemente com o intuito de capturar o comportamento das ações do mercado.

Uma grande parte dos investidores busca maximizar seus lucros operando no mercado em momentos específicos em que acreditam ter maiores chances de ganho (Kirkpatrick & Dahlquist [2006]). Muitos utilizam valores históricos dos preços. Entretanto Oliveira & Ziegelmann [2010] afirmam que o comportamento das séries financeiras é volátil dada a constante variação das mesmas. Muitos fatores podem influenciar na

oscilação das séries, tais como acontecimentos políticos, econômicos, sociais e a própria especulação, o que torna a tarefa de prever bons momentos de compra ou venda um desafio. A hipótese do mercado eficiente, proposta por Fama [1970], sugere que nenhum modelo de predição é capaz de prever preços futuros. Neste caso se a hipótese for verdadeira, seria impossível do investidor ter lucros acima do mercado. Contudo, vale ressaltar que essa teoria é bastante discutida na literatura, com correntes de pesquisadores tanto a favor quanto contra a hipótese e suas variações.

Este trabalho visa utilizar técnicas de ciência de dados e aprendizado de máquina para geração de estratégias financeiras eficientes e automatizadas. Propõe-se utilizar séries históricas reais de ativos da Bolsa de Valores, Mercadorias e Futuros de São Paulo (B3) para o processo de aprendizagem de uma rede neural. É utilizada uma rede recorrente *Long Short-Term Memory* devido ao seu bom desempenho em problemas envolvendo dados sequenciais na literatura (Chen et al. [2015], Zhao et al. [2017]).

Também é proposta uma análise detalhada das séries históricas, buscando destacar características e propriedades nas mesmas que evidenciam a presença de comportamentos lineares e não lineares ao longo do tempo, uma vez que estas séries são utilizadas como entradas para o algoritmo de previsão. É fundamental destacar evidências de dependências nas séries para corroborar a utilização de um modelo de previsão baseado em valores do passado.

Para validação do desempenho do modelo são analisadas os resultados em termos de classificação, como Acurácia e Precisão, e também os retornos financeiros alcançados pelo modelo. Os resultados obtidos são comparados com *baselines* comuns na literatura: previsor aleatório, estratégia de *Buy and Hold*, rendimentos das taxas SELIC e CDI.

Espera-se que com o desenvolvimento deste trabalho seja possível auxiliar na metodologia para elaboração de modelos de previsão de séries financeiras.

1.1 Motivação

Recentemente é possível encontrar muitos trabalhos na literatura que utilizam séries temporais financeiras com o objetivo de prever tendência nos preços das mesmas (Samarawickrama & Fernando [2017], Nelson et al. [2017], Huang et al. [2018]). São adotados modelos de previsão baseados nas mais diversas técnicas, como por exemplo: médias móveis (Gunasekarage & Power [2001]), médias exponenciais (Nakano et al. [2017]), modelos de regressão (Roy et al. [2015]), modelos de aprendizado de máquina (Huang et al. [2005], Kara et al. [2011], Selvin et al. [2017]) entre outros. Entretanto, até o presente momento não há nenhum registro de um trabalho que tenha conseguido al-

cançar resultados bons de forma robusta, em diversos cenários e para séries diferentes, conforme já mencionado na introdução do trabalho. Ainda é considerado um problema aberto conseguir um bom modelo de previsão.

Pode-se então questionar se é possível fazer uma previsão razoável para dias futuros com base apenas nos dados de preço do passado. Sendo assim, a primeira pergunta que deve-se ter em mente é: os dados de valores de preço do passado ajudam de alguma forma a prever os valores futuros? Ou será que os valores de preço são independentes entre si e são de certa forma aleatórios? Como verificar qual dessas duas hipóteses é mais provável? Se os dados de preço futuro forem independentes dos dados passados, não há sentido para tentar fazer uma previsão com esses dados. Mas se os dados forem dependentes, mesmo que a dependência seja pequena, é necessário entender esse relacionamento dos dados passados com os futuros para poder fazer uma previsão razoável. Neste caso, será que já existe algum modelo robusto que consiga capturar essa possível dependência? E para o caso dos dados serem dependentes surgem alguns questionamentos: qual modelo de previsão se deve escolher para tentar prever a tendência de uma série temporal específica? Por que um modelo escolhido a priori deverá obter resultados melhores sobre outros modelos? Quais características deste modelo o tornam mais apto para prever a série escolhida? Existe alguma metodologia para se escolher um modelo específico para este problema?

Este trabalho parte da premissa de que as séries históricas possuem, em certos períodos, padrões complexos que são possíveis de serem explorados. Esta hipótese se baseia no fato de que na literatura vários trabalhos foram feitos buscando refutar a hipótese de passeio aleatório nas mesmas, por exemplo Kim & Shamsuddin [2008] e Hoque et al. [2007]. Outro ponto que sustenta a hipótese de que as séries não são aleatórias é que recentemente alguns trabalhos que aplicaram algoritmos de aprendizado de máquina, obtiveram resultados significativos em comparação com previsores aleatórios, por exemplo Chen et al. [2015] e Zhao et al. [2017]. Já é sabido que se tais padrões realmente existirem, não são lineares, nem simples de serem entendidos ou explorados (Taylor [2007]). Por este motivo, uma possível opção para a construção de modelos de previsão seria aplicar algoritmos de aprendizado de máquina capazes de explorar padrões complexos e não lineares nos dados.

1.2 Objetivos

O principal objetivo desta dissertação é aplicar ciência de dados e aprendizado de máquina para previsões em séries financeiras, procurando explorar padrões e tendências

nas séries. A ideia é utilizar as previsões como base para elaboração de estratégias de investimento automatizadas, validando com dados reais do mercado financeiro. A fim de se obter bons resultados em termos de previsão, as séries são analisadas e destacadas suas principais características e propriedades estatísticas. Pode-se citar como objetivos específicos do trabalho:

- Analisar as propriedades estatísticas das séries financeiras e alguns fatos estilizados das mesmas;
- Testar hipóteses de dependência temporal nas séries por meio de testes estatísticos apropriados;
- Explorar a possível dependência temporal criando um novo atributo mais dependente;
- Aplicar um modelo de previsão de aprendizado de máquina (rede neural *Long Short Term Memory*) utilizando os dados históricos e o atributo criado para previsão de preços futuros;
- Analisar os resultados encontrados, tanto em termos de métricas de classificação quanto em retornos financeiros.

1.3 Contribuições

Espera-se que com este trabalho seja possível trazer novas ideias e abordagens no desenvolvimento de modelos de previsão baseados em dados históricos das séries financeiras. Pode-se citar como contribuições:

- Melhor entendimento do comportamento das séries financeiras da B3;
- Elaboração de uma metodologia sistemática para desenvolver melhores algoritmos de previsão de tendência;
- Demonstração de uma validação robusta para avaliar os resultados das previsões do trabalho.

1.4 Organização do trabalho

No Capítulo 2 são apresentados os conceitos e fundamentos em que o trabalho se baseia. Em seguida no Capítulo 3 são citados trabalhos da literatura relacionados ao tema da

dissertação, enfatizando os temas de análise e propriedades das séries financeiras, testes estatísticos aplicados as mesmas, o conceito de correlação de distância e suas aplicações, e técnicas de aprendizado de máquina aplicadas no mercado financeiro. No Capítulo 4 é descrita a metodologia desenvolvida na dissertação, com a explicação de cada etapa abordada no trabalho. O Capítulo 5 descreve e analisa os resultados obtidos pelos experimentos desenvolvidos. Por fim o Capítulo 6 apresenta a conclusão do trabalho e descreve tópicos que podem ser explorados em trabalhos futuros.

Capítulo 2

Fundamentação teórica

Neste Capítulo são abordados os principais temas que fundamentam a pesquisa, divididos em 9 seções. Na Seção 2.1 são apresentados conceitos básicos sobre mercado financeiro. Já a Seção 2.2 aprofunda o tema das séries temporais financeiras, que são os objetos de estudo do trabalho. Nas Seções 2.3 a 2.6 são apresentadas medidas de correlação utilizadas extensivamente ao longo do trabalho. Em seguida, na Seção 2.7 e 2.8 são apresentados os conceitos de redes neurais artificiais e uma arquitetura de rede neural específica (LSTM), responsável pela implementação do modelo de predição no trabalho. Por fim, na Seção 2.9 é apresentado o algoritmo *Differential Evolution* que é utilizado na etapa de criação de um novo atributo como entrada no modelo de predição.

2.1 Mercado Financeiro

O mercado financeiro é o local onde são negociadas as compras e vendas de valores mobiliários, mercadorias e câmbio. Perfeito [2011] afirma que as empresas abrem seu capital e negociam seus títulos visando atrair investimentos. Carvalho [2010] destaca que estes mercados possuem um papel fundamental na economia já que quanto mais investimentos nos mesmos, maior o aquecimento da economia. Ainda segundo Reis [2007] o mercado financeiro é constituído de um conjunto de vários mercados conectados em que são feitas negociações financeiras por meio dos investidores. As negociações são realizadas por sistemas de informações, permitindo um grande volume de negócios realizados em um curto intervalo de tempo.

Fortuna [2007] afirma que o preço de uma ação na bolsa de valores decorre da oferta e demanda do mercado no momento. E pode-se dizer que a maioria dos investidores busca maximizar seus ganhos baseando-se na ideia de comprar um ativo,

por exemplo, por um preço baixo e vende-lo por um valor maior, obtendo um retorno financeiro positivo. Entretanto saber o momento certo de comprar e vender o ativo se torna um desafio uma vez que os preços oscilam constantemente. Essa variação no preço é afetada por diversos fatores econômicos, políticos, notícias e especulação por parte dos investidores, tornando um desafio a tarefa de previsão (Perfeito [2011]).

No Brasil a Bolsa de Valores, Mercadorias e Futuros de São Paulo (B3) é responsável por administrar todas as negociações de títulos, valores mobiliários e contratos derivativos. A mesma realiza serviços de registro, compensação, liquidação e atua como contraparte garantidora da liquidação financeira das operações realizadas pelos investidores. Atualmente a B3 possui cerca de 500 empresas cadastradas e em sua página Web são disponibilizados arquivos de séries temporais de cotações históricas desde a década de 80 até os dias atuais. Estes arquivos são públicos e possuem as seguintes informações: nome da empresa, código, tipo de mercado (a vista, termo, opções), preços (abertura, fechamento, máximo, mínimo, médio e anterior), quantidade de negócios, volume negociado, dentre outas informações.

Fama [1970] propõe a hipótese do mercado eficiente (HME) afirmando que os preços das ações refletem instantaneamente toda a informação disponível, não sendo possível um investidor obter vantagens sobre o mercado. Pesquisas envolvendo esta hipótese demonstram que raramente o mercado cria oportunidades possibilitando valores altos de lucro. Para corroborar a hipótese, já foram realizados experimentos demonstrando que uma estratégia de investimento utilizando um agente aleatório consegue produzir resultados melhores do que indicadores técnicos tradicionais (Biondo et al. [2013]). Entretanto, existem anomalias que não podem ser explicadas pela HME, e há uma outra corrente de pesquisadores que acredita ser possível prever variações no comportamento do preço das ações até certo nível. Lo [2004] propõe a hipótese do mercado adaptativo, em que o mercado na maior parte do tempo é eficiente mas que há períodos de tempo em que nem todas as informações são refletidas no preço das ações de forma instantânea. Alguns trabalhos empíricos defendem essa segunda corrente, como por exemplo Lo & MacKinlay [2011].

2.2 Séries temporais financeiras

Uma série temporal pode ser definida como um conjunto de observações sequenciais ao longo do tempo. Morettin [2006] cita que um dos principais objetivos em estudá-las é poder realizar previsões de valores futuros da série, geralmente de curto prazo.

Nas mais variadas áreas de conhecimento pode-se citar exemplos de séries tempo-

rais: valores de temperatura ao longo do ano, valores diários de poluição, precipitação atmosférica em um local, registro de marés, registro de roubos de veículo em uma cidade, entre muitos outros. Pommeranzenbaum [2014] afirma que um grupo de séries que se destaca são as séries financeiras, geralmente compostas por cotações históricas de ativos da bolsa de valores. A Figura 2.1 ilustra uma série temporal financeira.

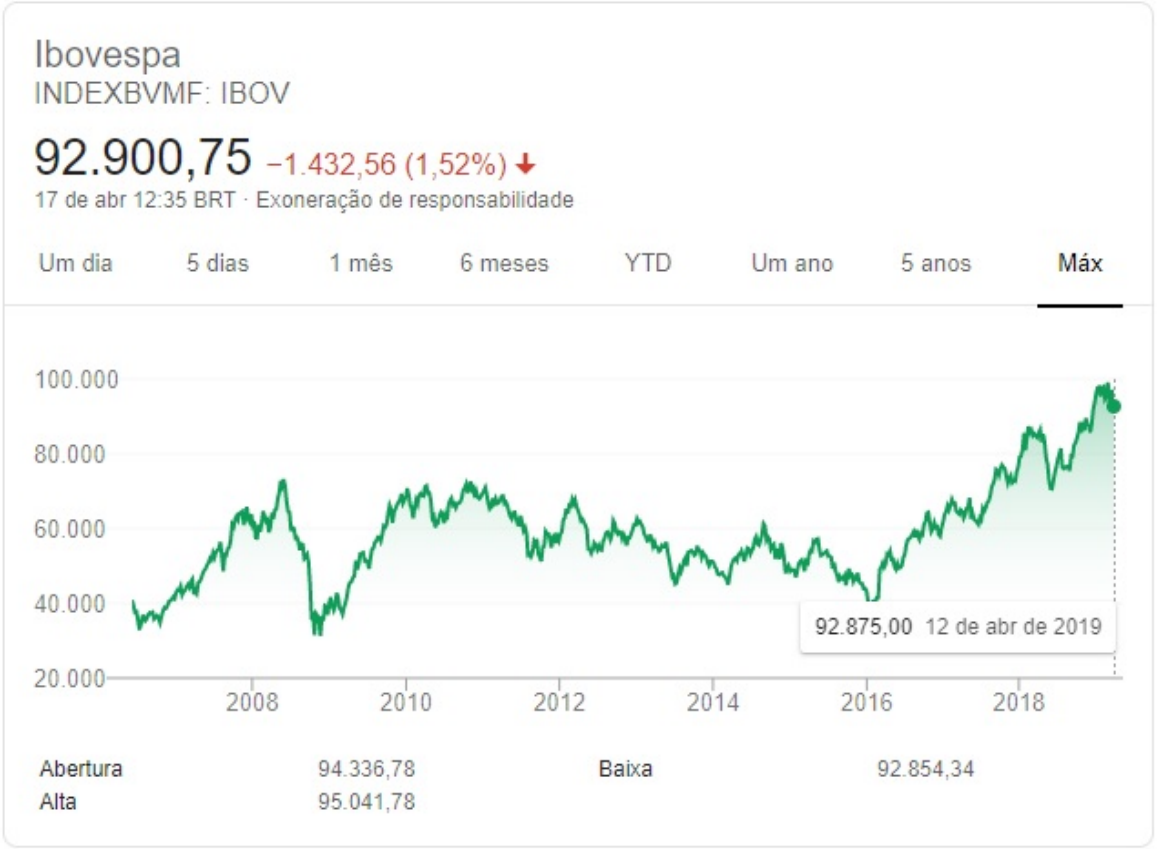

Figura 2.1. Exemplo de uma série temporal financeira: índice Ibovespa ao longo de 10 anos. Fonte Google [2019]

Mantegna & Stanley [2000] afirmam que a imprevisibilidade nas séries financeiras é um conceito majoritariamente aceito. E segundo Oliveira & Ziegelmann [2010], um dos motivos é a não linearidade das séries devido à natureza especulativa. Os autores também destacam a influência de fatores externos responsáveis por aumentar o comportamento aleatório.

Segundo Taylor [2007], a análise estatística diretamente aplicada nos preços originais é difícil, pois preços consecutivos são altamente correlacionados e a variância no preço aumenta ao longo do tempo. Preços não são estacionários e consequentemente se torna mais conveniente a análise na variação do preço (retornos). Resultados ob-

tidos para os retornos podem ser facilmente usados para fornecer resultados no preço original.

É conhecido na literatura que essas séries de retornos apresentam características singulares, os chamados fatos estilizados, Cont [2001]. A seguir são descritos alguns dos fatos estilizados abordados no trabalho:

- Distribuição dos retornos: as distribuições de probabilidade não são normais. Elas possuem caudas pesadas, são simétricas e possuem um pico elevado maior que uma distribuição normal. Há um número significativo de valores extremos nas distribuições.

- Estacionariedade e heterocedasticidade nas séries: as séries de retornos podem ser consideradas estacionárias em relação a média, entretanto é conhecido que a variância não é estacionária ao longo do tempo.

- Ausência de correlação linear das séries: valores de correlação linear são insignificantes para as séries de retornos.

- Correlação linear positiva das séries transformadas: ao analisar as transformações das séries em valores absolutos e elevados ao quadrado, são constatados valores significativos e positivos de correlação linear. Este fato sugere que existe uma dependência não linear nas séries de retornos.

- Agrupamento da volatilidade: ao analisar as séries é possível identificar períodos em que a volatilidade é alta. Nestes períodos, grandes alterações nos preços são percebidas consecutivamente. Esta característica está diretamente relacionada a dependência não linear das séries de retorno.

2.2.1 Passeio aleatório

A ideia de passeio aleatório sugere que os retornos dos preços das ações variam de uma maneira não previsível, em uma forma aleatória. Taylor [2007] afirma que existem diferentes definições para esta hipótese, sendo que uma delas assume que os retornos tenham distribuições idênticas e independentes (i.i.d). Entretanto, segundo o autor, a hipótese i.i.d. não é muito relevante em termos da previsibilidade dos retornos. Por esse motivo a definição adotada neste trabalho da hipótese de passeio aleatório é a mesma de Taylor [2007], sendo que os retornos possuem médias idênticas e distribuições não

correlacionadas

$$\begin{aligned} E[r_t] &= E[r_{t+\tau}] \\ cov(r_t, r_{t+\tau}) &= 0, \end{aligned} \tag{2.1}$$

para todo t e $\tau > 0$. Sendo o logaritmo do retorno financeiro igual a $r_t = log(p_t) - log(p_{t-1})$, em que p_t corresponde ao preço do ativo no dia t, e $r_{t+\tau} = log(p_{t+\tau}) - log(p_{t+\tau-1})$.

Ao longo do trabalho, tanto a hipótese das séries serem i.i.d.s, quanto a hipótese de passeio aleatório, são testadas com o objetivo de evidenciar possíveis dependências temporais nas séries.

2.2.2 Teste contra a hipótese de passeio aleatório: teste da razão da variância

As comparações entre a variância dos retornos de um período com a soma de retornos de multi períodos é utilizada para testar a hipótese do passeio aleatório (HPA). Vários testes tentam explorar quaisquer divergências desta predição, sendo um dos mais famosos o teste da razão da variância proposto por Lo & MacKinlay [1988]. O teste se baseia na seguinte propriedade: se a série de retornos financeiros segue um passeio aleatório, a variância da soma de retornos consecutivos deve ser igual a soma das variâncias individuais. Por exemplo, a variância de $r_t + r_{t+1}$ deve ser igual a duas vezes a variância de r_t. Seguindo o mesmo desenvolvimento adotado em Taylor [2007], a intuição por trás do teste é descrita a seguir.

Suponha que o processo gerador de retornos é estacionário (com média e variância constantes), com a variância do retorno de um período sendo igual $V(1) = var(r_t)$. A variância da soma dos retornos de dois períodos consecutivos é igual a

$$V(2) = var(r_t + r_{t+1}) = var(r_t) + var(r_{t+1}) + 2cov(r_t, r_{t+1}) = (2 + 2\rho_1)V(1), \tag{2.2}$$

com ρ_1 sendo igual ao coeficiente de autocorrelação de primeira ordem dos retornos $\{r_t\}$. A razão da variância dos dois períodos é então definida como

$$VR(2) = \frac{V(2)}{2V(1)} = 1 + \rho_1. \tag{2.3}$$

O termo da autocorrelação é zero quando HPA se aplica e então a razão da variância é igual a 1. Caso contrário, a hipótese de HPA é falsa e a razão pode ser tanto maior ou menor do que 1.

Considerando um período de N retornos, sendo N um inteiro maior ou igual a 2. Quando a hipótese de HPA é verdadeira,

$$V(N) = var(r_t + r_{t+1} + ... + r_{t+N-1}) = var(r_t) + var(r_{t1}) + ... + var(r_{t+N-1}) = NV(1) \tag{2.4}$$

e assim a variância é 1 para todo N

$$VR(N) = \frac{V(N)}{NV(1)} = 1. \tag{2.5}$$

Quando a hipótese de HPA é falsa, V(N) é igual a $NV(1)$ mais os termos de covariância entre todos os pares de retornos distintos, sendo assim

$$V(N) = NV(1) + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} cov(r_{t+i-1}, r_{t+j-1}) \tag{2.6}$$

$$V(N) = V(1)[N + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \rho_{j-i}] \tag{2.7}$$

$$VR(N) = 1 + \frac{2}{N}\sum_{\tau=1}^{N-1}(N-\tau)\rho_\tau. \tag{2.8}$$

O teste empírico utiliza os retornos observados para decidir se a amostra estimada da razão da variância é compatível com a predição teórica sendo igual a 1. O teste rejeita a hipótese de HPA quando o valor da razão diverge de 1. Isto acontece quando uma função linear das $(N-1)$ primeiras autocorrelações

$$(N-1)\rho_1 + (N-2)\rho_2 + (N-3)\rho_3 + ... + 2\rho_{N-2} + \rho_{N-1}, \tag{2.9}$$

se distanciam de zero. Todos os coeficientes são positivos e eles diminuem à medida que o lag aumenta. Sendo assim, para um conjunto de n observações de retornos financeiros com média igual a $\bar{r}$ e variância $\hat{V}(1) = \sum (r_t - \bar{r})^2/(n-1)$, a estimativa de $V(N)$ pode ser definida da seguinte forma

$$\hat{V}(N) = \frac{n}{(n-N)(n-N+1)} \sum_{t=1}^{n-N+1} (r_t + r_{t+1} + ... + r_{t+N-1} - N\bar{r})^2 \tag{2.10}$$

e a razão $VR(N) = \frac{\hat{V}(N)}{N\hat{V}(1)} = VR(r, N)$

$$VR(r, N) = \frac{1}{nN}\sum_{t=N}^{n}(r_t + r_{t+1} + ... + r_{t+N-1} - N\bar{r})^2 \div \frac{1}{n}\sum_{t=1}^{n}(r_t - \bar{r})^2 \tag{2.11}$$

Apesar do raciocínio acima ser adotado para um processo estacionário, Lo & MacKinlay [1988] demonstram que a propriedade da linearidade (variância da soma ser igual a soma das variâncias) se mantém mesmo para séries que não possuem a variância constante. Os autores definem então a estatística M2 que é utilizada quando a série exibe heterocedasticidade condicional, sendo igual a

$$M2(N) = \frac{VR(r, N) - 1}{\phi(N)^{0.5}}. \tag{2.12}$$

M2 segue uma distribuição normal assintoticamente sob a hipótese nula de que $VR(N) = 1$, em que

$$\phi(N) = \sum_{j=1}^{N-1}[\frac{2(N-j)}{N}]^2\delta(j) \tag{2.13}$$

e

$$\delta(j) = \{\sum_{t=j+1}^{n}(r_t - \hat{u})^2(r_{t-j} - \hat{u})^2\} \div \{[\sum_{t=1}^{n}(r_t - \hat{u})^2]^2\}. \tag{2.14}$$

Este teste é utilizado na etapa na caracterização das séries para testar a hipótese de passeio aleatório nas séries.

2.3 Correlação de Pearson

A medida de correlação foi proposta por Pearson [1895] e é responsável por medir a relação linear entre duas variáveis X e Y. Os valores obtidos variam entre -1 a 1. O sinal da medida indica a direção da associação entre X e Y. Se X aumenta e Y também aumenta, o valor da correlação é positivo. Já se X aumenta e Y diminui, o valor é negativo. Valores próximos de zero indicam que as variáveis possuem correlação linear muito fraca, sendo que valores iguais a zero indicam a ausência de correlação linear. Dadas duas amostras, pode-se utilizar a seguinte fórmula para calcular a correlação entre elas

$$cor(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{2.15}$$

Para testar a significância estatística da correlação entre duas amostras de variáveis aleatórias pode-se utilizar diversos testes como por exemplo, teste de permutação, informação de Fisher, entre outros.

Pode-se mencionar algumas desvantagens associadas à medida de correlação. Uma desvantagem é que a medida captura somente relações lineares. Outra desvantagem é que valores próximos de zero indicam independência linear mas pode existir dependências não lineares entre as variáveis. Levine et al. [2008] afirmam que as relações entre duas variáveis correlacionadas são de tendência e não de causa e efeito. Segundo os mesmos autores "A correlação por si só não consegue provar que existe causa e efeito - ou seja, que a alteração no valor de uma variável causou uma alteração na outra variável. Uma forte correlação pode ser produzida simplesmente pelo acaso, pelo efeito de uma terceira variável não considerada no cálculo da correlação ou, ainda, por uma relação de causa e efeito."

A medida de correlação é utilizada em várias etapas da caracterização das séries de retorno financeiro, de forma a destacar relações lineares ao longo do tempo.

2.4 Função de autocorrelação

Taylor [2007] afirma que a correlação entre duas variáveis aleatórias X_t e $X_{t+\tau}$ de um processo estacionário, é chamada de autocorrelação no *lag* τ, sendo a notação ρ_τ utilizada para esta correlação. Sendo as variâncias de X_t e $X_{t+\tau}$ ambas iguais a λ_0, define-se

$$\rho_\tau = \frac{cov(X_t, X_{t+\tau})}{\lambda_0} = \frac{\lambda_\tau}{\lambda_0}, \tag{2.16}$$

logo $\rho_0 = 1$ e $-1 \leq \rho_\tau \leq 1$.

A função de autocorrelação é utilizada no trabalho na parte de caracterização das séries com o objetivo de identificar relações lineares temporais. Também é utilizada como base em testes estatísticos para testar a hipótese das séries serem independentes e identicamente distribuídas.

2.5 Correlação de Distância

Visando superar algumas das desvantagens da correlação de Pearson e obter uma medida que fosse capaz de capturar dependências lineares e não lineares, Székely et al. [2007] desenvolveram a correlação de distância. O valor da medida varia entre 0 e 1, em que quanto mais próximo de 1 mais forte a dependência entre as duas variáveis, e

quanto mais próximo de zero mais fraca é a dependência. Uma vantagem da correlação de distância sobre a correlação de Pearson é que para valores iguais a 0 pode-se afirmar que ambas variáveis são independentes. A seguir é definido o conceito da correlação de distância.

Sejam (X_k, Y_k) com $k = 1, 2, ..., n$ amostras de um par de duas variáveis aleatórias (X, Y). As matrizes de distâncias em pares $n\ x\ n$ são definidas como sendo igual a

$$a_{jk} =|| X_j - X_k ||, j, k = 1, 2, ..., n; \qquad b_{jk} =|| Y_j - Y_k ||, j, k = 1, 2, ..., n \tag{2.17}$$

em que $||\ .\ ||$ é igual à norma Euclidiana. Sejam também

$$A_{jk} = a_{jk} - \bar{a_{j.}} - \bar{a_{.k}} + \bar{a_{..}}; \qquad B_{jk} = b_{jk} - \bar{b_{j.}} - \bar{b_{.k}} + \bar{b_{..}} \tag{2.18}$$

em que $\bar{a_{j.}}$ é igual à média da linha j, $\bar{a_{k.}}$ é igual à média da coluna k e $\bar{a_{..}}$ é igual a *grand mean* (média das médias) da matriz de distância da amostra X. Seguindo a mesma notação para os valores de b, pode-se definir a distância de covariância ao quadrado como sendo igual a

$$dCov(X, Y)^2 = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} A_{jk} B_{jk} \tag{2.19}$$

e a distância de variância ao quadrado como sendo igual a

$$dVar(X)^2 = dCov(X, X)^2 = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} A_{jk}^2. \tag{2.20}$$

Segue-se então a fórmula da correlação de distância

$$dCor(X, Y) = \frac{dCov(X, Y)}{\sqrt{dVar(X) dVar(Y)}} \tag{2.21}$$

SzéKely & Rizzo [2013] afirmam que uma desvantagem que pode se citar em relação a medida original de correlação de distância é que principalmente para variáveis com alta dimensionalidade essa medida pode ser enviesada, apresentando valores próximos de 1 mesmo para variáveis independentes. Este fato torna difícil analisar a medida sem um teste formal. Os mesmos autores propõem então uma medida modificada com objetivo de evitar esse viés na medida original. Sendo

$$\mathrm{A}_{ij}^{*} = \begin{cases} \frac{n}{n-1}(A_{ij} - \frac{a_{ij}}{n}),\ i \neq j \\ \frac{n}{n-1}(\bar{a_i} - \bar{a}),\ i = j \end{cases} \qquad B_{ij}^{*} = \begin{cases} \frac{n}{n-1}(B_{ij} - \frac{b_{ij}}{n}),\ i \neq j \\ \frac{n}{n-1}(\bar{b_i} - \bar{b}),\ i = j \end{cases} \tag{2.22}$$

a medida de distância de covariância modificada é igual a

$$V_n^{*}(X,Y) = \frac{1}{n(n-3)}\{\sum_{i,j=1}^{n} A_{ij}^{*}B_{ij}^{*} - \frac{n}{n-2}\sum_{i=1}^{n} A_{ij}^{*}B_{ij}^{*}\} \tag{2.23}$$

e consequentemente a medida de correlação de distância modificada é igual a

$$R_n^{*}(X,Y) = \frac{V_n^{*}(X,Y)}{\sqrt{V_n^{*}(X,X)V_n^{*}(Y,Y)}} \tag{2.24}$$

se $V_n^{*}(X,X)V_n^{*}(Y,Y) > 0$, caso contrário $R_n^{*}(X,Y) = 0$. Em seguida os autores apresentam a estatística

$$\tau_n = \sqrt{v-1}\frac{R_n^{*}}{\sqrt{1-(R_n^{*})^2}}, \tag{2.25}$$

que segue uma distribuição t de Student com $v-1$ graus de liberdade, sob a hipótese de independência e com as dimensões de X e Y tendendo ao infinito. Neste caso, $v = \frac{n(n-3)}{2}$. Os mesmos autores discutem a aplicação da estatística t de correlação de distância com o objetivo de testar a independência entre duas séries temporais. Entretanto para utilização do teste é necessário possuir observações i.i.d.s das séries. É então apresentado o seguinte procedimento para obtenção de uma amostra aleatória i.i.d. das séries:

- Fixe $p < N$ e seja $T_1, T_2, ..., T_n$ inteiros entre $\{1, ..., N-p+1\}$. Defina X_j a subsequência de tamanho p começando em $X(T_j)$ e similarmente defina Y_j, ou seja
- $X_j = \{X(T_j), X(T_j+1), ..., X(T_j+p-1)\}, j = 1, ..., n,$
- $Y_j = \{Y(T_j), Y(T_j+1), ..., Y(T_j+p-1)\}, j = 1, ..., n.$

Se T_j são retirados de forma aleatória com igual probabilidade de $\{1, ..., N-p+1\}$ esses vetores $\{X_j\}$ são observações i.i.d.s. Similarmente os vetores $\{Y_j\}$ também são observações i.i.d. e pode-se aplicar o t teste de independência incondicionalmente.

A medida da correlação de distância é utilizada no trabalho para verificar dependências não lineares ao longo das séries e para criação de um novo atributo mais correlacionado, na etapa de utilização da rede neural como modelo de predição. O teste T apresentado acima é utilizado para testar a independência entre o novo atributo e a série de valores futuros do conjunto de treinamento da rede neural.

2.6 Função de auto correlação de distância (ADCF)

Zhou [2012] propõe estender o conceito de correlação de distância para o campo de séries temporais a fim de explorar e testar as dependências não lineares nas séries. É desenvolvida então a função de auto correlação de distância (ADCF) que, de acordo com o autor, seu valor é igual a zero se os componentes da série temporal forem independentes. O autor também afirma que a ADCF é capaz de medir dependências não lineares complexas que não são possíveis por meio da correlação de Pearson.

A definição da ADCF amostral é análoga à equação (2.21) com a segunda amostra sendo substituída pela série atrasada, ou seja $Y_j = X_{j+k}$ para $j > 0$. Segundo o autor, a ACDF amostral entre X_j e X_{j+k} é o número não negativo $R_x^n(k)$ definido por

$$[R_x^n(k)]^2 = \frac{V_x^n(k)}{\sqrt{V_y^n(0)V_x^n(0)}}, \tag{2.26}$$

se $V_y^n(0)V_x^n(0) \neq 0$, sendo $V_y^n(0)$ igual à distância de variância amostral de $\{Y_j\}_{j=1}^{n-k}$ e $V_x^n(0)$ igual a distância de variância amostral de $\{X_j\}_{j=1}^{n-k}$. Caso contrário, $R_x^n(k) = 0$. O autor ainda sugere uma metodologia baseada em subamostragem das séries para calcular os valores críticos de 95% de confiança com o objetivo de testar a independência entre os *lags*.

Fokianos & Pitsillou [2017] propuseram uma estatística baseada em ADCF para testar a hipótese das séries serem i.i.d.s. A estatística T_n é definida da seguinte forma

$$T_n = \sum_{j=1}^{n-1}(n-j)k^2(\frac{j}{p})R_X^2(j), \tag{2.27}$$

sendo $k(.)$ uma função kernel $k : R \longrightarrow [-1, 1]$ simétrica e contínua, com $k(0) = 1$,

$$\int_{-\infty}^{\infty} k^2(z)dz \leq \infty$$

e $\mid k(z) \mid < C \mid z \mid^{-b}$ para z grande e $b > 0.5$. E seja $p = cn^{\lambda}$ (largura de banda da

função kernel) para $c > 0$ e $\lambda \in (0, 1)$. Sob essas condições e a hipótese nula de que a série é i.i.d., a estatística T_n segue uma distribuição normal com média igual a 0 e desvio padrão igual a 1.

A função de auto correlação de distância é utilizada no trabalho com o objetivo de identificar correlações não lineares ao longo das séries, e verificar se os *lags* são independentes. Também é utilizada a estatística T_n para testar a hipótese i.i.d. das séries.

2.7 Redes Neurais Artificiais

A ideia por trás das redes neurais artificiais surgiu em Mcculloch & Pitts [1943] quando os pesquisadores da época buscavam simular computacionalmente o funcionamento das células de um cérebro. Rosenblatt [1958] afirma que uma rede neural é um modelo matemático com a capacidade de aprendizagem e generalização. Este modelo se assemelha com os sistemas nervosos dos seres vivos. Oliveira & Ziegelmann [2010] sugerem que as redes neurais buscam simular a capacidade de aprendizado na aquisição de conhecimento do cérebro humano.

Segundo de Pádua Braga [2007], uma rede neural é composta por nodos(neurônios) paralelos responsáveis por calcular uma determinada função matemática relacionada ao aprendizado da rede. Estes neurônios são organizados em camadas que se interligam geralmente de forma unidirecional por meio de diversas conexões. Na maioria das vezes, a cada neurônio está associado um peso que armazena o aprendizado do modelo e é responsável por ponderar a entrada recebida a cada neurônio da rede.

As redes neurais possuem a capacidade de realizar previsões. Para isso é necessário treinar a rede com com uma série de dados de forma que ocorra o aprendizado e a rede possa em seguida inferir os próximos valores [Basheer & Hajmeer, 2001]. As redes neurais artificiais (RNAs) possuem a capacidade de aproximar funções, são robustas e tolerante a falhas [Neto et al., 2010]. Por possuir essas características, as RNAs se tornam boas candidatas para previsão de sistemas não lineares e séries temporais não estacionárias como é o caso das séries temporais financeiras.

2.8 Rede Neural *Long Short Term Memory*

As redes neurais *Long-Short Term Memory* (LSTM) foram propostas por Hochreiter & Schmidhuber [1997] e são um tipo de rede neural recorrente de aprendizado profundo. Redes recorrentes possuem conexões de realimentação entre processamentos

passados da rede e as entradas do momento presente, obtendo-se então a característica de memória. Dessa maneira é possível encontrar correlações entre eventos separados por um longo intervalo de tempo. Por este motivo essas redes são ideais para tarefas de classificação, processamento e predição em séries temporais de dados.

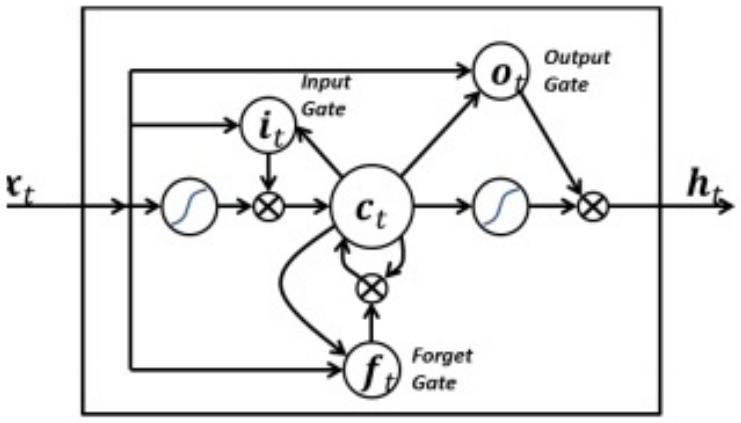

Figura 2.2. Representação de uma unidade LSTM com suas respectivas portas, entradas e saídas. Figura do trabalho de Greff et al. [2015]

O desafio em utilizar redes recorrentes comuns para processar longas sequências de dados é que elas sofrem com um problema conhecido na literatura como desaparecimento do gradiente. Ao utilizar métodos de treinamento baseados na retropropagação do gradiente do erro, a atualização de pesos de camadas mais distantes pode não acontecer e neste caso a rede pode interromper seu aprendizado. As redes LSTM foram então propostas com a motivação de oferecer melhores desempenhos para estes casos, ao utilizar portas (*input gate, forget gate, output gate*) capazes de descartar, manter, adicionar ou atualizar informações no tempo. A utilização destas portas ajuda na preservação do erro que pode ser retropropagado através do tempo e das camadas. Mantendo o erro mais constate é possível continuar o aprendizado no treinamento ao longo de muitos instantes de tempo e associar nos dados causas e efeitos remotos.

A porta *forget gate* é responsável por decidir quanto da informação deve ser esquecida, ao se comparar as entradas do momento atual x_t com os os valores do estado passado h_{t-1}. É utilizada uma função sigmoide em que a saída é composta por números no intervalo entre 0 e 1 indicando a quantidade que a informação deve fluir. Já a porta *input gate* é responsável por escolher os valores a serem atualizados, e uma camada *tanh* cria um vetor de novos candidatos que poderão ser adicionados ao estado atual. A porta *output gate* é responsável por fornecer a saída baseando-se no estado

atual. É utilizada uma combinação de uma função sigmoide e uma camada *tanh* para filtrar a saída. A Figura 2.2 ilustra os componentes de uma unidade LSTM.

Neste trabalho é utilizado como modelo de predição uma rede neural LSTM.

2.9 Algoritmo de otimização: *Differential Evolution*

O algoritmo *differential evolution* (DE) é uma abordagem heurística para problemas de otimização não lineares e não diferenciáveis. Foi proposto por Storn & Price [1997] e de acordo com os autores é um algoritmo rápido, robusto, fácil de utilizar e com um nível de confiança maior que outros métodos de otimização global. O DE é divido em quatro etapas: inicialização dos vetores, mutação , cruzamento e seleção. A Figura 2.3 ilustra o processo.

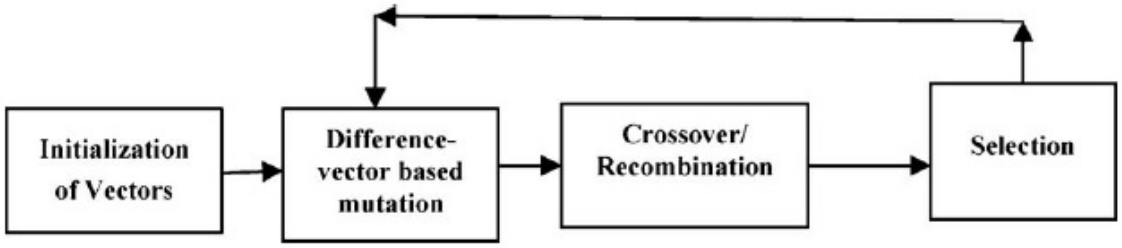

Figura 2.3. Etapas do algoritmo DE. Figura do trabalho de Das & Suganthan [2011]

Na etapa da inicialização dos vetores é gerada aleatoriamente uma população de NP vetores D-dimensinais, x_{iG} com $i = 1, 2, ..., NP$. Cada vetor representa uma solução para o problema de otimização. A população inicial de vetores é gerada aleatoriamente. Cada geração seguinte é representada por $G = 0, 1, ..., Gmax$. Uma nomenclatura é utilizada para diferenciar os vetores durante as etapas do algoritmo. Um vetor pai de uma geração corrente é denominado vetor alvo, um vetor mutante obtido por meio da operação de mutação é denominado vetor doador e o vetor obtido pela recombinação do vetor alvo e doador é chamado de vetor experimental.

Na etapa da mutação são gerados os vetores doadores. Para cada vetor alvo da população atual, três outros vetores diferentes X_{r1}, X_{r2}, X_{r3} são amostrados aleatoriamente da população atual, sendo que os índices r_1, r_2, r_3 são inteiros mutualmente exclusivos escolhidos aleatoriamente no intervalo $[1, NP]$. A diferença entre dois destes

três vetores é escalada por um fator F$\in [0, 2]$ e o resultado é adicionado ao terceiro vetor obtendo-se o vetor doador $V_{i,G+1}$

$$V_{i,G+1} = X_{r1,G} + F.(X_{r2,G} - X_{r3,G}). \tag{2.28}$$

A Figura 2.4 ilustra este processo.

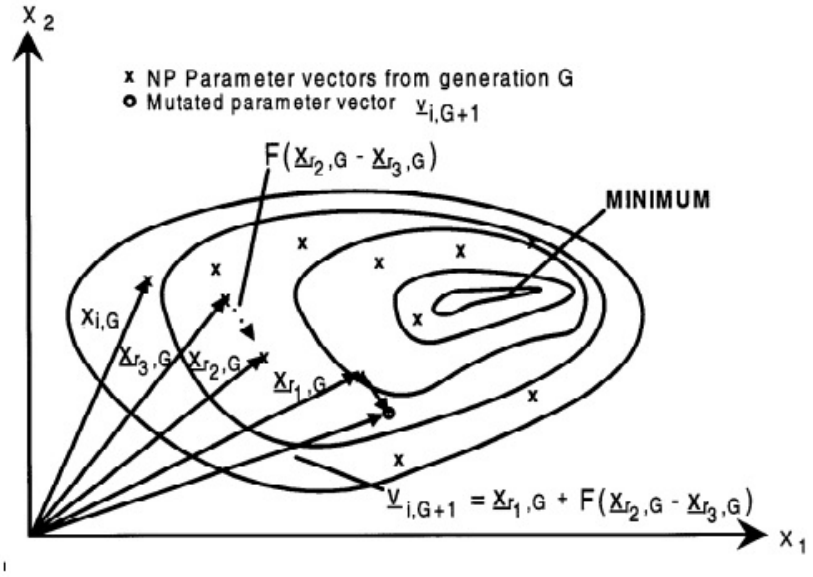

Figura 2.4. Exemplo de uma função bidimensional demonstrando as linhas de contorno e o processo de gerar o vetor doador. Figura do trabalho de Storn & Price [1997].

Em seguida é realizada a etapa de cruzamento para aumentar a diversidade. É nesta etapa que são gerados os vetores experimentais. Um vetor experimental $u_{i,G+1} = [u_{1i,G+1}, u_{2i,G+1}, ..., u_{Di,G+1}]$ é formado da seguinte forma

$$u_{ji,G+1} = \begin{cases} v_{ji,G+1}, & \text{se (randb(j)} \leq CR) \text{ ou } j = \text{rnbr(i)} \\ x_{ji,G}, & \text{se (randb(j)} > CR) \text{ ou } j \neq \text{rnbr(i)} \end{cases}$$

(2.29)

com $j = 1, 2, ..., D$. Na equação acima, $randb(j)$ corresponde à avaliação j de um gerador de número aleatório uniforme com saída $\in [0, 1]$, CR é a constante de cruzamento $\in [0, 1]$ que deve ser determinada pelo usuário, e $rnbr(i)$ é um índice escolhido de forma aleatória $\in 1, 2, ..., D$ que garante que o vetor $u_{i,G+1}$ receba pelo menos um

parâmetro de $v_{i,G+1}$. A Figura 2.5 ilustra o processo de cruzamento para vetores com 7 dimensões.

A última etapa consiste na seleção em que são escolhidos os membros da geração $G + 1$. O vetor experimental $u_{i,G+1}$ é comparado com o vetor alvo $x_{i,G}$ utilizando o critério guloso. Se o vetor $u_{i,G+1}$ produz um valor de função de custo menor que x_{iG}, então $x_{i,G+1}$ é escolhido para $u_{i,G+1}$, caso contrário o valor $x_{i,}G$ é mantido.

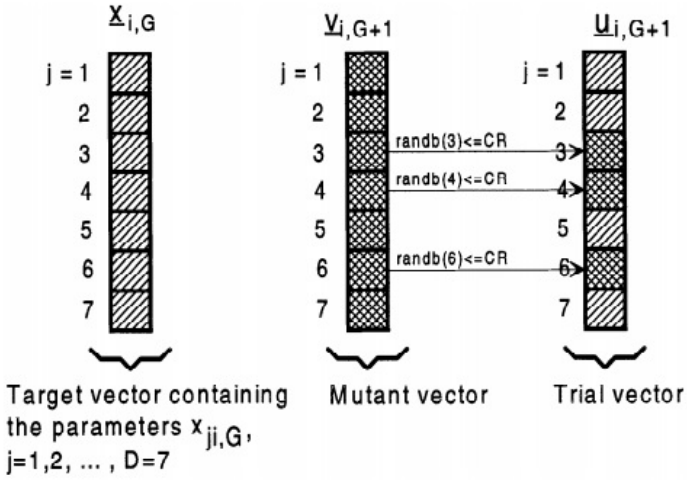

Figura 2.5. Processo de cruzamento dom D = 7. Figura do trabalho de Storn & Price [1997]

Neste trabalho o algoritmo DE é utilizado como método de otimização da medida de correlação de distância entre os atributos originais, com o objetivo de criar um novo atributo como entrada do modelo de predição.

Capítulo 3

Trabalhos relacionados

O estudo do comportamento do mercado financeiro é bem antigo. Bachelier [1900], foi um dos precursores da utilização de probabilidades no estudo das séries financeiras. O mesmo afirmou que variações no preço possuíam distribuições independentes e normais. Fama [1965] removeu a premissa de que as distribuições eram normais, e assumiu a hipótese de que as variações no preço eram independentes e possuíam a mesma distribuição. [Granger & Morgenstern, 1970] afirmaram que as variações no preço não seguiam a mesma distribuição. Ao longo da história, o entendimento do comportamento das séries de preço foi se modificando assim como as suposições da previsibilidade das mesmas.

Muitos trabalhos surgiram para testar a hipótese de que as séries financeiras seguiam um passeio aleatório. Working [1934] demonstrou que várias séries de preços futuros de commodities possuíam um comportamento muito parecido com uma série gerada artificialmente que seguia um passeio aleatório. Já Kendall & Hill [1953] analisaram os preços de trigo, algodão e índices, concluindo que os investidores deveriam assumir que os preços seguiam um passeio aleatório. E Fama [1965] analisou 30 ações do índice Dow Jones em detalhes, e concluiu que ou os preços seguiam um passeio aleatório ou algo muito similar.

Destacam-se na literatura diferentes testes estatísticos criados para testar a hipótese de passeio aleatório nas séries temporais. Taylor [1982] desenvolveu um teste de tendência de preços ao analisar as autocorrelações das séries. Já Fama & French [1988] demonstram que uma reversão nos preços induz em uma maior autocorrelação em *lags* futuros e esta ideia é explorada para a construção de um teste contra a hipótese de passeio aleatório. Jegadeesh [1991] desenvolveu um teste baseado na regressão da série em k períodos. Fama [1965] analisou as sequências de retornos positivos, neutros e negativos para construir o teste *Run*. Pode-se citar Praetz [1979] que elaborou um

teste estatístico baseado na função espectral de densidade das séries.

[Dryden, 1970] aplicou diversos testes estatísticos em 15 séries financeiras diárias do mercado do Reino Unido. O autor ainda sugere uma nova forma de teste baseado na autocorrelação da série para detectar dependência não linear. Cunningham [1973] afirmou que existe dependência em duas séries analisadas no mercado do Reino Unido e que a correlação em variações é positiva. Os padrões de subida e queda dos preços são explicados pelo autor por meio de probabilidades conjuntas diferentes de subida e queda. Jennergren & Korsvold [1974] utilizaram 45 ações entre os mercados de Oslo e Estocolmo para analisar a distribuição dos preços e testar a hipótese de passeio aleatório por meio de testes de correlação.

Demonstrar que as séries não seguem um passeio aleatório é uma evidência de que pode haver uma dependência temporal nas mesmas e justifica a utilização de alguma estratégia de predição. Um dos testes que se destacou neste contexto foi o da razão da variância, originalmente proposto por Lo & MacKinlay [1988]. Em seguida, outros testes baseados no mesmo conceito foram desenvolvidos, podendo citar os trabalhos de Chow [1993], Wright [2000], Whang [2003]. Vários trabalhos foram realizados buscando confrontar a hipótese de passeio aleatório nas séries financeiras por meio destes testes, como por exemplo Hoque et al. [2007] e Kim & Shamsuddin [2008] no mercado asiático.

Cont [2001] analisa em profundidade as propriedades em comum das séries financeiras, assim como são mencionados fatos estilizados que a grande maioria das séries possuem. [Taylor, 2007] aborda diversos aspectos da análise de séries financeiras diárias. É desenvolvida uma metodologia extensa analisando detalhadamente diversas características das séries, assim como o desenvolvimento dos testes de passeio aleatório.

Székely et al. [2007] apresentam uma nova medida, a correlação de distância, que é capaz medir todos os tipos de dependência entre vetores aleatórios em dimensões não necessariamente iguais. A medida varia entre o intervalo de 0 a 1 e é igual a 0 somente quando os vetores são independentes. Neste caso é uma medida mais geral que a correlação de Pearson. Os autores ainda apresentam um teste estatístico de independência ao se utilizar amostras independentes e identicamente distribuídas dos vetores. Huo & Székely [2016] abordam o fato de que a correlação de distância como foi definida possui a complexidade computacional de $O(n^2)$, sendo uma desvantagem comparada com outros métodos mais rápidos. Os mesmos desenvolvem uma metologia para o cálculo da medida utilizando um algoritmo com complexidade $O(nlogn)$.

Zhou [2012] estende o conceito de correlação de distância para séries temporais e apresenta o conceito de auto correlação de distância com a finalidade de medir a dependência temporal nas séries. O autor faz uma comparação com a função de autocorrelação demonstrando que a nova medida é capaz de capturar dependências em que

a função de autocorrelação não consegue. Também é apresentada uma metodologia de subamostragem da série para construção de um teste estatístico de independência, utilizando um intervalo de *lag* fixo. Fokianos & Pitsillou [2017] estudam o comportamento da função de auto correlação de distância em um número crescente de *lags* por meio de métodos que utilizam o domínio espectral das séries já estudados em Hong [1999]. Sendo assim, o trabalho de Fokianos & Pitsillou [2017] estende o trabalho de Zhou [2012] e desenvolve um teste estatístico que captura todas as dependências par a par das séries ao combinar os conceitos desenvolvidos em Hong [1999] e Székely et al. [2007].

Modelos complexos envolvendo diferentes abordagens para previsão de tendência de preços estão sendo utilizados em trabalhos recentes. Melo [2012] compara técnicas de inteligência artificial e aprendizado de máquina em séries temporais do mercado financeiro, como redes neurais, análise textual de notícias, inteligência coletiva, entre outras. O autor também faz uma análise sobre algumas pesquisas na área. [Allen & Karjalainen, 1999] utilizam Algoritmos Genéticos aplicados no mercado financeiro, enquanto que [j. Kim, 2003] utilizou Máquina de Vetor Suporte para predição. Pommeranzenbaum [2014] utiliza redes neurais aplicadas ao índice Ibovespa para predição de valores futuros e criação de diferentes estratégias financeiras.

Redes neurais recorrentes têm sido utilizadas em séries temporais financeiras. Pode-se citar [Kamijo & Tanigawa, 1990] que utilizaram os dados de preço e a também a utilização de indicadores técnicos em [Wang & Leu, 1996]. Bebarta et al. [2015] e Rather et al. [2015] utilizam redes neurais recorrentes para prever preços futuros no mercado indiano, obtendo em ambos os casos resultados melhores que os *baselines* escolhidos.

Há um considerável número de trabalhos recentes que utilizam a rede LSTM com o objetivo de prever tendências nos preços das séries financeiras. Chen et al. [2015] utilizaram 10 atributos de entrada na rede para previsão dos retornos do mercado chinês. A utilização da rede neural demonstrou aumentos na Acurácia de 14.3% a 27.2% ao se comparar com um previsor aleatório. Zhao et al. [2017] utilizaram uma função de peso temporal de acordo com a proximidade com o dado a ser previsto combinando com uma rede LSTM. Os mesmos afirmam ter alcançado uma Acurácia de 83.91% ao utilizar os dados do índice CSI 300. O índice CSI 300 também foi utilizado como base de dados em Yao et al. [2018] que aplicaram uma rede LSTM em dados de alta frequência e encontraram resultados melhores que previsores aleatórios.

Shao et al. [2017] utilizaram uma combinação de redes LSTM com o algoritmo *k-means*, encontrando resultados melhores que a rede individualmente. O algoritmo de agrupamento foi utilizado com o intuito de dividir a série temporal em sub-sequências e

utilizar essas novas sequências como entradas nas redes. Samarawickrama & Fernando [2017] compararam diferentes tipos de arquiteturas de redes neurais recorrentes utilizando dados históricos do mercado de Sri Lanka. Foram utilizados como entradas nas redes os valores do preço de fechamento, máximo e mínimo diários das séries escolhidas. Os autores afirmam que a rede LSTM foi uma das que obteve menor erro comparada às demais. Nelson et al. [2017] utilizaram uma combinação de rede LSTM juntamente com indicadores técnicos de mercado para realizar previsões na bolsa de valores B3. Os resultados encontrados foram promissores, com valores de Acurácia de até 55.9%

Huang et al. [2018] implementaram um modelo Bayesiano de rede LSTM utilizando seis indicadores do mercado chinês. Os resultados obtidos demonstraram que o modelo proposto aumentou em cerca de 25% os resultados da rede neural original. Lin & Chen [2018] utilizaram um algoritmo genético para otimizar os pesos internos da rede LSTM e demonstraram reduzir com a técnica o tempo de treinamento da rede. Pawar et al. [2019] compararam o desempenho de uma rede LSTM com diversos algoritmos de aprendizado (*Support Vector Machine, Random Forest*, regressão) e destacaram o desempenho melhor da rede LSTM em relação as outras técnicas utilizando séries financeiras. A comparação da rede LSTM com outras técnicas de aprendizado também é realizada em Karmiani et al. [2019] e os autores destacam os melhores resultados em termos de Acurácia para a rede neural LSTM.

A maioria dos trabalhos relacionados citados foca-se somente em um dos seguintes aspectos:

- Demonstrar as características estatísticas e fatos estilizados das séries,
- Testar a hipótese de passeio aleatório por meio de testes apropriados,
- Realizar previsões por meio de diferentes modelos,
- Simulação de estratégias de operação no mercado.

Como diferencial, este trabalho procura passar por todas estas etapas em conjunto por meio de experimentos com dados reais. Assume-se que para ser capaz de obter um modelo de previsão satisfatório no contexto de séries financeiras, é necessário entender e analisar cada uma destas etapas. Também destaca-se no trabalho a combinação de testes estatísticos e análises clássicas conhecidas na literatura, com a medida de correlação de distância, que até então foi pouco explorada neste contexto. Por fim, como principal diferencial, este trabalho ainda desenvolve uma metodologia nova para criação de um atributo de entrada no modelo de previsão mais correlacionado com

a série de valores futuros, ao se maximizar a correlação de distância dos atributos originais por meio de um algoritmo de otimização evolucionário.

Capítulo 4

Metodologia

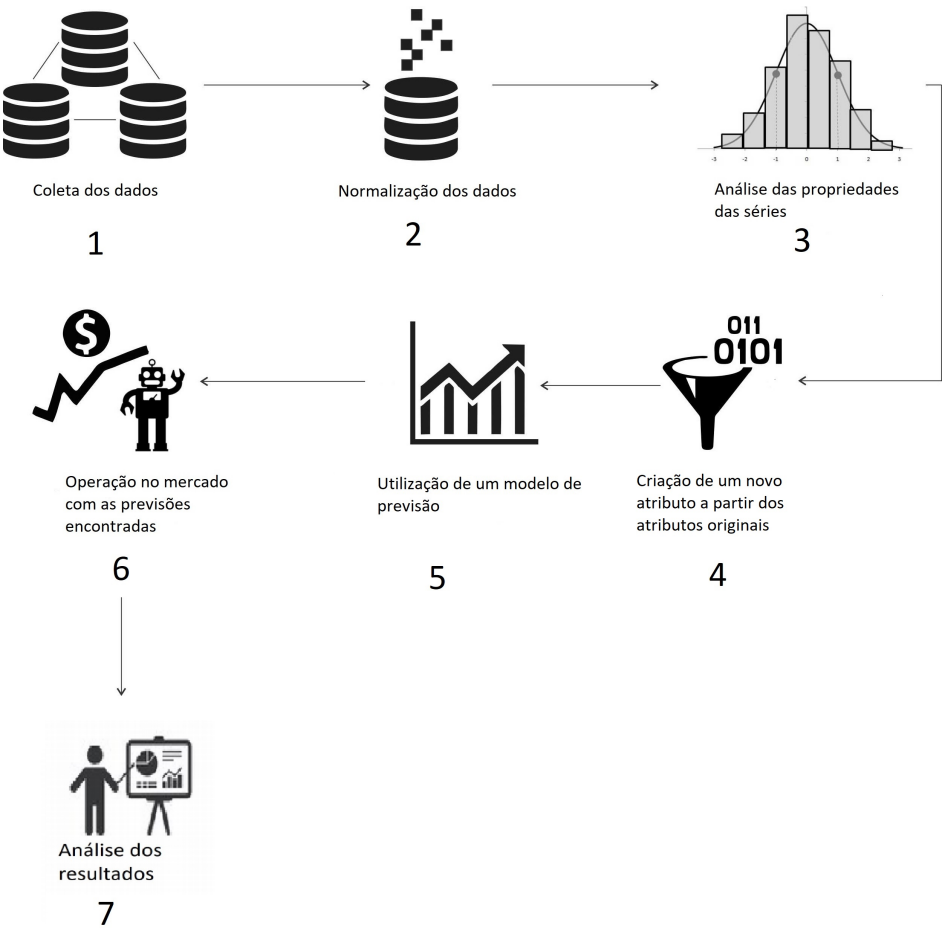

Figura 4.1. Etapas a serem seguidas na metodologia do trabalho

Este Capítulo descreve a metodologia utilizada neste trabalho. Os métodos propostos são validados em experimentos usando dados de diferentes ativos da bolsa de

valores de forma a averiguar se podem ser generalizados. São descritas a seguir as 7 etapas que consistem em: coleta e armazenamento dos dados, tratamento e normalização das séries de preço, análise das propriedades estatísticas e fatos estilizados, criação de um novo atributo mais correlacionado com os valores futuros das séries normalizadas, utilização de um modelo de previsão, implementação da estratégia financeira baseada nas previsões e análise dos resultados obtidos. A Figura 4.1 ilustra as etapas da metodologia do trabalho.

4.1 Coleta e armazenamento dos dados

Para este trabalho, são utilizados dados históricos diários refentes a bolsa de valores B3. Utiliza-se no trabalho as séries financeiras relativas ao preço de fechamento das ações. Essa coleta foi feita manualmente por meio da página Web da B3.

Os arquivos fornecidos pela B3 possuem para cada dia os seguintes campos: código do ativo, preço de abertura, preço de fechamento, preço máximo, preço mínimo, preço de fechamento anterior, quantidade de negócios, quantidade de papéis, volume financeiro e horário (hora e minutos) das transações realizadas. Neste sentido é necessário realizar uma filtragem pelos campos de código do ativo e preço de fechamento, e armazenamento dos mesmos.

4.2 Tratamento e normalização das séries de preços

Os dados coletados na etapa anterior são tratados buscando-se obter séries consistentes. É realizada uma verificação com a finalidade de remover valores nulos, incorretos ou inconsistentes encontrados.

Após os tratamentos nas séries, as mesmas são normalizadas. É utilizado o logaritmo do retorno financeiro ao invés dos preços, já que as séries do logaritmo do retorno financeiro são estacionárias em relação a média (mas não necessariamente em relação a variância, sendo um fato estilizado já mencionado), e possuem propriedades estatísticas mais fáceis de se analisar, ao contrário das séries de preços.

Seguindo a mesma metodologia de Taylor [2007], foi adotado o logaritmo do retorno financeiro podendo ser definido da seguinte forma

$$r_t = log(p_t) - log(p_{t-1}) \tag{4.1}$$

em que p_t é o preço no dia t. No restante do trabalho, as séries do logaritmo do retorno financeiro serão referenciadas apenas como séries de log-retorno financeiro,

para abreviação.

4.3 Análise das propriedades estatísticas e fatos estilizados

Nesta etapa são analisadas as propriedades estatísticas que as séries de log-retorno financeiro têm em comum, conhecidas na literatura como fatos estilizados. Também são analisadas algumas características das séries com o objetivo de identificar possíveis dependências temporais nas mesmas. É utilizada uma amostra grande de dados com o intuito de se obter melhores estimativas das propriedades estatísticas.

Essa primeira etapa é muito importante para o trabalho, uma vez que os resultados encontrados podem corroborar na utilização de modelos de previsão a partir dos dados históricos.

4.3.1 Momentos e distribuição das séries

São analisados os valores dos 4 primeiros momentos (média, desvio padrão, assimetria e curtoses) e também valores extremos de forma a tentar compreender melhor a distribuição das séries. As distribuições são comparadas com distribuições normais e é utilizado o teste Jarque Bera para testar a hipótese de que as séries de log-retornos financeiros seguem uma distribuição normal.

O teste de Jarque Bera utiliza os valores de assimetria e curtose das distribuições, e pode ser definido da seguinte forma

$$JB = n(\frac{S^2}{6} + \frac{(C-3)^2}{24}) \tag{4.2}$$

em que n corresponde ao número de observações, S corresponde ao valor da assimetria da amostra e C corresponde ao valor da curtose da amostra. Se as séries possuem distribuição normal, a estatística JB possui assintoticamente uma distribuição qui-quadrado com dois graus de liberdade. A hipótese nula do teste assume que a assimetria da amostra é igual a zero e que a curtose da amostra é igual a 3.

4.3.2 Análise das funções de autocorrelação (ACF) e auto correlação de distância (ACDF)

Outro ponto fundamental da caracterização das séries é a análise da ACF que possibilita a identificação de dependências lineares temporais. É realizada uma comparação entre

os valores da ACF para a série de log-retornos financeiros, log-retornos financeiros absolutos e log-retornos financeiros ao quadrado. Essa comparação pode evidenciar a presença de dependências não lineares.

Por meio dos valores da função ACF é calculada a estatística Q de Ljung-Box para testar a hipótese de que as séries de log-retornos financeiros são independentes e identicamente distribuídas (i.i.d). A estatística Q utiliza um conjunto de valores de autocorrelações da série e pode ser definida da seguinte forma

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k} \tag{4.3}$$

em que n corresponde ao tamanho da amostra, $\hat{\rho}_k$ corresponde ao valor da autocorrelação no *lag* k, e h corresponde ao número de *lags* a serem testados. A hipótese nula i.i.d. da estatística Q segue uma distribuição chi-quadrado assintoticamente.

Para testar a hipótese de passeio aleatório é utilizado o teste da razão da variância abordado no Capítulo da Fundamentação Teórica.

Em seguida são analisados os valores da ADCF buscando identificar dependências não lineares que a função ACF não é capaz de detectar. Valores significativos da função indicam que os *lags* não são independentes. São analisados os valores da ADCF também para as séries de log-retornos financeiros elevadas ao quadrado e sobre o valor absoluto, a fim de comparação. A hipótese i.i.d. das séries novamente é testada utilizando o teste da ADCF de correlação também abordado no Capítulo da Fundamentação Teórica.

4.4 Explorando a correlação não linear: criação de um atributo mais correlacionado por meio dos atributos originais

Nesta etapa é proposto um método para explorar a dependência não linear nas séries ao se criar um novo atributo como entrada do modelo de predição, a partir dos atributos originais. O método se baseia na técnica de filtragem da seleção de atributos, em que os atributos selecionados para um algoritmo de aprendizado de máquina são aqueles que exibem valores mais significativos de alguma métrica estatística (correlação, entropia, informação mútua, etc) com a classe alvo, ao analisar o conjunto de treino. No trabalho é utilizada como métrica a correlação de distância e os atributos originais do modelo são combinados buscando maximizar esta medida com a série de valores futuros do conjunto de treinamento.

Partindo-se da hipótese de que é possível haver uma dependência temporal nas séries, os atributos originais de entrada do modelo de predição são definidos como sendo iguais as séries atrasadas em 1, 2, ..., m *lags*, da série alvo. Utilizando o conjunto de treinamento, é calculada então uma combinação linear destas m séries atrasadas, que busca maximizar a correlação de distância com a série de valores futuros, criando-se dessa forma o novo atributo (uma nova série temporal a partir das séries atrasadas). Pode-se modelar então um problema de otimização irrestrito, com as variáveis de decisão sendo os pesos ws, e a função objetivo igual à correlação de distância entre a combinação linear dos atributos de entrada com a série de valores futuros (do conjunto de treinamento).

Seja a série de valores futuros (alvos) do conjunto de treinamento $X = x_1, x_2, ..., x_n$. E sejam as entradas do modelo de predição as séries atrasadas $X_{-1} = x_{1-1}, x_{2-1}, ..., x_{n-1}, X_{-2} = x_{1-2}, x_{2-2}, ..., x_{n-2}, ..., X_{-m} = x_{1-m}, x_{2-m}, ..., x_{n-m}$. Sejam $w_1, w_2, ..., w_m$ os números reais que correspondem aos pesos associados a cada uma das entradas. Sendo $S = (w_1 * X_{-1} + w_2 * X_{-2} + ... + w_{m*}X_{-m})$ a combinação linear dos atributos originais, o problema de otimização pode ser definido da seguinte forma

$$Max \frac{dCov(X, S)}{\sqrt{dVar(X)dVar(S)}} \tag{4.4}$$

em que $dCov(X, S)$ é igual ao valor da distância de covariância entre X e S, e $dVar(X)$, $dVar(S)$ é igual ao valor da distância de variância de X e de S respectivamente.

Para resolver o problema de otimização é utilizado o algoritmo *differential evolution*.

Espera-se obter resultados melhores pelo modelo de classificação ao substituir os atributos originais $(X_{-1}, X_{-2}, ..., X_{-m})$ pelo novo atributo (S), caso a combinação linear no conjunto de treinamento obtenha valores de correlação de distância mais significativos com a série alvo, do que os atributos originais. O princípio é o mesmo da técnica de filtragem em seleção de atributos. O novo atributo S substitui as entradas do modelo inclusive para o conjunto de teste ao se manter os mesmos valores dos pesos encontrados.

Para exemplificar a criação do novo atributo S, a seguir é demonstrado um exemplo bem simples em duas dimensões. São utilizados 11 valores exemplos de log-retorno financeiro da ação ABEV3, correspondentes ao ano de 2008. Estes valores correspondem à série alvo do exemplo (X), e são utilizadas como atributos originais as duas séries atrasadas em 1 e 2 *lags* respectivamente. Em seguida é calculada a combinação linear dos atributos originais buscando encontrar os valores dos pesos que maximizam correlação de distância entre a série X e o novo atributo S.

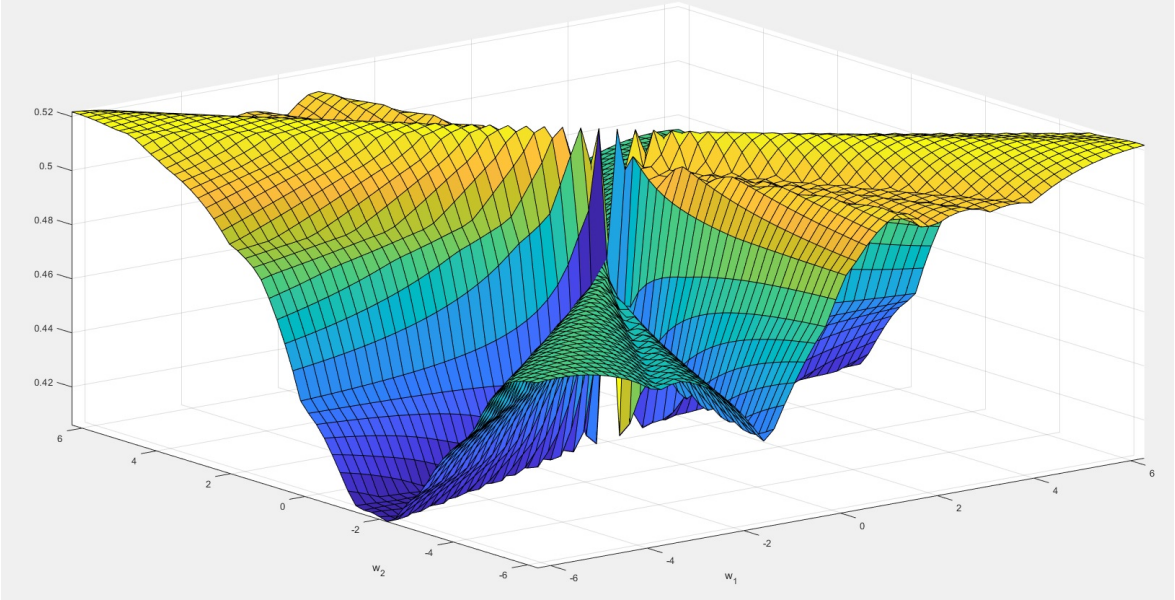

Figura 4.2. Valores da função da correlação de distância da combinação linear ao se variar w_1 e w_2

- Valores da série alvo exemplo (X): [0.008163311, -0.016393810, 0.005494519, 0.040273899, -0.015915455, -0.008053735, -0.027324104, -0.016760169, -0.034387342, 0.020202707, -0.049790665];

- Valores atrasados em 1 intervalo de tempo da série exemplo (atributo original): [-0.026955810, 0.008163311, -0.016393810, 0.005494519, 0.040273899, -0.015915455, -0.008053735, -0.027324104, -0.016760169, -0.034387342, 0.020202707];

- Valores atrasados em 2 intervalos de tempo da série exemplo (atributo original): [0.021506205, -0.026955810, 0.008163311, -0.016393810, 0.005494519, 0.040273899, -0.015915455, -0.008053735, -0.027324104, -0.016760169, -0.034387342];

- Novo atributo criado (S) = [-0.026955810*w1 + 0.021506205*w2, 0.008163311*w1 - 0.026955810*w2, -0.016393810*w1 - 0.008163311*w2, 0.005494519*w1 + -0.016393810*w2, 0.040273899*w1 – 0.005494519*w2, -0.015915455*w1 + 0.040273899*w2, -0.008053735*w1 + -0.015915455*w2, -0.027324104*w1 -0.008053735*w2, -0.016760169*w1 -0.027324104*w2, -0.034387342*w1 - 0.016760169*w2, -0.034387342*w1 - 0.034387342*w2];

A Figura 4.2 ilustra os valores da função de correlação de distância ao se variar os pesos w_1 e w_2. Ao se observar a Figura pode-se notar que o valor máximo da função de correlação de distância é aproximadamente 5.21. Ao se aplicar o algoritmo *differential evolution* foi encontrado exatamente o valor de 0.521 para $w_1 = -0.80002321$ e $w_2 = 0.79164731$. Neste exemplo o valor da correlação de distância entre X e X_{-1} é igual 0.438, e entre X e X_{-2} é igual a 0.49, demonstrando que foi possível aumentar a correlação de distância com a série alvo, ao utilizar o novo atributo S.

4.5 Modelo de previsão aplicado às séries financeiras

Essa etapa é responsável por criar um modelo de predição de valores futuros das séries que serve de base para a implementação das estratégias financeiras. Neste trabalho é desenvolvida uma metodologia que utiliza um algoritmo de aprendizado de máquina supervisionado como modelo de previsão. São utilizadas cinco etapas para a implementação do algoritmo:

- Modelagem em um problema de classificação binário: para cada dia futuro o algoritmo procura classificar aquele dia como uma classe de Altas ou uma classe de Baixas. A classe de Altas (tendência de alta nos preços) correspondem os valores de log-retorno financeiro positivos e a classe de Baixas (tendência de queda nos preços) correspondem os valores de log-retorno financeiro negativos. Essa classificação é realizada baseada no aprendizado obtido na etapa de treinamento por meio das classificações passadas.

- Definição do conjunto de treinamento e conjunto de teste: é necessário obter um conjunto grande o suficiente para o treinamento do algoritmo, assim como um conjunto de teste utilizado para avaliar as métricas de classificação do modelo de predição. No trabalho é adotado um total de 250 dias (correspondendo a 1 ano de dados), sendo 166 dias utilizados para o treinamento do modelo e os 84 dias restantes utilizados para o teste.

- Definição dos atributos de entrada: correspondem aos dados de entrada do modelo que são utilizados na etapa de aprendizagem do algoritmo. São utilizadas duas configurações distintas como entrada do modelo de predição. Na primeira configuração os atributos de entrada correspondem as séries atrasadas da série alvo futura. Já na segunda configuração, os atributos correspondem as combinações lineares das séries atrasadas, conforme a Seção anterior (4.4) da metodologia.

Os resultados para ambas as configurações são comparados entre si a fim de comparação.

- Configuração dos parâmetros do algoritmo: escolha dos parâmetros utilizados para o algoritmo de classificação.
- Execução do algoritmo: número de vezes que o algoritmo é executado. Alguns algoritmos de aprendizado inicialmente atribuem pesos aleatórios ao modelo (por exemplo redes neurais) e dessa forma é ideal realizar várias execuções para se obter valores mais precisos do desempenho do algoritmo.

4.6 Estratégia de operação

Os resultados do modelo de predição fornecem uma classificação para cada dia futuro do conjunto de teste do algoritmo de aprendizado de máquina. Como são utilizados os valores relativos ao preço de fechamento, pode-se implementar uma estratégia que toma a decisão de comprar ou vender um ativo a partir da classificação do dia futuro, minutos antes do preço de fechamento do dia atual. Neste caso, assume-se que o valor do preço de fechamento do dia corrente será praticamente o mesmo do momento em que a decisão é tomada. Sendo assim, a estratégia de operação no mercado adotada no trabalho é definida na seguinte forma:

- classe de Altas: sendo uma previsão de alta, é emitida uma ordem de compra ao final do dia atual t, e é realizada uma ordem de venda ao final do próximo dia $t + 1$,
- classe de Baixas: sendo uma previsão de queda, é emitida uma ordem de venda ao final do dia atual t, e é realizada uma ordem de compra ao final do próximo dia $t + 1$. Nota-se que neste caso é realizada a operação de venda a descoberto, uma vez que o investidor não possuía a ação previamente.

Vale ressaltar que ao utilizar a estratégia acima, há uma negociação para cada dia futuro previsto pelo algoritmo de aprendizado na etapa anterior.

4.6.1 Combinação da estratégia de operação com *Stop Loss*

A estratégia de operação também é implementada em conjunto com *Stop Loss*, um valor limite que represente o máximo que o investidor esteja disposto a perder. Caso o valor da ação ultrapasse este limite a operação é realizada imediatamente com o objetivo de diminuir as perdas.

4.7 Análise dos resultados

Nesta etapa são analisados os resultados do desempenho do algoritmo de classificação e dos retornos financeiros ao se aplicar as estratégias de operação. Em termos de classificação, os resultados são comparados entrei si e também com um previsor aleatório. Já em termos de retornos financeiros, os resultados novamente são comparados entre si, com um previsor aleatório, com a estratégia de negociação *Buy and Hold*, e rendimentos da taxa SELIC e CDI. Por fim são analisados os custos financeiros de se operar no mercado em comparação com os retornos financeiros obtidos ao se aplicar as estratégias propostas. A seguir são descritas estas etapas.

4.7.1 Análise do desempenho de classificação

Para avaliação do algoritmo de classificação foram utilizadas as seguintes métricas: Precisão, Revocação, medida F1 e Acurácia. A Precisão é calculada como o número de acertos da classe positiva (TP) sobre o número total de previsões desta classe, ou seja, TP mais as falsas positivas (FP)

$$P = \frac{TP}{TP + FP}. \tag{4.5}$$

A Revocação é calculada pelo número de classificações corretas da classe positiva sobre o número total de ocorrências desta classe, ou seja, TP mais as falsas negativas (FN)

$$R = \frac{TP}{TP + FN}. \tag{4.6}$$

Em seguida pode-se definir a medida F1 como sendo igual a média harmônica entre Precisão e Revocação, sendo muito útil em casos de problemas de enviesamento de uma determinada classe

$$F1 = 2 * \frac{P * R}{P + R}. \tag{4.7}$$

Apesar das medidas acima serem definidas ao se selecionar a classe positiva, os mesmos cálculos foram realizados para a classe negativa. Por fim, pode-se calcular como medida de desempenho do algoritmo de classificação a Acurácia que é definida como o número classificações corretas sobre o número total de previsões

$$A = \frac{TP + TN}{TP + FP + TN + FN}, \tag{4.8}$$

sendo TN igual ao número de classificações corretas da classe negativa.

Os valores da Acurácia obtidos utilizando os atributos originais e o novo atributo são comparados entre si e também com um previsor aleatório. Para cada dia futuro este previsor atribui aleatoriamente ou a classe 1 ou a classe 0 com igual probabilidade para ambas de 50%. Para validação estatística dos resultados é calculado o intervalo de confiança da medida de Acurácia para o previsor aleatório. Seguindo a formulação de Levine et al. [2008]

$$p - Z\sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + Z\sqrt{\frac{p(1-p)}{n}} \tag{4.9}$$

em que

- p = proporção da amostra (Acurácia),
- π = proporção da população,
- Z = valor crítico da distribuição normal padronizada,
- n = tamanho da amostra.

São calculados diferentes níveis de confiança e comparado o número de ações que obtiveram Acurácia superior ao intervalo para as diferentes entradas do modelo de previsão.

4.7.2 Análise do resultado financeiro

Após obter as previsões dos dias futuros, são aplicadas as estratégias de operação. Nesta etapa é avaliado o retorno financeiro em percentual. Para comparação, são utilizados como *baselines* os retornos financeiros de um classificador aleatório e com os valores da estratégia de *Buy and Hold* em relação ao período de teste. A estratégia de *Buy and Hold* consiste em comprar o ativo e aguardar por um longo período independente das flutuações do mercado. Também são utilizados a fim de comparação os valores da porcentagem da taxa SELIC e CDI em relação ao período de teste.

Para validação estatística é calculado o teste Z de diferença de proporções entre as classificações do algoritmo de aprendizado e do classificador aleatório. A proporção é calculada em relação ao número de vezes que os modelos de predição superaram os *baselines Buy and Hold*, SELIC e CDI. De acordo com Levine et al. [2008] a estatística

Z segue aproximadamente uma distribuição normal padronizada e pode ser definida da seguinte forma

$$Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\bar{p}(1-\bar{p})(\frac{1}{n_1} + \frac{1}{n_2})}} \tag{4.10}$$

e

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2}, p_1 = \frac{X_1}{n_1}, p_2 = \frac{X_2}{n_2} \tag{4.11}$$

em que

- p_1 = proporção de sucessos na amostra 1,
- X_1 = número de sucessos na amostra 1,
- n_1 = tamanho da amostra 1,
- π_1 = proporção de sucessos na população 1,
- p_2 = proporção de sucessos na amostra 2,
- X_2 = número de sucessos na amostra 2,
- n_2 = tamanho da amostra 2,
- π_2 = proporção de sucessos na população 2,
- $\bar{p}$ = estimativa agrupada para a proporção de sucessos na população,

É feita a comparação entre os modelos de predição com o previsor aleatório.

4.7.3 Análise dos custos de operação

Mesmo obtendo um retorno financeiro positivo é necessário avaliar o custo operacional de cada negociação, uma vez que a estratégia adotada no trabalho utiliza operações diárias. Esta análise se faz necessária já que os custos de operação podem ultrapassar os lucros obtidos pelos investidores dependendo do tamanho do lote negociado.

Para cada operação são cobradas 4 taxas: Imposto de Renda recolhido na fonte pela B3 (0.5%), Taxa de Corretagem sobre as operações (R$2.50), Imposto Sobre Serviços em relação a taxa de corretagem (12%) e os Emolumentos sobre o valor investido (0.025%). Caso o investidor tenha um lucro na negociação, também é cobrado o Imposto de Renda recolhido pelo investidor sobre o lucro (20%). Nota-se que o percentual

correspondente a soma de todos os custos de operação pode ser bem alto em relação ao lucro bruto obtido pelo investidor.

No trabalho é avaliada a relação dos custos de operações com o número de lotes necessários para se obter lucro líquido pelo investidor. São comparadas as simulações das operações da estratégia financeira com os resultados dos retornos financeiros dos *baselines*, incluindo os custos de operação. Nesta etapa os resultados são apresentados em valores de Reais (R$).

Capítulo 5

Resultados

Neste Capítulo foram instanciadas todas as etapas da metodologia, conforme descrita no Capítulo 4. Foram executados experimentos correspondentes a cada etapa e os resultados encontrados foram analisados buscando uma validação robusta por meio de testes estatísticos e comparação com *baselines* da literatura.

5.1 Coleta e armazenamento dos dados

Foram coletados dados históricos diários reais refentes à bolsa de valores B3 no período de Janeiro de 2008 a Março de 2017. Foram utilizadas as séries de preço de fechamento de 38 ações da bolsa de valores.

Também foram utilizados os dados históricos intra-diários com periodicidade de 15 minutos, referentes ao último quadrimestre de 2016 na etapa da estratégia de operação com utilização de *Stop Loss*.

5.2 Tratamento e normalização das séries preços

Todas as séries correspondentes aos 38 ativos da B3 foram tratadas e normalizadas, obtendo-se ao final as séries consistentes relativas aos log-retornos financeiros.

As Figuras 5.1 e 5.2 a seguir ilustram o comportamento do preço de fechamento e do log-retorno financeiro da ação BBAS3 entre o período de 2010 a 2016, como exemplo. Pode-se notar a diferença no comportamento das curvas e na escala dos gráficos para os valores de preço e do log-retorno financeiro.

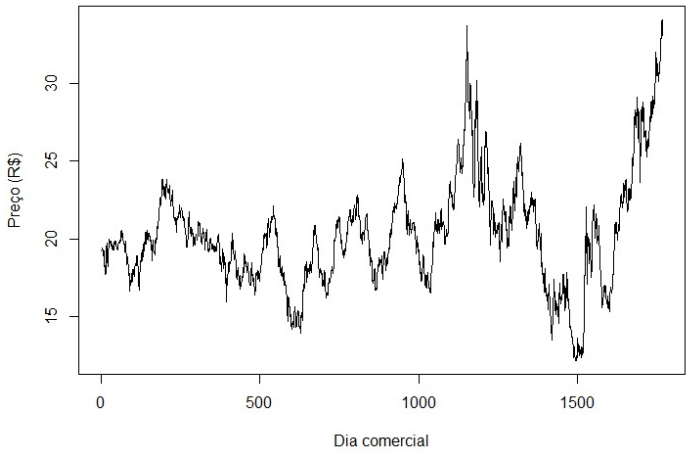

Figura 5.1. Valores de preço de fechamento da ação BBAS3 entre 2010 e 2016

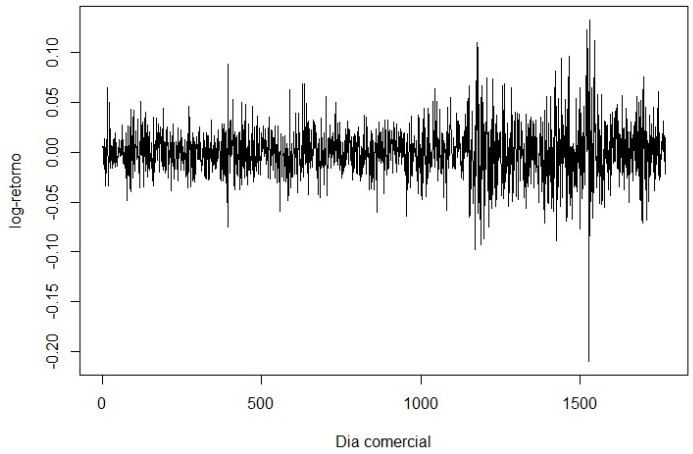

Figura 5.2. Valores de log-retorno da ação BBAS3 entre 2010 e 2016

5.3 Análise das propriedades estatísticas e fatos estilizados

Foram analisadas as propriedades estatísticas do período de Janeiro de 2008 até Março de 2017 (total de 2267 dias). As Seções a seguir demonstram cada uma das análises realizadas. Foi utilizado o software R juntamente com os pacotes *moments*, *vrtest*, *energy* e *dCovTS*.

Tabela 5.1. Estimação dos quatro primeiros momentos e p-valor para o teste Jarque-Bera

Ações	Média	Desvio	Curtose	Assimetria	p-valor
ABEV3	6.8E-4	0.017	9.28	0.07	2.20E-16
BBAS3	2.7E-4	0.028	7.97	0.11	2.20E-16
BBDC4	3.4E-4	0.023	9.15	0.54	2.20E-16
BRAP4	-1.6E-4	0.029	6.26	-5.1E-05	2.20E-16
BRFS3	3.1E-4	0.021	8.27	0.15	2.20E-16
BRKM5	4.9E-4	0.028	7.89	0.20	2.20E-16
BRML3	2.6E-4	0.028	10.45	0.41	2.20E-16
CCRO3	5.6E-4	0.023	7.19	0.02	2.20E-16
CMIG4	3.2E-4	0.025	11.84	-0.64	2.20E-16
CPFE3	4.0E-4	0.019	5.90	-0.02	2.20E-16
CPLE6	2.2E-4	0.023	7.22	-0.10	2.20E-16
CSAN3	4.0E-4	0.026	7.73	-0.03	2.20E-16
CSNA3	-1.6E-4	0.036	7.18	0.24	2.20E-16
CYRE3	-1.7E-4	0.031	7.18	0.24	2.20E-16
EGIE3	4.6E-4	0.019	8.12	0.26	2.20E-16
ELET3	2.3E-4	0.028	6.52	0.22	2.20E-16
EMBR3	-6.64E-6	0.024	7.11	-0.10	2.20E-16
ENBR3	3.9E-4	0.021	5.26	-0.10	2.20E-16
FIBR3	-2.8E-4	0.029	6.21	0.05	2.20E-16
GGBR4	-2.2E-4	0.030	5.85	0.13	2.20E-16
GOAU4	-7.0E-4	0.032	6.26	0.08	2.20E-16
ITSA4	4.1E-4	0.023	9.69	0.41	2.20E-16
ITUB4	3.9E-4	0.024	9.26	0.58	2.20E-16
JBSS3	3.3E-4	0.033	7.43	0.20	2.20E-16
LAME4	3.6E-4	0.027	9.30	0.33	2.20E-16
LREN3	6.6E-4	0.026	6.68	0.13	2.20E-16
MRFG3	-3.7E-4	0.031	9.36	-0.41	2.20E-16
MULT3	5.7E-4	0.022	15.80	0.77	2.20E-16
NATU3	3.2E-4	0.022	4.59	0.16	2.20E-16
PCAR4	2.8E-4	0.021	6.53	0.23	2.20E-16
PETR3	-4.3E-4	0.030	5.83	0.16	2.20E-16
PETR4	-3.5E-4	0.029	5.92	0.03	2.20E-16
SUZB5	-1.6E-4	0.025	5.47	0.12	2.20E-16
TIMP3	1.2E-4	0.027	13.19	0.19	2.20E-16
USIM5	-6.9E-4	0.036	7.84	0.56	2.20E-16
VALE3	-1.1E-4	0.029	6.72	-0.03	2.20E-16
VIVT4	2.4E-4	0.017	5.19	-0.01	2.20E-16
WEGE3	3.1E-4	0.021	11.49	-0.38	2.20E-16

Tabela 5.2. Frequência e magnitude dos *outliers* das distribuições. Na Tabela são demonstrados os números de observações encontradas ao se variar o valor de k em $|r_t - \bar{r}| > ks$. Os parâmetros $r_t, \bar{r}, s, k$ representam respectivamente cada valor da distribuição, a média amostral, desvio padrão amostral e um inteiro positivo.

Ações	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9
ABEV3	112	24	6	6	4	2	1	0
BBAS3	116	30	9	5	2	2	0	0
BBDC4	115	31	13	4	1	1	1	1
BRAP4	121	32	8	3	1	1	0	0
BRFS3	120	37	8	3	1	1	1	0
BRKM5	110	25	11	3	3	1	0	0
BRML3	93	26	13	10	7	2	0	0
CCRO3	109	25	9	5	3	0	0	0
CMIG4	111	30	10	5	3	2	2	1
CPFE3	117	26	6	2	1	1	0	0
CPLE6	110	33	9	5	2	1	0	0
CSAN3	109	37	12	4	1	1	0	0
CSNA3	129	32	17	4	1	0	0	0
CYRE3	116	29	12	5	2	1	1	1
EGIE3	113	30	9	1	1	1	1	1
ELET3	117	37	12	3	1	0	0	0
EMBR3	110	36	15	5	2	0	0	0
ENBR3	110	26	3	2	1	0	0	0
FIBR3	112	38	11	3	0	0	0	0
GGBR4	117	31	9	3	0	0	0	0
GOAU4	122	29	13	3	0	0	0	0
ITSA4	102	35	13	6	2	1	1	1
ITUB4	102	34	11	5	1	1	1	1
JBSS3	113	30	13	4	2	1	0	0
LAME4	102	28	11	8	2	1	1	0
LREN3	120	33	11	4	1	0	0	0
MRFG3	116	25	12	3	1	1	1	1
MULT3	103	27	10	6	2	2	1	1
NATU3	109	20	7	1	0	0	0	0
PCAR4	108	28	7	2	1	1	0	0
PETR3	133	41	10	0	0	0	0	0
PETR4	121	42	13	1	0	0	0	0
SUZB5	123	23	9	2	0	0	0	0
TIMP3	98	31	10	7	4	3	3	0
USIM5	114	34	9	3	1	1	1	0
VALE3	131	32	8	2	1	1	0	0
VIVT4	113	28	5	1	0	0	0	0
WEGE3	96	25	11	5	3	2	2	0

Tabela 5.3. Comparação da distribuição da ação BBAS3 com uma distribuição normal com mesma média e desvio padrão. Na Tabela são demonstrados os números de observações encontrados ao se variar o valor de k em $|r_t - \bar{r}| > ks$

	$\mathbf{k \leq 0.25}$	$\mathbf{0.25 < k \leq 0.5}$	$\mathbf{0.5 < k \leq 1}$	$\mathbf{1 < k \leq 2}$	$\mathbf{2 < k \leq 3}$	$\mathbf{k > 3}$
BBAS3	586	476	669	420	86	30
Normal	453	437	657	616	99	5
Diferença	133	39	12	-196	-13	25

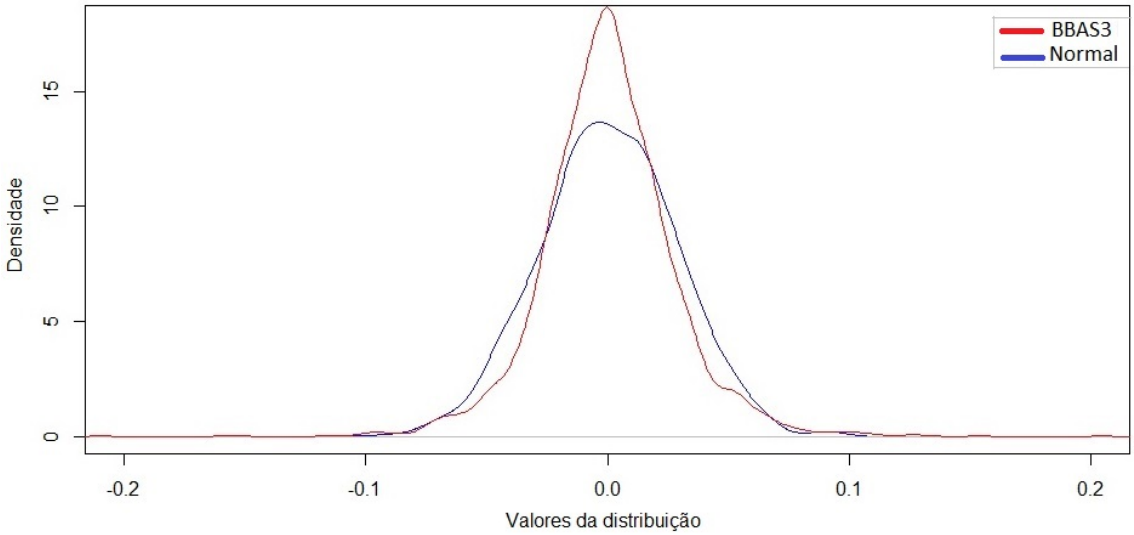

Figura 5.3. Comparação da densidades da distribuição da ação BBAS3 com uma distribuição normal de mesma média e desvio da ação

5.3.1 Momentos e distribuição das séries

Foram calculados os valores para os primeiros quatro momentos das distribuições das séries de log-retornos financeiros: média, desvio padrão, assimetria e curtose, assim como os p-valores para o teste Jaque-Bera com o intuito de verificar se as distribuições se aproximam de uma distribuição normal.

A Tabela 5.1 demonstra os valores estimados para as ações. Nota-se que os valores de média são aproximadamente zero. A estimativa da assimetria demonstra que para quase todas as ações seu valor se aproxima de zero, o que indica que as distribuições tendem a ser simétricas. Na Tabela se destaca os valores encontrados para a curtose das ações. Distribuições normais possuem curtose igual a 3. Os valores encontrados são bem maiores que 3, sendo que todas as ações possuem curtose maior que 4.5 e a maioria possui valores maior que 6. Os p-valores obtidos rejeitam a hipótese de distribuições normais. Em nenhum caso foi obtido um p-valor maior que 0.05 ou 0.01. Fica evidente

que as distribuições não se aproximam de uma distribuição normal. Ao se analisar os resultados pode-se dizer que a distribuição é simétrica pelos valores da assimetria, a distribuição possui caudas pesadas pelos altos valores de curtose e que a distribuição possui picos altos, maiores do que o esperado de uma distribuição normal.

Altos valores para curtose são causados por muitas observações que estão mais longe de desvios padrões do que os previstos para uma distribuição normal. Na Tabela 5.2 são demonstrados os resultados da frequência e magnitude dos *outliers* encontrados para cada ação. A frequência do evento é calculada pela seguinte fórmula: $|x_t - \bar{x}| > ks$, em que s corresponde ao desvio padrão e k é um inteiro positivo. Os resultados encontrados demonstram que a frequência de *outliers* é muito maior do que a esperada em uma distribuição normal, havendo casos de observações distantes em mais de 9 desvios padrões (CMIG4, CYRE3, EGIE3, por exemplo). A regra empírica 3-sigma afirma que a probabilidade de se encontrar observações maiores que 3 desvios padrões é aproximadamente igual a 0.3% para uma distribuição normal . Das 2267 amostras de cada ação, essa probabilidade corresponde a aproximadamente entre 6 e 7 observações. Entretanto se observa que na maioria dos casos as ações apresentaram um número de observações muito maior, chegando em alguns casos em 30 observações, o que corresponderia a 1.32%.

A Tabela 5.3 demonstra como exemplo a comparação da distribuição da ação BBAS3 com uma distribuição normal com mesma média e desvio padrão da ação. Pode-se perceber que existem mais observações no intervalo entre, a média menos um desvio padrão até a média mais um desvio padrão, do que o esperado em uma distribuição normal, correspondendo ao pico alto na distribuição. O último intervalo também demonstra que há mais observações extremas correspondendo às duas caudas pesadas. Os altos valores da curtose são causados pelos *outliers* das caudas. Vale ressaltar que ambos fatores estão relacionados. Os valores extremos contribuem com a variância da distribuição, o que implica que deve haver mais observações perto do centro da distribuição do que é encontrado em uma normal com a mesma média e variância. A Figura 5.3 demonstra a comparação das funções de densidade da distribuição real da série de log-retorno financeiro da ação BBAS3, juntamente com a função de densidade de uma normal com mesma média e desvio padrão. Percebe-se exatamente o pico maior da distribuição dos dados em comparação com a normal, assim como as caudas pesadas.

Estes resultados corroboram o fato estilizado de que a distribuição dos log-retornos financeiros diários não é normal, é simétrica, possui caudas pesadas, apresenta um pico maior do que uma normal e exibe um formato de sino.

5.3.2 Análise das funções de autocorrelação (ACF) e auto correlação de distância (ADCF)

A segunda característica analisada é a autocorrelação. A amostra de autocorrelação dos log-retornos financeiros é geralmente próxima de zero independente do intervalo de *lags*. A Tabela 5.4 demonstra as estimativas de autocorrelação entre os *lags* de 1 a 30. Para resumir seus sinais e magnitudes os resultados foram atribuídos a um dos 6 seguintes intervalos, sendo $\hat{p}$ o valor da função de autocorrelação:

- intervalo 1 : $\hat{p} < -0.1$;
- intervalo 2 : $-0.1 \leq \hat{p} < -0.05$;
- intervalo 3 : $-0.05 \leq \hat{p} < 0$;
- intervalo 4 : $0 \leq \hat{p} < 0.05$;
- intervalo 5 : $0.05 \leq \hat{p} < 0.1$;
- intervalo 6 :$0.1 < \hat{p}$

As estimativas das autocorrelações demonstram ser pequenas. Os valores críticos correspondem a 0.0411 e -0.0411. Das 38 ações, 27 possuem pelo menos 90% das estimativas no intervalo de $[-0.05, 0.05]$, sendo que na maioria dos casos os valores não são significativos. O número total de autocorrelações significativas é baixo, sugerindo que qualquer dependência linear nos log-retornos financeiros diários deve ser considerada pequena, ou ainda o caso dos valores significativos serem um erro de amostra. Pode-se notar também um número similar entre autocorrelações positivas e negativas.

Entretanto, ao se analisar os log-retornos financeiros elevados ao quadrado e em valores absolutos, esse comportamento se altera. Pode-se notar que a maioria das autocorrelações estimadas são significativas, indicando uma dependência linear nas séries transformadas e consequentemente uma dependência não linear nas séries de log-retornos financeiros. As Tabelas 5.5 e 5.6 demonstram estes resultados. Percebe-se que em ambas as Tabelas, em quase todas as ações a maioria dos valores de autocorrelação é maior que 0.1 e praticamente não há correlações negativas. Este resultado demonstra um fato estilizado já mencionado e que provavelmente é causado pelos agrupamentos de períodos de alta volatilidade, em que dias consecutivos possuem variações maiores nos preços.

As estimativas das autocorrelações dos log-retornos financeiros e suas transformações podem ser utilizadas para testar a hipótese das séries serem independentes e

Tabela 5.4. Valores dos 5 primeiros *lags* da função de autocorrelação dos log-retornos financeiros. Na Tabela também são demonstrados os números totais de observações encontradas nos intervalos definidos de 1 a 6 para os 30 primeiros *lags* das séries

	Lags					**Intervalos**					
Ações	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**0**
ABEV3	-0.089	-0.067	-0.058	0.013	0.046	0	5	13	11	1	0
BBAS3	0.013	-0.010	-0.042	-0.023	0.001	0	0	15	15	0	0
BBDC4	-0.045	-0.022	-0.082	0.001	-0.022	0	2	12	16	0	0
BRAP4	-0.005	-0.017	-0.040	-0.014	0.043	0	0	15	14	1	0
BRFS3	-0.032	-0.050	-0.027	0.005	-0.010	0	2	15	12	1	0
BRKM5	-0.047	0.016	0.000	-0.025	-0.005	0	0	13	17	0	0
BRML3	0.050	-0.035	-0.053	-0.032	-0.018	0	3	13	13	1	0
CCRO3	-0.057	-0.042	-0.048	-0.052	-0.009	0	2	16	12	0	0
CMIG4	-0.038	0.007	-0.050	-0.030	0.056	0	0	14	15	1	0
CPFE3	-0.104	-0.042	-0.064	-0.026	0.033	1	3	13	13	0	0
CPLE6	-0.075	-0.042	-0.021	0.003	0.029	0	1	15	14	0	0
CSAN3	-0.005	-0.039	-0.041	-0.058	0.046	0	2	15	10	3	0
CSNA3	0.033	0.010	-0.002	0.017	0.016	0	0	11	18	1	0
CYRE3	0.013	-0.002	-0.070	-0.020	0.012	0	2	11	16	1	0
EGIE3	-0.180	-0.004	-0.015	-0.087	0.045	2	17	10	0	0	0
ELET3	0.019	0.010	-0.025	-0.025	0.050	0	0	11	19	0	0
EMBR3	-0.062	0.033	-0.035	0.004	0.040	0	2	15	12	1	0
ENBR3	-0.139	-0.025	-0.024	-0.008	-0.013	1	0	17	12	0	0
FIBR3	0.038	0.038	-0.036	-0.006	0.033	0	0	12	16	2	0
GGBR4	0.014	0.031	-0.060	-0.015	0.003	0	1	10	17	2	0
GOAU4	0.038	0.037	-0.036	-0.001	0.032	0	0	9	21	0	0
ITSA4	-0.013	0.002	-0.065	-0.015	-0.024	0	3	13	14	0	0
ITUB4	-0.004	-0.003	-0.076	-0.004	-0.047	0	2	15	13	0	0
JBSS3	-0.078	-0.036	-0.068	-0.001	0.052	0	2	13	14	1	0
LAME4	-0.076	-0.057	-0.027	-0.016	0.036	0	4	13	11	2	0
LREN3	-0.084	-0.050	-0.063	0.013	0.042	0	3	14	12	1	0
MRFG3	-0.006	0.028	-0.001	-0.011	0.014	0	0	14	15	1	0
MULT3	-0.036	-0.054	-0.064	-0.024	-0.027	1	5	12	9	3	0
NATU3	-0.027	-0.019	-0.041	-0.019	0.022	0	0	17	12	1	0
PCAR4	-0.053	0.002	-0.049	0.011	0.004	0	1	15	13	1	0
PETR3	0.000	-0.008	-0.046	0.014	0.018	0	0	13	16	1	0
PETR4	0.000	-0.001	-0.034	0.013	0.018	0	0	11	19	0	0
SUZB5	0.011	-0.009	-0.020	-0.044	-0.016	0	0	13	16	1	0
TIMP3	-0.030	-0.021	-0.058	-0.038	-0.020	0	4	15	9	2	0
USIM5	0.078	-0.004	-0.032	-0.001	0.061	0	1	10	15	4	0
VALE3	0.034	-0.055	-0.081	-0.014	-0.010	0	3	14	11	2	0
VIVT4	-0.115	0.000	-0.050	-0.007	-0.024	1	1	17	11	0	0
WEGE3	-0.067	-0.008	-0.074	-0.021	0.021	0	3	16	11	0	0

Tabela 5.5. Valores dos 5 primeiros *lags* da função de autocorrelação dos log-retornos financeiros absolutos. Na Tabela também são demonstrados os números totais de observações encontradas nos intervalos definidos de 1 a 6 para os 30 primeiros *lags* das séries

	Lags					**Intervalos**					
Ações	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**6**
ABEV3	0.163	0.229	0.188	0.249	0.198	0	0	0	0	0	30
BBAS3	0.218	0.241	0.208	0.187	0.259	0	0	0	0	1	29
BBDC4	0.221	0.245	0.168	0.233	0.217	0	0	0	0	0	30
BRAP4	0.187	0.221	0.213	0.192	0.249	0	0	0	0	0	30
BRFS3	0.227	0.179	0.173	0.161	0.213	0	0	0	0	3	27
BRKM5	0.171	0.171	0.131	0.173	0.116	0	0	0	4	14	12
BRML3	0.288	0.285	0.195	0.214	0.177	0	0	0	0	4	26
CCRO3	0.184	0.151	0.178	0.145	0.085	0	0	0	0	18	12
CMIG4	0.231	0.167	0.152	0.195	0.161	0	0	0	0	5	25
CPFE3	0.236	0.223	0.202	0.17	0.2	0	0	0	0	0	30
CPLE6	0.177	0.218	0.093	0.161	0.125	0	0	0	0	11	19
CSAN3	0.239	0.259	0.242	0.218	0.253	0	0	0	0	0	30
CSNA3	0.276	0.247	0.239	0.205	0.226	0	0	0	0	0	30
CYRE3	0.301	0.297	0.278	0.303	0.303	0	0	0	0	0	30
EGIE3	0.214	0.19	0.155	0.168	0.127	0	0	0	1	7	22
ELET3	0.207	0.196	0.115	0.141	0.128	0	0	0	0	16	14
EMBR3	0.205	0.149	0.115	0.098	0.162	0	0	0	0	10	20
ENBR3	0.148	0.177	0.1	0.115	0.053	0	0	0	1	14	15
FIBR3	0.246	0.221	0.267	0.203	0.186	0	0	0	0	0	30
GGBR4	0.225	0.213	0.254	0.2	0.246	0	0	0	0	0	30
GOAU4	0.238	0.198	0.267	0.214	0.208	0	0	0	0	0	30
ITSA4	0.226	0.285	0.193	0.248	0.208	0	0	0	0	1	29
ITUB4	0.207	0.238	0.178	0.236	0.216	0	0	0	0	1	29
JBSS3	0.199	0.213	0.159	0.177	0.145	0	0	0	0	3	27
LAME4	0.2	0.303	0.192	0.237	0.272	0	0	0	0	2	28
LREN3	0.263	0.25	0.244	0.254	0.253	0	0	0	0	0	30
MRFG3	0.153	0.159	0.077	0.084	0.087	0	0	0	6	19	5
MULT3	0.275	0.224	0.194	0.183	0.114	0	0	0	0	6	24
NATU3	0.116	0.121	0.076	0.077	0.066	0	0	0	1	18	11
PCAR4	0.187	0.206	0.189	0.129	0.16	0	0	0	0	1	29
PETR3	0.171	0.232	0.189	0.22	0.199	0	0	0	0	0	30
PETR4	0.22	0.252	0.211	0.222	0.228	0	0	0	0	0	30
SUZB5	0.128	0.2	0.124	0.143	0.108	0	0	0	0	18	12
TIMP3	0.202	0.147	0.145	0.128	0.14	0	0	0	0	4	26
USIM5	0.191	0.19	0.158	0.167	0.214	0	0	0	0	1	29
VALE3	0.227	0.256	0.229	0.194	0.23	0	0	0	0	0	30
VIVT4	0.153	0.131	0.113	0.134	0.157	0	0	0	1	10	19
WEGE3	0.247	0.263	0.235	0.151	0.218	0	0	0	0	0	30

Tabela 5.6. Valores dos 5 primeiros *lags* da função de autocorrelação dos log-retornos financeiros ao quadrado. Na Tabela também são demonstrados os números totais de observações encontradas nos intervalos definidos de 1 a 6 para os 30 primeiros *lags* das séries

	Lags					**Intervalos**					
Ações	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**6**
ABEV3	0.102	0.251	0.152	0.295	0.27	0	0	0	0	5	25
BBAS3	0.147	0.253	0.159	0.135	0.294	0	0	0	1	6	23
BBDC4	0.174	0.213	0.081	0.196	0.215	0	0	0	0	7	23
BRAP4	0.139	0.247	0.208	0.143	0.253	0	0	0	0	1	29
BRFS3	0.191	0.173	0.131	0.142	0.223	0	0	0	1	11	18
BRKM5	0.107	0.183	0.09	0.116	0.065	0	0	1	14	11	4
BRML3	0.467	0.355	0.251	0.266	0.211	0	0	0	0	6	24
CCRO3	0.209	0.136	0.299	0.161	0.055	0	0	0	6	14	10
CMIG4	0.201	0.109	0.116	0.145	0.091	0	0	0	13	12	5
CPFE3	0.188	0.162	0.2	0.088	0.151	0	0	0	0	9	21
CPLE6	0.14	0.211	0.075	0.128	0.124	0	0	0	5	9	16
CSAN3	0.166	0.317	0.162	0.215	0.203	0	0	0	0	0	30
CSNA3	0.282	0.229	0.167	0.148	0.175	0	0	0	1	0	29
CYRE3	0.213	0.331	0.223	0.406	0.274	0	0	0	0	1	29
EGIE3	0.222	0.249	0.222	0.176	0.085	0	0	0	6	11	13
ELET3	0.277	0.161	0.067	0.086	0.071	0	0	0	11	15	4
EMBR3	0.18	0.128	0.092	0.071	0.133	0	0	0	5	18	7
ENBR3	0.098	0.189	0.082	0.069	0.019	0	0	0	5	13	12
FIBR3	0.313	0.248	0.309	0.257	0.181	0	0	0	0	1	29
GGBR4	0.198	0.225	0.247	0.205	0.249	0	0	0	0	1	29
GOAU4	0.21	0.19	0.271	0.205	0.21	0	0	0	0	1	29
ITSA4	0.221	0.372	0.161	0.278	0.226	0	0	0	0	3	27
ITUB4	0.192	0.247	0.167	0.221	0.2	0	0	0	0	2	28
JBSS3	0.171	0.351	0.127	0.189	0.116	0	0	0	0	8	22
LAME4	0.127	0.414	0.139	0.297	0.273	0	0	0	0	5	25
LREN3	0.25	0.289	0.306	0.254	0.277	0	0	0	0	1	29
MRFG3	0.065	0.17	0.029	0.039	0.034	0	0	0	22	7	1
MULT3	0.244	0.263	0.195	0.112	0.064	0	0	0	5	12	13
NATU3	0.107	0.132	0.054	0.07	0.032	0	0	0	9	16	5
PCAR4	0.218	0.21	0.166	0.1	0.134	0	0	0	0	5	25
PETR3	0.144	0.201	0.152	0.167	0.194	0	0	0	0	0	30
PETR4	0.218	0.213	0.177	0.166	0.226	0	0	0	0	0	30
SUZB5	0.079	0.217	0.108	0.209	0.087	0	0	0	1	15	14
TIMP3	0.094	0.069	0.052	0.113	0.069	0	0	0	6	20	4
USIM5	0.114	0.143	0.108	0.259	0.182	0	0	0	0	9	21
VALE3	0.171	0.231	0.208	0.128	0.191	0	0	0	0	1	29
VIVT4	0.136	0.12	0.084	0.122	0.151	0	0	0	4	11	15
WEGE3	0.205	0.237	0.214	0.075	0.175	0	0	0	1	10	19

Tabela 5.7. Estatística Q de Ljung-Box para as séries de log-retornos financeiros (R_t), log-retornos financeiros absolutos ($|R_t|$) e log-retornos financeiros elevados ao quadrado ($(R_t)^2$)

Ações	R_t	$\|R_t\|$	$(R_t)^2$
ABEV3	93.314	1862.046	2137.381
BBAS3	32.653	2329.112	1858.297
BBDC4	62.842	2442.965	2131.163
BRAP4	52.891	2576.067	2225.968
BRFS3	61.492	1413.581	1112.144
BRKM5	28.170	673.275	382.542
BRML3	62.914	1727.282	2897.589
CCRO3	37.745	790.5744	843.093
CMIG4	54.645	1256.018	406.329
CPFE3	78.408	2060.532	1399.371
CPLE6	45.132	1043.536	897.283
CSAN3	78.683	2659.759	2206.326
CSNA3	41.005	3073.656	2279.187
CYRE3	56.313	4314.368	3648.889
EGIE3	121.626	976.394	955.588
ELET3	45.626	805.191	518.683
EMBR3	62.805	920.775	573.210
ENBR3	79.885	803.629	837.988
FIBR3	48.446	2228.765	2977.084
GGBR4	52.254	2849.242	2823.648
GOAU4	50.357	2751.245	2348.391
ITSA4	57.511	2518.185	3260.629
ITUB4	52.225	2358.931	2528.773
JBSS3	54.853	1263.899	1230.712
LAME4	97.204	2389.009	2991.267
LREN3	77.973	2810.158	3513.287
MRFG3	27.989	481.683	171.588
MULT3	143.393	1438.575	1210.174
NATU3	45.895	558.434	360.558
PCAR4	40.998	1628.726	1650.831
PETR3	27.125	2482.795	1985.783
PETR4	21.637	2822.176	2404.296
SUZB5	43.984	673.085	865.562
TIMP3	90.581	1151.716	425.175
USIM5	71.459	1565.242	1109.921
VALE3	80.693	2932.774	2112.595
VIVT4	55.112	808.526	662.632
WEGE3	70.796	2059.389	1449.501

Tabela 5.8. Resultado do teste da razão da variância. O símbolo (*) destaca os valores que rejeitam a hipótese de passeio aleatório.

Ações	**N = 2**	**N = 5**	**N = 20**	**Total**
ABEV3	-3.168*	-3.805*	-2.543*	30
BBAS3	0.440	-0.483	-0.892	0
BBDC4	-1.376	-2.286*	-2.022*	16
BRAP4	-0.206	-1.067	0.059	0
BRFS3	-0.990	-1.887	-1.929	10
BRKM5	-1.707	-1.065	-0.949	0
BRML3	1.011	-0.219	-0.990	0
CCRO3	-1.805	-2.944*	-2.981*	19
CMIG4	-1.024	-1.386	-0.448	0
CPFE3	-3.562*	-4.406*	-2.949*	30
CPLE6	-2.631*	-2.913*	-1.561	14
CSAN3	-0.173	-1.590	-0.888	0
CSNA3	0.950	0.943	1.231	0
CYRE3	0.354	-0.564	-0.816	0
EGIE3	-5.343*	-4.556*	-3.603*	30
ELET3	0.548	0.176	0.059	0
EMBR3	-2.033*	-1.351	-0.768	1
ENBR3	-5.564*	-4.871*	-3.393*	30
FIBR3	1.113	1.009	1.100	0
GGBR4	0.467	0.081	0.277	0
GOAU4	1.256	1.144	1.286	2
ITSA4	-0.377	-0.941	-1.773	0
ITUB4	-0.134	-0.967	-1.777	6
JBSS3	-2.574*	-3.124*	-1.834	15
LAME4	-2.550*	-2.831*	-1.579	13
LREN3	-2.561*	-3.315*	-1.693	16
MRFG3	-0.235	0.304	0.247	0
MULT3	-0.811	-1.859	-1.910	12
NATU3	-1.111	-2.001*	-2.261*	21
PCAR4	-1.704	-1.785	-1.379	0
PETR3	0.001	-0.700	-0.432	0
PETR4	-0.021	-0.396	0.033	0
SUZB5	0.431	-0.491	-0.084	0
TIMP3	-0.989	-2.073*	-2.002*	20
USIM5	2.767*	1.478	1.750	2
VALE3	1.131	-1.258	-1.230	0
VIVT4	-4.371*	-4.025*	-3.267*	30
WEGE3	-1.800	-2.259*	-1.471	8

Tabela 5.9. Total de correlações não lineares significativas para os 30 primeiros *lags* das séries de log-retornos financeiros (retorno), log-retornos financeiros ao quadrado (retorno ao quadrado), e log-retornos financeiros absolutos (retorno absoluto) ao utilizar a função ADCF

	Total de correlações significativas		
Ações	**retorno**	**retorno ao quadrado**	**retorno absoluto**
ABEV3	26	30	30
BBAS3	30	30	30
BBDC4	30	30	30
BRAP4	30	30	30
BRFS3	27	30	30
BRKM5	15	30	30
BRML3	23	30	30
CCRO3	16	30	30
CMIG4	29	30	30
CPFE3	30	30	30
CPLE6	21	30	30
CSAN3	30	30	30
CSNA3	30	30	30
CYRE3	30	30	30
EGIE3	11	30	30
ELET3	22	30	30
EMBR3	21	30	30
ENBR3	14	30	30
FIBR3	26	30	30
GGBR4	30	30	30
GOAU4	30	30	30
ITSA4	28	30	30
ITUB4	29	30	30
JBSS3	22	30	30
LAME4	26	30	30
LREN3	29	30	30
MRFG3	10	30	30
MULT3	22	30	30
NATU3	22	30	30
PCAR4	27	30	30
PETR3	30	30	30
PETR4	30	30	30
SUZB5	17	30	30
TIMP3	27	30	30
USIM5	29	30	30
VALE3	30	30	30
VIVT4	14	30	30
WEGE3	29	30	30

Tabela 5.10. Resultados dos p-valores para o teste ADCF contra a hipótese i.i.d. das séries de log-retornos financeiros, ao se variar o parâmetro p. Para o teste foi utilizada a função kernel *bartlett* e 499 réplicas de *bootsrap*

	Teste ADCF		
Ações	**p = 4.68**	**p = 21.98**	**p = 103.1**
ABEV3	2.20E-16	2.20E-16	2.20E-16
BBAS3	2.20E-16	2.20E-16	2.20E-16
BBDC4	2.20E-16	2.20E-16	2.20E-16
BRAP4	2.20E-16	2.20E-16	2.20E-16
BRFS3	0.002	2.20E-16	2.20E-16
BRKM5	2.20E-16	2.20E-16	2.20E-16
BRML3	2.20E-16	2.20E-16	2.20E-16
CCRO3	2.20E-16	2.20E-16	2.20E-16
CMIG4	2.20E-16	2.20E-16	2.20E-16
CPFE3	2.20E-16	2.20E-16	2.20E-16
CPLE6	2.20E-16	2.20E-16	2.20E-16
CSAN3	2.20E-16	2.20E-16	2.20E-16
CSNA3	2.20E-16	2.20E-16	2.20E-16
CYRE3	2.20E-16	2.20E-16	2.20E-16
EGIE3	2.20E-16	2.20E-16	2.20E-16
ELET3	2.20E-16	2.20E-16	2.20E-16
EMBR3	2.20E-16	2.20E-16	2.20E-16
ENBR3	2.20E-16	2.20E-16	2.20E-16
FIBR3	2.20E-16	2.20E-16	2.20E-16
GGBR4	2.20E-16	2.20E-16	2.20E-16
GOAU4	2.20E-16	2.20E-16	2.20E-16
ITSA4	2.20E-16	2.20E-16	2.20E-16
ITUB4	2.20E-16	2.20E-16	2.20E-16
JBSS3	2.20E-16	2.20E-16	2.20E-16
LAME4	2.20E-16	2.20E-16	2.20E-16
LREN3	2.20E-16	2.20E-16	2.20E-16
MRFG3	2.20E-16	2.20E-16	2.20E-16
MULT3	0.004	2.20E-16	2.20E-16
NATU3	0.018	2.20E-16	2.20E-16
PCAR4	2.20E-16	2.20E-16	2.20E-16
PETR3	2.20E-16	2.20E-16	2.20E-16
PETR4	2.20E-16	2.20E-16	2.20E-16
SUZB5	6.00E-03	6.00E-03	6.00E-03
TIMP3	2.20E-16	2.20E-16	2.20E-16
USIM5	2.20E-16	2.20E-16	2.20E-16
VALE3	2.20E-16	2.20E-16	2.20E-16
VIVT4	2.20E-16	2.20E-16	2.20E-16
WEGE3	2.20E-16	2.20E-16	2.20E-16

identicamente distribuídas (i.i.d.). A Tabela 5.7 demonstra os resultados encontrados ao se calcular a estatística Q de Ljung-Box. A distribuição assintótica da estatística Q é qui-quadrada com k graus de liberdade, quando o processo é i.i.d. A hipótese nula é rejeitada no intervalo de 5% se $Q > 43.77$ e no nível de 1% se $Q > 50.89$. Ao analisar as autocorrelações dos log-retornos financeiros, 30 ações rejeitam a hipótese a nível de 5% e 24 rejeitam a nível de 1%. Já ao se analisar tanto os log-retornos financeiros absolutos quanto elevados ao quadrado, todas as 38 ações rejeitam a hipótese a nível de 1%. Este teste sugere que apesar da maioria dos *lags* individuais das séries de log-retornos financeiros não apresentar valores significativos, existe uma dependência linear que não pode ser desprezível, já que ao analisar os 30 primeiros *lags* em conjunto os mesmos rejeitam a hipótese nula. Fica evidente também que há uma maior dependência linear nas transformações dos log-retornos financeiros, confirmando mais um fato estilizado conhecido na literatura. A Figura 5.4 demonstra como exemplo as comparações dos valores de autocorrelação para as séries de log-retorno financeiro, log-retorno financeiro absoluto e log-retorno financeiro ao quadrado da ação BBAS3.

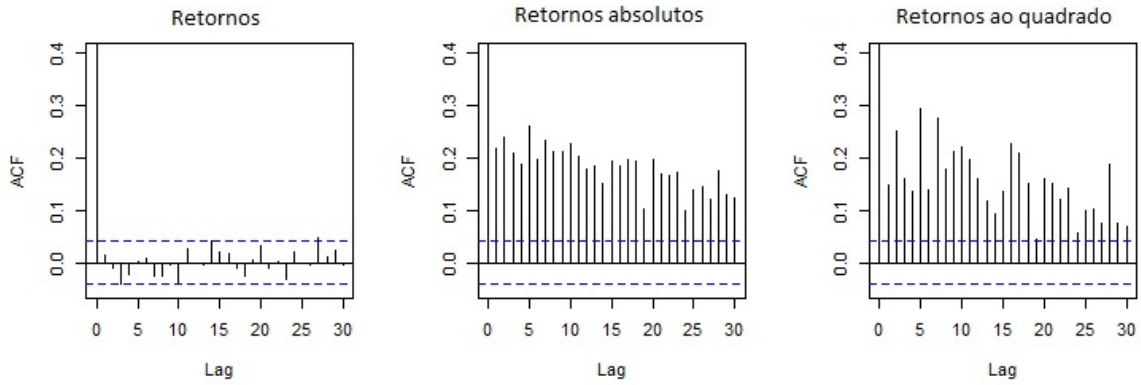

Figura 5.4. Comparação dos valores da função de autocorrelação (ACF) para as séries de log-retorno financeiro, log-retorno financeiro absoluto e log-retorno financeiro ao quadrado da ação BBAS3

A Tabela 5.8 exibe os resultados da aplicação do teste da razão da variância nas séries dos log-retornos financeiros. Na Tabela são demonstrados os valores obtidos pelo teste para $N = 2, N = 5$ e $N = 20$ uma vez que estes valores representam respectivamente um período diário, semanal e mensal. A escolha do parâmetro N para o teste é arbitrária e por este motivo o teste foi aplicado também variando N de 2 até 31. Ao final é obtido para cada ação quantas vezes o teste foi rejeitado ao se variar o valor de N. Do total das 38 ações, 20 séries de log-retornos financeiros apresentaram valores significativos para a estatística indicando uma rejeição da hipótese nula. E ao

se observar os resultados, percebe-se que a maiora das vezes em que o teste rejeitou a hipótese nula, a mesma foi rejeitada para diferentes valores de N.

Os resultados do teste da razão da variância reafirmam que 20 ações possuem uma dependência linear temporal que não pode ser desprezada. Além do teste de Ljung-Box refutar a hipótese de que a maioria das séries de log-retorno financeiro não são i.i.d.s por seus valores da ACF, os resultados do teste da razão da variância demonstram que 20 séries de log-retornos financeiros também refutam a hipótese de não serem correlacionadas linearmente.

Com o objetivo de compreender as dependências não lineares nas séries de log-retornos financeiros, foram analisados os valores de função de auto correlação de distância (ADCF). A Tabela 5.9 demonstra para cada ação o número de *lags* significativos para as séries de log-retornos financeiros, log-retornos financeiros absolutos e log-retornos financeiros ao quadrado. Nota-se que ao contrário dos resultados anteriores, as séries de log-retornos financeiros demonstram valores significativos para a maioria dos 30 primeiros *lags*, indicando a presença de correlações não lineares. Estes valores significativos indicam que os *lags* não são independentes. Percebe-se também que as séries transformadas dos log-retornos financeiros continuam possuindo mais *lags* significativos do que a série de log-retorno financeiro, e com valores maiores. A Figura 5.5 ilustra o comportamento da função ADCF para a série de log-retorno financeiro da ação BBAS3 e suas respectivas transformações.

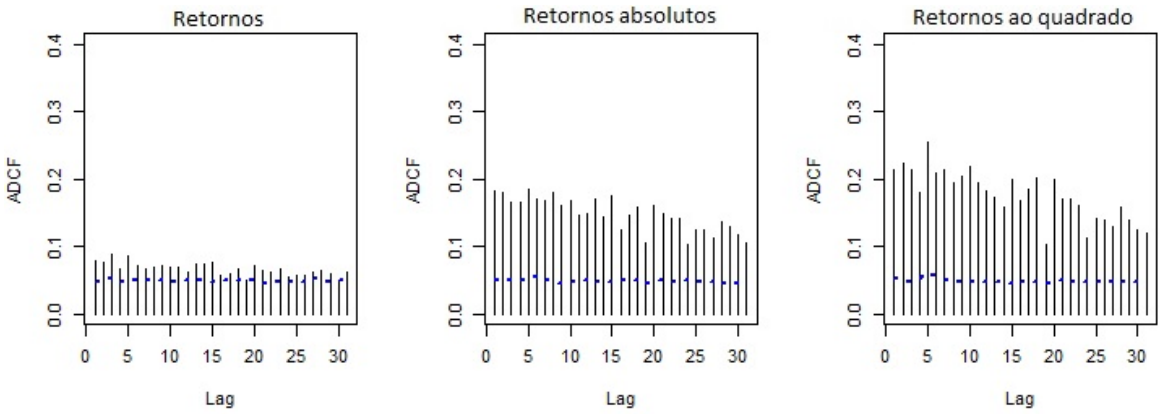

Figura 5.5. Comparação dos valores da função de auto correlação de distância (ADCF) para as séries de log-retorno financeiro , log-retorno financeiro absoluto e retorno log-financeiro ao quadrado da ação BBAS3

Os valores da função ADCF também são utilizados para testar a hipótese i.i.d. dos log-retornos financeiros. Foi aplicado o teste ADCF variando o parâmetro p. Foi utilizada a função *kernel Barrtlet* com 500 réplicas de *bootstrap* e variando o parâmetro

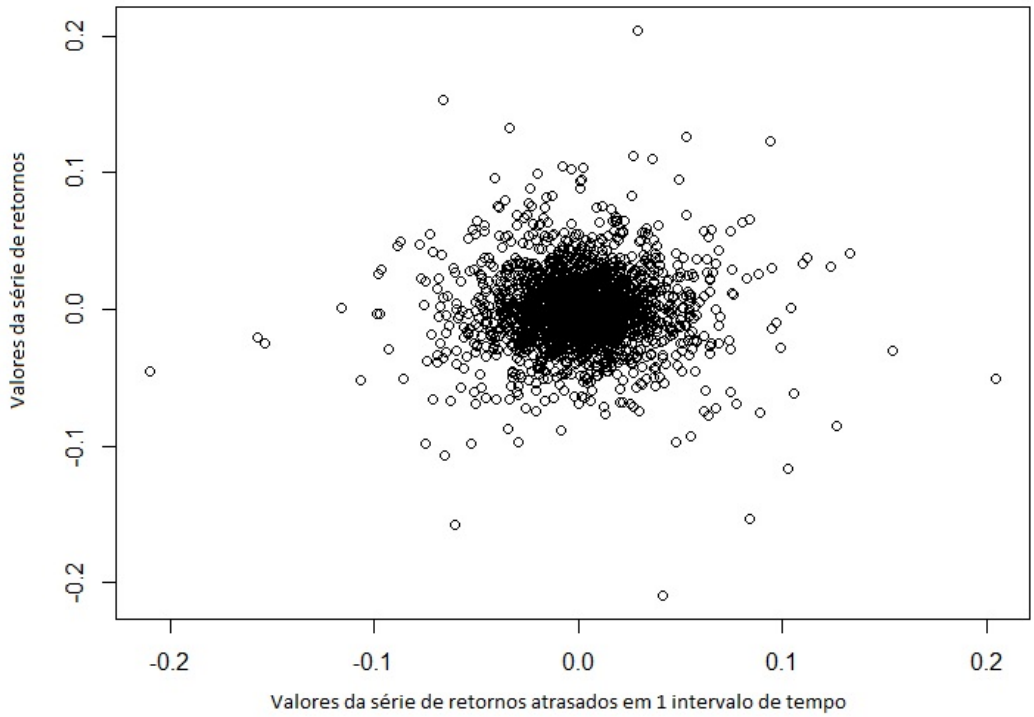

Figura 5.6. Gráfico de dispersão da série de log-retornos financeiros pela série de log-retornos financeiros atrasados em 1 intervalo de tempo. Valores relacionados à ação BBAS3.

p em 0.2, 0.4 e 0.6. A configuração dos parâmetros aplicada ao teste segue a mesma metodologia de Fokianos & Pitsillou [2017]. A Tabela 5.10 demonstra os resultados encontrados. Para todas as ações novamente foi rejeitada a hipótese i.i.d. no nível de 5%.

Apesar dos resultados da função de auto correlação de distância demonstrarem que existe uma dependência não linear temporal nas séries, esta dependência não parece ser forte. Os valores encontrados são significativos mas possuem uma magnitude baixa, muito próxima da fronteira do valor crítico. Este fato também pode ser observado por meio da Figura 5.6 em que são demonstrados nos eixos do gráfico a série de log-retorno financeiro da ação BBAS3 e o *lag* 1 da mesma série. Não é fácil encontrar um padrão não linear na Figura.

A análise da autocorrelação das séries reafirma um fato estilizado já mencionado na literatura: maior evidência de correlação linear nas séries de log-retorno financeiro ao quadrado e valores absolutos em comparação com a série de log-retornos financeiros.

A análise da função ADCF demonstra que as séries de log-retorno financeiro possuem correlações não lineares significativas ao longo do tempo, indicando a dependência temporal nos valores passados de log-retorno financeiro.

5.4 Explorando a correlação não linear: criação de um atributo mais correlacionado por meio dos atributos originais

Os 4 atributos originais de entrada do modelo de predição correspondem à série de valores futuros atrasados em 1, 2, 3 e 4 *lags* respectivamente. Nesta etapa foi calculada uma combinação linear destes 4 atributos ($X_{-1} = x_{1-1}, x_{2-1}, ..., x_{166-1}, X_{-2} = x_{1-2}, x_{2-2}, ..., x_{166-2}, X_{-3} = x_{1-3}, x_{2-3}, ..., x_{166-3}$ e $X_{-4} = x_{1-4}, x_{2-4}, ..., x_{166-4}$) que busca maximizar a correlação de distância com a série de valores futuros do conjunto de treinamento ($X = x_1, x_2, ..., x_{166}$). Para resolver o problema de otimização foi utilizado o algoritmo *differential evolution*, utilizando a linguagem *python* juntamente com a biblioteca *scipy optimize*.

A Tabela 5.11 demonstra os valores dos pesos ws encontrados pelo algoritmo de otimização. Também são demonstrados os resultados dos p-valores ao aplicar o teste T de correlação de distância entre o novo atributo gerado e a série alvo de valores futuros no conjunto de treino. Os resultados demonstram que das 38 ações, 29 rejeitaram a hipótese de independência.

Na Tabela 5.12 são demonstrados os resultados da medida de correlação de distância das 4 entradas do modelo de previsão com a série alvo de valores futuros do modelo. Os valores são comparados com o valor da correlação de distância do novo atributo gerado a partir dos atributos originais, novamente com a série alvo de valores futuros. Pode-se perceber que os novos atributos gerados apresentam valores de correlação de distância maiores do que os atributos originais para todas as séries.

Os resultados desta Seção demonstram que é possível explorar a correlação não linear existente nas séries e que o novo atributo gerado demonstra possuir mais dependência com a série de valores futuros em comparação aos atributos originais.

Tabela 5.11. Valores dos pesos encontrados pelo algoritmo *differential evolution* que buscam maximizar a correlação de distância entre a combinação dos atributos originais e a série alvo de valores futuros do conjunto de treinamento. Também é demonstrado na tabela os p-valores para o teste T da correlação de distância para testar a hipótese do novo atributo ser independente da série alvo de valores futuros do conjunto de treinamento.

	Pesos				**Teste T**
Ações	**w1**	**w2**	**w3**	**w4**	**p-value**
ABEV3	0.920	-0.186	-0.107	0.542	2.0E-4
BBAS3	-0.636	-0.870	-0.457	-0.112	0.075
BBDC4	-0.631	0.484	-0.661	-0.655	4.8E-4
BRAP4	0.449	0.443	0.716	0.455	0.006
BRFS3	-0.294	-0.268	-0.877	0.211	0.104
BRKM5	0.885	-0.255	-0.280	-0.203	0.088
BRML3	0.041	0.594	-0.013	0.035	0.003
CCRO3	0.539	-0.049	0.248	0.669	0.014
CMIG4	0.542	-0.641	-0.440	0.353	0.082
CPFE3	0.776	0.952	-0.134	0.783	2.20E-16
CPLE6	0.905	0.636	-0.173	0.484	1.99E-07
CSAN3	0.145	0.599	-0.244	-0.853	0.044
CSNA3	0.205	-0.886	-0.328	-0.142	3.2E-4
CYRE3	-0.448	0.265	0.462	0.937	4.48E-06
EGIE3	-0.854	0.153	-0.109	-0.372	1.99E-11
ELET3	0.355	-0.304	0.917	-0.416	0.112
EMBR3	0.966	0.101	-0.042	0.137	0.013
ENBR3	0.695	0.499	-0.448	0.726	0.032
FIBR3	0.110	-0.307	-0.811	-0.364	0.004
GGBR4	-0.195	0.682	-0.147	-0.613	0.001
GOAU4	0.650	-0.868	0.511	-0.346	0.052
ITSA4	-0.952	0.185	-0.395	-0.716	0.009
ITUB4	0.124	-0.658	0.903	0.858	0.006
JBSS3	0.857	-0.005	0.235	-0.115	6.13E-10
LAME4	0.244	-0.946	-0.696	0.886	0.043
LREN3	-0.214	0.732	0.870	-0.873	0.001
MRFG3	-0.842	-0.787	-0.365	-0.571	2.20E-16
MULT3	-0.020	-0.903	-0.381	-0.267	0.014
NATU3	-0.223	0.836	-0.368	-0.489	0.197
PCAR4	0.905	0.292	-0.149	-0.024	0.001
PETR3	-0.748	-0.369	0.075	-0.073	0.039
PETR4	-0.750	-0.582	-0.515	-0.237	0.087
SUBZ5	-0.555	-0.013	-0.482	0.738	2.48E-05
TIMP3	-0.778	0.033	0.262	0.224	2.87E-07
USIM5	0.483	0.748	0.953	0.262	0.034
VALE3	-0.103	0.177	0.237	0.694	0.003
VIVT4	0.840	0.321	-0.702	-0.545	0.129
WEGE3	0.907	-0.343	-0.712	0.302	8.6E-4

Tabela 5.12. Comparação dos valores de correlação de distância das entradas originais e do novo atributo com os valores futuros da série alvo do conjunto de treinamento

Ações	*Lag* 1	*Lag* 2	*Lag* 3	*Lag* 4	Novo atributo
ABEV3	0.201	0.145	0.118	0.168	0.227
BBAS3	0.156	0.155	0.141	0.153	0.205
BBDC4	0.126	0.206	0.15	0.17	0.232
BRAP4	0.165	0.123	0.153	0.14	0.213
BRFS3	0.151	0.134	0.148	0.142	0.176
BRKM5	0.183	0.144	0.109	0.14	0.193
BRML3	0.128	0.213	0.158	0.143	0.213
CCRO3	0.177	0.143	0.121	0.145	0.202
CMIG4	0.127	0.132	0.101	0.111	0.188
CPFE3	0.254	0.187	0.216	0.221	0.325
CPLE6	0.198	0.141	0.151	0.162	0.257
CSAN3	0.116	0.155	0.118	0.135	0.18
CSNA3	0.183	0.197	0.145	0.152	0.227
CYRE3	0.165	0.141	0.158	0.211	0.257
EGIE3	0.27	0.124	0.116	0.138	0.29
ELET3	0.143	0.126	0.144	0.13	0.175
EMBR3	0.212	0.146	0.136	0.135	0.224
ENBR3	0.143	0.126	0.128	0.179	0.192
FIBR3	0.139	0.121	0.18	0.171	0.222
GGBR4	0.15	0.153	0.154	0.149	0.223
GOAU4	0.122	0.149	0.156	0.146	0.19
ITSA4	0.153	0.141	0.147	0.155	0.207
ITUB4	0.13	0.144	0.132	0.144	0.214
JBSS3	0.249	0.171	0.178	0.122	0.275
LAME4	0.115	0.113	0.131	0.123	0.194
LREN3	0.14	0.154	0.095	0.154	0.204
MRFG3	0.217	0.183	0.124	0.147	0.304
MULT3	0.138	0.16	0.105	0.143	0.194
NATU3	0.136	0.144	0.12	0.156	0.168
PCAR4	0.194	0.137	0.147	0.12	0.219
PETR3	0.16	0.136	0.143	0.111	0.191
PETR4	0.144	0.147	0.126	0.098	0.177
SUZB5	0.149	0.136	0.192	0.189	0.248
TIMP3	0.234	0.123	0.093	0.117	0.252
USIM5	0.157	0.154	0.151	0.151	0.201
VALE3	0.209	0.119	0.159	0.199	0.214
VIVT4	0.156	0.109	0.11	0.147	0.177
WEGE3	0.195	0.149	0.126	0.141	0.217

5.5 Instanciação do modelo de predição: rede neural *Long Short Term Memory*

Essa etapa é responsável por instanciar um modelo de predição de valores futuros das séries que servem de base para a implementação das estratégias financeiras. A seguir são descritas as etapas da implementação do modelo. Foi utilizada uma rede neural LSTM uma vez que este algoritmo apresentou melhor desempenho que outros também previamente testados, como por exemplo, rede neural *Extreme Learning Machine* (ELM), rede neural *Multilayer Perceptron* (MLP) e *Random Forest*. A rede neural foi implementada utilizando a linguagem *python* com auxílio da biblioteca *Keras*.

5.5.1 Modelagem do problema em classes binárias

A rede neural LSTM foi utilizada para previsão de tendência de subida ou queda nas observações das séries temporais. Foram atribuídos a classe de Altas todos os valores da série de logaritmo do retorno financeiro maiores ou iguais a zero, e a classe de Baixas os valores menores que zero.

5.5.2 Definição do conjunto de treinamento e conjunto de teste

Foi escolhido o período de 1 ano (aproximadamente 250 dias), relativo ao ano de 2016, para aplicar o modelo de previsão da rede neural. O conjunto de treinamento corresponde a dois terços (aproximadamente dos meses de Janeiro a Agosto) com 166 dias. Já o conjunto de teste dos dados corresponde aos meses seguintes de Setembro a Dezembro, com 84 dias. Foi escolhido o ano de 2016 para aplicação do modelo de previsão já que o conjunto de dados representava o ano mais atual.

A Figura 5.7 ilustra como exemplo, algumas séries originais de preços de 4 ações divididas entre conjunto de treinamento e conjunto de teste. Pela Figura percebe-se os mais diversos tipos de tendência nos preços (movimento de subida, descida e lateralidade) utilizados para o algoritmo de aprendizado de máquina. Vale ressaltar que na Figura são demonstradas as séries de preços mas o algoritmo utiliza as séries de log-retornos financeiros normalizadas para a etapa do aprendizado.

Conforme a metodologia do trabalho, são escolhidas como entradas valores passados das séries, uma vez que estes podem conter dependências temporais entre si. O número de entradas foi definido como 4 pois este valor de entradas apresentou previamente um bom desempenho e não prejudicou muito no tempo de execução do algoritmo.

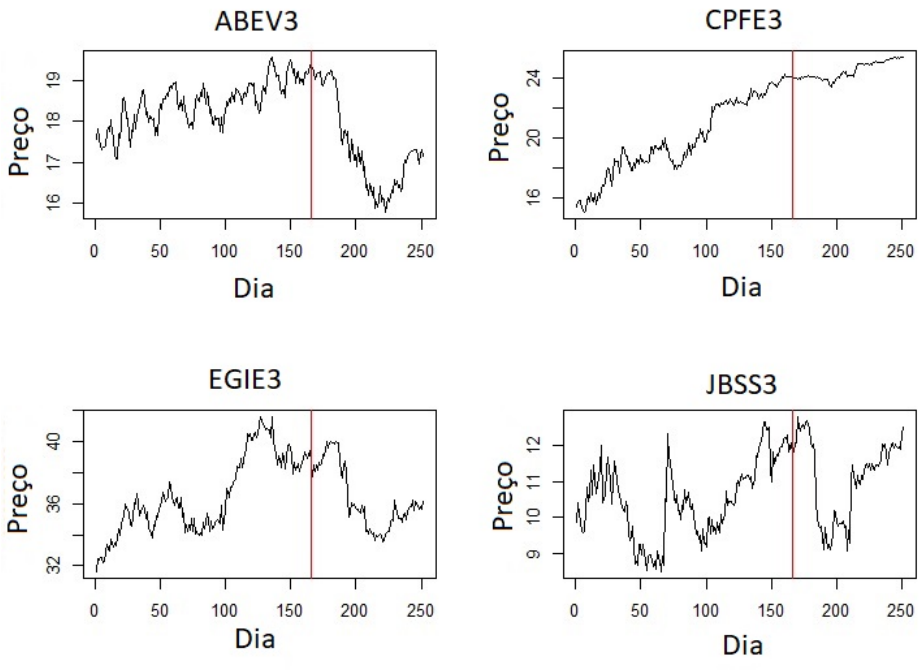

Figura 5.7. Comportamento de algumas séries de preços de fechamento relativas aos conjuntos de treino e teste

Sendo assim, o conjunto de treino de cada ação corresponde a uma matriz de 166 linhas e 4 colunas com duas diferentes configurações: utilizando as séries originais do log-retorno financeiro atrasadas em 1, 2, 3 e 4 *lags*, e ao utilizar 4 novas séries criadas pelo método da Seção 4.4 do capítulo da Metodologia. Dessa forma, o método da Seção 4.4 é executado 4 vezes (para gerar 4 novos atributos para cada ação), sendo que os pesos encontrados para cada uma das combinações lineares são diferentes.

5.5.3 Arquitetura da rede LSTM

A rede LSTM utiliza 4 camadas. A primeira camada contem 8 unidades LSTM, a segunda camada contém 4 unidades LSTM com regularizador L1 de valor λ igual a 0.02 e a terceira camada contém 2 unidades LSTM também com regularizador L1 de valor λ igual a 0.02. A camada de saída contém um neurônio com função de ativação sigmoide. Portanto, a saída da rede é um valor no intervalo [0,1]. A rede é treinada utilizando otimizador Adamax e função de perda entropia cruzada binária. O treinamento é executado por 250 épocas com batches de tamanho 100 e conjunto de validação de tamanho 50. A arquitetura da rede neural foi baseada no trabalho de Nelson et al. [2017] e em experimentos empíricos para calibração dos parâmetros.

5.5.4 Execução da rede LSTM

Para cada ação, foram realizadas 10 execuções da rede neural e calculada a média aritmética dos resultados. Isso se deve ao fato de que os pesos associados as entradas da rede são inicializados de forma aleatória e consequentemente cada execução produz um resultado diferente.

Este procedimento foi realizado duas vezes: ao se utilizar como entrada os atributos originais (séries atrasadas em até 4 *lags*) e ao utilizar as 4 novas séries correspondentes as combinações lineares dos atributos originais. Para abreviação, no restante do trabalho o primeiro caso será referenciado como **LSTM1** e o segundo caso como **LSTM2**.

5.6 Análise das métricas de classificação no conjunto de teste

Nesta etapa é analisado o desempenho do modelo de classificação ao se utilizar como entrada os valores passados das séries e também ao utilizar os novos atributos criados a partir destes valores. Para comparação, foi utilizado um classificador aleatório e comparadas as medidas de Precisão, Revocação e F1 de ambas as classes (Altas e Baixas). Também foram registrados os valores da Acurácia, sendo esta a principal medida de desempenho nesta etapa. Isso se deve ao fato de que a Acurácia é capaz de medir o desempenho de ambas as classes. A Precisão das classes poderia ser analisada exclusivamente se não fosse o fato de que em vários conjuntos de treino e teste as classes estarem desbalanceadas, enviesando o modelo a favor de uma delas. Por esse motivo também foram registrados os valores relativos a porcentagem de classe de Altas no conjunto de treinamento das séries (%To) e a porcentagem da classe de Altas no conjunto de teste (%Te). As medidas de Revocação e F1 em conjunto com a Precisão são utilizadas para verificar se as predições não foram muito enviesadas para somente uma das classes.

As Tabelas 5.13, 5.14 e 5.15 demonstram os resultados encontrados da média de 10 execuções, relativos a LSTM1, LSTM2 e o classificador aleatório, respectivamente. Todas as 3 Tabelas estão ordenadas pelos valores da Acurácia.

Ao se comparar os valores da Acurácia das Tabelas 5.13 e 5.14 nota-se que em ambos os casos o desempenho do modelo foi similar, não há uma diferença significativa dos melhores valores. Na Tabela 5.13 o maior valor encontrado foi igual a 0.61 e 15 ações obtiveram valores acima de 0.53. Já para a Tabela 5.14, o maior valor encontrado

Tabela 5.13. Resultado das métricas de Acurácia (A), Precisão (P), Revocação (R) e F1 utilizando a rede LSTM1. Também é demonstrada a porcentagem de valores da classe de Altas encontradas no conjunto de Treino (%To) e a porcentagem de valores da classe de Altas encontradas no conjunto de Teste (%Te). O símbolo (▲) demonstra um valor de Acurácia maior que o previsor aleatório enquanto que o símbolo (▼) demonstra uma valor menor.

				Altas			Baixas		
Ações	**%To**	**%Te**	**A**	**P**	**R**	**F1**	**P**	**R**	**F1**
WEGE3	52.41	47.62	▲0.61	0.56	0.73	0.63	0.69	0.49	0.57
ABEV3	53.01	53.57	▲0.59	0.61	0.6	0.6	0.61	0.58	0.55
PETR4	61.44	58.33	▲0.59	0.58	0.59	0.58	0.61	0.59	0.58
CPFE3	52.40	69.05	▲0.58	0.67	0.56	0.59	0.48	0.65	0.53
EGIE3	53.61	53.57	▲0.58	0.57	0.65	0.61	0.61	0.5	0.55
ENBR3	54.81	58.33	▲0.58	0.67	0.42	0.5	0.54	0.82	0.65
ITSA4	55.42	58.33	▲0.57	0.63	0.54	0.57	0.51	0.6	0.54
SUZB5	45.18	52.38	▲0.57	0.55	0.62	0.58	0.61	0.51	0.55
BBDC4	55.42	60.71	▲0.56	0.61	0.63	0.62	0.47	0.44	0.45
NATU3	51.20	54.76	▲0.56	0.55	0.5	0.52	0.57	0.63	0.59
CPLE6	53.61	57.14	▲0.55	0.59	0.46	0.51	0.52	0.66	0.57
JBSS3	51.20	48.81	▲0.55	0.52	0.79	0.63	0.62	0.31	0.41
PCAR4	53.01	54.76	▲0.55	0.6	0.45	0.51	0.51	0.67	0.58
LAME4	58.43	46.43	▲0.54	0.48	0.51	0.49	0.6	0.57	0.57
MRFG3	46.39	64.29	▲0.54	0.7	0.34	0.46	0.46	0.88	0.6
CMIG4	50.60	48.81	▲0.53	0.53	0.51	0.51	0.52	0.55	0.52
TIMP3	55.42	54.76	▲0.53	0.56	0.38	0.44	0.51	0.72	0.59
VALE3	57.22	55.95	▲0.53	0.52	0.57	0.54	0.47	0.48	0.45
BRFS3	53.01	50	▲0.52	0.49	0.39	0.38	0.56	0.64	0.55
FIBR3	48.19	59.52	▲0.52	0.58	0.61	0.59	0.42	0.39	0.4
BRAP4	59.64	55.95	▲0.51	0.54	0.48	0.51	0.47	0.55	0.5
BRKM5	48.79	53.57	▲0.51	0.5	0.41	0.43	0.51	0.62	0.55
CSAN3	57.83	55.95	▲0.51	0.69	0.37	0.39	0.46	0.68	0.53
ELET3	60.24	51.19	▼0.51	0.47	0.49	0.48	0.55	0.53	0.53
LREN3	58.43	50	▲0.51	0.47	0.28	0.35	0.52	0.74	0.61
CCRO3	51.2	54.76	0.5	0.51	0.56	0.53	0.5	0.42	0.43
MULT3	57.23	57.14	▲0.5	0.5	0.48	0.47	0.48	0.52	0.49
BBAS3	54.22	59.52	▲0.49	0.57	0.43	0.47	0.41	0.58	0.47
GGBR4	57.23	60.71	▼0.47	0.57	0.45	0.48	0.37	0.5	0.41
VIVT4	51.2	57.14	▼0.47	0.52	0.35	0.42	0.43	0.62	0.51
BRML3	55.42	51.19	▼0.46	0.47	0.52	0.49	0.46	0.4	0.43
ITUB4	54.21	59.52	▼0.46	0.55	0.34	0.42	0.41	0.65	0.5
EMBR3	46.39	60.71	▼0.45	0.58	0.25	0.35	0.4	0.76	0.53
GOAU4	56.02	65.48	▼0.45	0.62	0.31	0.41	0.36	0.7	0.48
USIM5	54.21	54.76	▼0.44	0.42	0.2	0.27	0.44	0.74	0.55
CSNA3	54.21	55.95	▼0.43	0.47	0.27	0.3	0.39	0.65	0.47
CYRE3	53.61	55.95	▼0.42	0.41	0.33	0.36	0.42	0.54	0.47
PETR3	52.41	46.43	▼0.41	0.39	0.46	0.42	0.44	0.36	0.4

Tabela 5.14. Resultado das métricas de Acurácia (A), Precisão (P), Revocação (R) e F1 utilizando a rede LSTM2. Também é demonstrada a porcentagem de valores da classe de Altas encontradas no conjunto de Treino (%To) e a porcentagem de valores da classe de Altas encontradas no conjunto de Teste (%Te). O símbolo (▲) demonstra um valor de Acurácia maior que o previsor aleatório enquanto que o símbolo (▼) demonstra uma valor menor.

				Altas			Baixas		
Ações	**%To**	**%Te**	**A**	**P**	**R**	**F1**	**P**	**R**	**F1**
ABEV3	51.20	53.57	▲0.63	0.69	0.53	0.6	0.59	0.75	0.66
SUZB5	46.98	52.38	▲0.62	0.6	0.58	0.59	0.64	0.65	0.63
WEGE3	46.98	47.62	▲0.61	0.57	0.59	0.57	0.65	0.62	0.63
BBDC4	42.16	60.71	▲0.6	0.6	0.87	0.71	0.58	0.16	0.24
CPLE6	55.42	57.14	▲0.59	0.62	0.61	0.61	0.57	0.57	0.57
TIMP3	47.59	54.76	▲0.58	0.62	0.43	0.51	0.56	0.76	0.64
ELET3	48.19	51.19	▲0.57	0.57	0.18	0.28	0.57	0.97	0.72
MULT3	43.97	57.14	▲0.57	0.59	0.54	0.55	0.59	0.6	0.54
CMIG4	48.79	48.81	▲0.56	0.58	0.37	0.45	0.55	0.74	0.63
CPFE3	57.22	69.05	▲0.56	0.61	0.73	0.66	0.33	0.19	0.24
EGIE3	46.98	53.57	▲0.56	0.55	0.59	0.57	0.57	0.52	0.54
ENBR3	54.21	58.33	▲0.56	0.6	0.57	0.55	0.51	0.55	0.49
ITSA4	46.98	58.33	▲0.56	0.64	0.48	0.55	0.5	0.68	0.58
NATU3	51.80	54.76	▲0.56	0.54	0.77	0.64	0.6	0.29	0.39
CCRO3	54.21	54.76	▲0.55	0.55	0.63	0.54	0.6	0.44	0.46
LAME4	49.39	46.43	▲0.54	0.47	0.73	0.56	0.67	0.37	0.44
PETR4	40.96	58.33	▲0.54	0.63	0.13	0.2	0.53	1.1	0.71
FIBR3	53.01	59.52	▲0.53	0.59	0.62	0.6	0.42	0.39	0.41
GGBR4	53.01	60.71	▲0.53	0.61	0.61	0.61	0.4	0.41	0.4
CSAN3	46.38	55.95	▲0.52	0.58	0.53	0.54	0.46	0.51	0.48
JBSS3	48.19	48.81	▲0.52	0.51	0.8	0.62	0.59	0.27	0.36
BRKM5	46.98	53.57	▲0.51	0.46	0.24	0.26	0.47	0.81	0.58
PCAR4	52.41	54.76	▲0.51	0.56	0.42	0.48	0.48	0.62	0.54
VIVT4	50.62	57.14	▲0.51	0.59	0.36	0.43	0.47	0.71	0.55
BRFS3	48.79	50	▲0.5	0.45	0.36	0.4	0.53	0.63	0.57
CSNA3	47.59	55.95	▲0.5	0.54	0.81	0.64	0.25	0.11	0.14
ITUB4	55.42	59.52	▼0.48	0.55	0.4	0.46	0.43	0.6	0.49
MRFG3	62.04	64.29	▼0.48	0.64	0.27	0.38	0.42	0.87	0.57
VALE3	53.01	55.95	0.48	0.53	0.45	0.48	0.41	0.5	0.44
BRAP4	59.63	55.95	▼0.47	0.51	0.34	0.41	0.45	0.63	0.52
BRML3	54.22	51.19	▼0.47	0.47	0.64	0.54	0.44	0.29	0.35
CYRE3	50.60	55.95	▼0.47	0.48	0.47	0.47	0.46	0.48	0.46
LREN3	50.60	50	▼0.47	0.42	0.26	0.32	0.5	0.68	0.57
EMBR3	46.98	60.71	▼0.46	0.58	0.37	0.44	0.39	0.6	0.47
USIM5	54.21	54.76	▼0.46	0.45	0.25	0.3	0.45	0.71	0.54
PETR3	40.96	46.43	▼0.44	0.41	0.47	0.44	0.47	0.42	0.44
BBAS3	46.98	59.52	▼0.41	0.46	0.04	0.08	0.41	0.96	0.58
GOAU4	50.60	65.48	▼0.41	0.64	0.13	0.18	0.37	0.93	0.52

Tabela 5.15. Resultado das métricas de Acurácia (A), Precisão (P), Revocação (R) e F1 utilizando um classificador aleatório. Também é demonstrada a porcentagem de valores da classe Alta encontradas no conjunto de Teste (%Te)

			Altas			Baixas		
Ações	**%Te**	**A**	**P**	**R**	**F1**	**P**	**R**	**F1**
MRFG3	64.29	0.53	0.63	0.48	0.54	0.44	0.63	0.52
BRML3	51.19	0.52	0.52	0.5	0.51	0.53	0.55	0.54
ELET3	51.19	0.52	0.47	0.45	0.46	0.56	0.59	0.57
EMBR3	60.71	0.52	0.61	0.54	0.57	0.41	0.48	0.44
ENBR3	58.33	0.52	0.58	0.5	0.53	0.46	0.55	0.5
ITSA4	58.33	0.52	0.57	0.48	0.52	0.46	0.56	0.51
NATU3	54.76	0.52	0.51	0.47	0.49	0.53	0.59	0.55
PETR4	58.33	0.52	0.52	0.42	0.46	0.51	0.66	0.57
CMIG4	48.81	0.51	0.5	0.47	0.48	0.53	0.55	0.54
CPFE3	69.05	0.51	0.62	0.42	0.5	0.42	0.71	0.52
CYRE3	55.95	0.51	0.52	0.46	0.49	0.49	0.56	0.52
GOAU4	65.48	0.51	0.64	0.51	0.57	0.36	0.51	0.42
ITUB4	59.52	0.51	0.58	0.51	0.54	0.44	0.51	0.47
TIMP3	54.76	0.51	0.54	0.48	0.5	0.49	0.54	0.51
BBDC4	60.71	0.5	0.58	0.45	0.5	0.43	0.57	0.49
BRAP4	55.95	0.5	0.55	0.46	0.5	0.46	0.56	0.5
BRKM5	53.57	0.5	0.49	0.46	0.47	0.51	0.55	0.53
CCRO3	54.76	0.5	0.5	0.45	0.47	0.49	0.56	0.52
FIBR3	59.52	0.5	0.58	0.5	0.53	0.42	0.51	0.46
PCAR4	54.76	0.5	0.54	0.51	0.52	0.47	0.51	0.49
SUZB5	52.38	0.5	0.49	0.49	0.49	0.51	0.52	0.51
VIVT4	57.14	0.5	0.55	0.45	0.49	0.45	0.56	0.5
CPLE6	57.14	0.49	0.54	0.44	0.48	0.46	0.57	0.1
CSAN3	55.95	0.49	0.55	0.49	0.52	0.43	0.49	0.46
CSNA3	55.95	0.49	0.55	0.5	0.53	0.43	0.49	0.46
LAME4	46.43	0.49	0.44	0.46	0.45	0.55	0.53	0.54
LREN3	50	0.49	0.47	0.48	0.47	0.5	0.49	0.49
PETR3	46.43	0.49	0.45	0.47	0.46	0.52	0.51	0.51
USIM5	54.76	0.49	0.53	0.52	0.52	0.45	0.47	0.46
ABEV3	53.57	0.48	0.51	0.46	0.48	0.46	0.51	0.48
BRFS3	50	0.48	0.44	0.45	0.44	0.51	0.5	0.5
GGBR4	60.71	0.48	0.59	0.46	0.52	0.39	0.52	0.44
VALE3	55.95	0.48	0.54	0.49	0.51	0.43	0.47	0.45
BBAS3	59.52	0.47	0.56	0.46	0.5	0.38	0.48	0.42
MULT3	57.14	0.47	0.49	0.4	0.44	0.45	0.57	0.51
EGIE3	53.57	0.47	0.44	0.45	0.47	0.51	0.49	0.47
WEGE3	47.62	0.47	0.43	0.46	0.44	0.51	0.48	0.49
JBSS3	48.81	0.45	0.44	0.45	0.44	0.47	0.46	0.46

Tabela 5.16. Resultados dos intervalos de confiança para a Acurácia de um previsor aleatório, ao se variar os níveis de confiança. Na Tabela são demonstrados os valores do limite superior, limite inferior, níveis dos intervalos e número total de ações que estão acima do limite superior para LSTM1 e LSTM2

Nível de confiança	Z	Inferior	Superior	LSTM1	LSTM2
0.70	1.04	0.443	0.556	13	15
0.75	1.15	0.437	0.562	10	14
0.80	1.28	0.430	0.569	8	8
0.85	1.44	0.421	0.578	6	6
0.90	1.645	0.410	0.589	3	5
0.92	1.75	0.404	0.595	1	4
0.95	1.96	0.393	0.606	1	4

corresponde a 0.63 e 17 ações apresentaram valores acima de 0.53. É possível observar que na Tabela 5.14, 4 ações obtiveram valores acima de 0.6 enquanto que na Tabela 5.13 apenas uma ação apresentou resultado acima de 0.6. Comparando-se as duas Tabelas, apesar da Tabela 5.14 apresentar uma pequena vantagem, não houve uma diferença notável dos resultados mesmo utilizando o novo atributo que possuía valores de correlação de distância maiores que os dados originais. Ao analisar os resultados encontrados na Tabela 5.12 e comparando-os com os resultados da Tabela 5.13 e 5.14, pode-se refutar a hipótese de que quanto maior a correlação de distância entre as entradas e os valores de previsão no conjunto de treinamento, melhores as classificações do modelo. Para o modelo de classificação implementado no trabalho isso não ocorreu.

É interessante notar que há uma diferença considerável nos resultados encontrados ao se comparar os valores da Tabela 5.13 e 5.14 em relação a Tabela 5.15. Na Tabela 5.15 o maior valor de Acurácia encontrado foi igual a 0.53, chegando a ser aproximadamente até 10% pior que o melhor caso das outras Tabelas. De todas as ações analisadas, 15 ações da Tabela 5.13 e 17 ações da Tabela 5.14 apresentaram valores melhores que o melhor caso da Tabela 5.15. Os resultados dos modelos utilizando a rede neural em ambos os casos foram melhores, em termos de Acurácia, em 27 ações comparando com o classificador aleatório.

Os resultados da Tabela 5.16 demonstram o número total de ações dos modelos LSTM1 e LSTM2 que possuem Acurácia maior que o intervalo de confiança da Acurácia de um previsor aleatório, para diferentes níveis de confiança. Percebe-se que para um nível de 70% de confiança todas as ações do previsor aleatório já estão incluídas (nenhuma obteve Acurácia superior a 0.55). Pode-se afirmar que os modelos LSTM1 e LSTM2 apresentam uma superioridade estatisticamente significativa em relação ao previsor aleatório, uma vez que a probabilidade do previsor aleatório ter uma Acurácia

Tabela 5.17. Valores do rendimento financeiro da taxa SELIC e CDI. Fonte: BancoCentral [2019]

Período	Rendimento taxa SELIC	Rendimento CDI
09/2016 - 12/2016	4.34%	4.29%

maior que 0.6 é menor que 5%. Nota-se também uma pequena vantagem ao se utilizar a LSTM2 em relação a LSTM1.

Estes resultados demonstram que é possível obter modelos de predição superiores a um previsor aleatório, para pelo menos algumas das ações.

5.7 Análise do retorno financeiro no conjunto de teste

Esta etapa é responsável por analisar o retorno financeiro obtido ao se utilizar a estratégia baseada nos modelos de classificação implementados. Os resultados obtidos correspondem a média de 10 execuções para cada ação. Foram calculados os valores de: média dos retornos financeiros, valor máximo de retorno financeiro, valor mínimo de retorno financeiro e retorno financeiro acumulado de cada uma das classes (Altas, Baixas). Também foi calculado o valor do retorno financeiro total acumulado, ao utilizar ambas as classes. Os valores dos retornos financeiros acumulados foram comparados com os valores obtidos pela estratégia *Buy and Hold*. As Tabelas 5.18, 5.19 e 5.20 apresentam estes valores ordenados pela Acurácia dos modelos da rede LSTM1, LSTM2 e o classificador aleatório, respectivamente. Na Tabela 5.17 são apresentados os valores de rendimento da taxa SELIC e CDI, correspondente ao período de Setembro a Dezembro de 2016. Estes valores são utilizados como alguns dos *baselines* a fim de comparação. Vale ressaltar que para calcular o retorno financeiro acumulado foi utiliza a fórmula do retorno para n períodos: $1 + R_t(n) = \prod_{i=0}^{n-1}(1 + R_{t-i})$.

Ao se comparar os valores do retorno financeiro obtidos pela LSTM1 com os valores da estratégia de *Buy and Hold* pela Tabela 5.18, nota-se que 11 (28.95%) das 38 ações apresentaram valores superiores ao baseline. Entretanto dos 19 primeiros resultados, 11 (57.89%) obtiveram melhores resultados e dos 10 melhores, 6 (60%) também apresentaram valores maiores. Observando apenas a classe de Altas, 13 ações superaram o *Buy and Hold* e para a classe de Baixas 9 ações obtiveram retorno financeiro maior ao *baseline*. Já ao se comparar com os rendimentos da taxa SELIC e CDI, 12 ações apresentaram retornos financeiros maiores, ao observar ambas as classes. Somente observando a classe de Altas, 20 ações superaram as taxas de juros, enquanto

Tabela 5.18. Resultado dos valores percentuais (%) de: retorno financeiro do *Buy and Hold* (B&H), retorno financeiro total acumulado (RFT), média dos retornos financeiros (Média), valor máximo do retorno financeiro (Max), valor mínimo do retorno financeiro (Min), e retorno financeiro acumulado (RF) por classe, para LSTM1. Os símbolos (▲) (▼) demonstram a comparação do Retorno Financeiro entre LSTM1 e *Buy and Hold*

			Altas				Baixas			
Ações	**B&H**	**RFT**	**Média**	**Max**	**Min**	**RF**	**Média**	**Max**	**Min**	**RF**
WEGE3	▼-9.57	▲47.33	0.32	4.98	-6.34	17.06	0.75	3.15	-2.64	25.83
ABEV3	▼-10.72	▲29.58	0.22	3.17	-2.95	8.41	0.58	3.27	-1.64	19.28
PETR4	▲10.93	▼-4.76	0.09	8.27	-9.61	3.49	-0.25	4.96	-7.85	-8.53
CPFE3	▲6.01	▼3.08	0.11	1.87	-1.05	4.67	-0.06	1.15	-1.36	-1.53
EGIE3	▼-6.25	▲12.85	0.08	3.7	-4.23	3.55	0.27	4.57	-1.93	8.94
ENBR3	▲0.29	▼-6.95	-0.07	2.94	-5.22	-2.00	-0.08	3.17	-4.21	-5.12
ITSA4	▼13.31	▲14.59	0.36	4.44	-3.92	15.63	-0.01	4.86	-3.01	-1.05
SUZB5	▲23.19	▼14.65	0.36	8.4	-5.52	18.73	-0.06	3.74	-10.47	-4.38
BBDC4	▼5.02	▲6.88	0.17	3.6	-8.17	7.73	-0.03	5.49	-4.1	-1.09
NATU3	▼-19.37	▲-10.44	-0.31	5.42	-5.34	-13.47	0.09	11.76	-4.88	3.25
CPLE6	▼-4.05	▲3.49	0.05	5.47	-12.36	1.04	0.09	5.22	-6.91	2.03
JBSS3	▼3.39	▲13.02	0.2	19.07	-11.45	8.46	0.19	3.83	-6.81	3.99
PCAR4	▼7.76	▲16.24	0.41	5.76	-5.03	13.91	0.07	7.04	-5.26	1.8
LAME4	▼-19.75	▲-2.05	-0.24	7.36	-7.86	-9.15	0.21	6.88	-5.31	7.47
MRFG3	▲24.02	▼3.04	0.6	9.89	-4.44	15.66	-0.17	5.9	-6.71	-10.95
CMIG4	▲7.41	▼0.05	0.19	7.94	-7.73	6.12	-0.18	6.54	-9.63	-6.25
TIMP3	▲12.41	▼4.63	0.3	4.4	-5.63	9.18	-0.07	4.82	-4.19	-4.39
VALE3	▲80.24	▼19.52	0.38	6.44	-6.04	43.76	-0.49	5.67	-7.52	-18.37
BRFS3	▼-20.84	▲27.93	0.36	4.01	-3.3	1.73	0.56	5.24	-3.11	25.06
FIBR3	▲18.45	▼-30.45	-0.14	4.37	-8.97	-8.47	-0.82	3.4	-11.12	-24.05
BRAP4	▲107.07	▼-12.21	0.78	7.3	-6.9	34.69	-0.95	6.66	-6.65	-35.14
BRKM5	▲29.68	▼-5.80	0.29	11.45	-6.27	9.63	-0.32	5.07	-6.38	-14.30
CSAN3	▲10	▼9.54	1.14	5.25	-5.61	13.02	0	6.58	-5.35	-3.41
ELET3	▲9.5	▼-2.43	0.15	6.31	-5.61	5.83	-0.18	3.96	-7.36	-7.87
LREN3	▲-5.11	▼-10.34	-0.21	3.32	-4.87	-5.23	-0.08	3.72	-6.68	-5.73
CCRO3	▲-4.56	▼-9.22	-0.1	5.65	-8.19	-6.21	-0.09	4.7	-4.05	-3.25
MULT3	▲-2.43	▼-4.39	-0.07	4.76	-4.14	-1.11	-0.06	3.72	-5.39	-3.54
BBAS3	▲38.05	▼-19.71	0.17	7.67	-6.47	6.57	-0.58	6.03	-6.6	-25.00
GGBR4	▲41.76	▼-8.46	0.37	10.44	-6.39	15.62	-0.51	6.08	-6.89	-21.48
VIVT4	▲2.03	▼-7.65	-0.05	2.96	-3.81	-1.73	-0.1	6.74	-3.73	-6.07
BRML3	▲16.4	▼-10.58	0.1	8.27	-8.1	3.38	-0.38	2.96	-6.38	-13.61
ITUB4	▲15.28	▼-9.68	0.13	4.97	-4.99	3.86	-0.25	4.57	-3.72	-13.16
EMBR3	▲25.9	▼-14.08	0.28	3.47	-3.84	5.80	-0.31	4.44	-6.85	-18.93
GOAU4	▲63.28	▼-12.75	0.68	7.75	-7.71	22.01	-0.59	7.92	-8.4	-30.3
USIM5	▲46.22	▼-38.09	-0.19	6.41	-5.5	-1.37	-0.68	7.76	-11.3	-37.90
CSNA3	▲27.82	▼-27.51	-0.14	5.96	-6.09	-3.25	-0.62	7.1	-6.7	-25.16
CYRE3	▲23.14	▼-24.15	0.03	4.55	-4.56	0.41	-0.57	8.09	-8.46	-24.70
PETR3	▲1.99	▼-15.61	-0.07	10.6	-5.82	-5.24	-0.3	5.02	-4.93	-11.06

Tabela 5.19. Resultado dos valores percentuais (%) de: retorno financeiro do *Buy and Hold* (B&H), retorno financeiro total acumulado (RFT), média dos retornos financeiros (Média), valor máximo do retorno financeiro (Max), valor mínimo do retorno financeiro (Min), e retorno financeiro acumulado (RF) por classe, para LSTM2. Os símbolos (▲) (▼) demonstram a comparação do Retorno Financeiro entre LSTM2 e *Buy and Hold*

			Altas				Baixas			
Ações	**B&H**	**RFT**	**Média**	**Max**	**Min**	**RF**	**Média**	**Max**	**Min**	**RF**
ABEV3	▼-10.72	▲43.71	0.4	3.02	-2.38	14.41	0.47	3.32	-3.18	25.61
SUZB5	▼23.19	▲46.53	0.75	13.57	-5.52	34.10	0.25	3.74	-3.34	9.16
WEGE3	▼-9.57	▲33.49	0.28	4.98	-6.34	11.51	0.45	3.15	-3.36	19.52
BBDC4	▼5.02	▲21.01	0.21	4.24	-8.94	14.35	0.79	5.63	-2.01	5.79
CPLE6	▼-4.05	▲18.8	0.21	5.81	-12.36	8.20	0.29	5.22	-6.64	9.72
TIMP3	▼12.41	▲16.71	0.45	4.4	-5.63	15.33	0.04	4.82	-4.19	1.05
ELET3	▲9.5	▼8.81	0.93	5.56	-3.02	12.70	-0.02	5.61	-7.76	-3.48
MULT3	▼-2.43	▲0.20	0.04	6.14	-4.3	1.27	0.01	3.24	-3.86	-1.13
CMIG4	▲7.41	▼-14.75	-0.01	3.49	-8.82	-1.24	-0.21	6.3	-12.06	-13.71
CPFE3	▲6.01	▼-0.32	0.04	1.87	-1.39	2.91	-0.24	0.35	-1.34	-3.14
EGIE3	▼-6.25	▲8.78	0.05	3.87	-4.23	1.69	0.2	4.57	-1.98	6.92
ENBR3	▼0.29	▲2.94	0.02	3.83	-4.99	2.81	0.01	3.01	-2.69	0.08
ITSA4	▲13.31	▼11.30	0.38	4.48	-2.86	14.17	-0.04	5.1	-2.8	-2.51
NATU3	▼-19.37	▲-5.52	-0.16	4.88	-5.34	-11.47	0.41	11.76	-5.42	6.70
CCRO3	▼-4.56	▲4.00	0.23	5.97	-7.13	0.25	0.07	4.22	-3.08	3.37
LAME4	▼-19.75	▲-6.86	-0.23	6.74	-5.95	-11.53	0.23	8.47	-5.54	5.15
PETR4	▲10.93	▼-5.97	0.64	4.57	-3.59	4.62	-0.12	9.07	-8.77	-10.67
FIBR3	▲18.45	▼-29.63	-0.13	4.37	-8.97	-7.91	-0.8	3.4	-11.12	-23.60
GGBR4	▲41.76	▼20.87	0.63	10.44	-5.71	33.03	-0.24	6.76	-6.84	-9.29
CSAN3	▲10	▼2.23	0.2	5.63	-9.1	8.56	-0.13	4.05	-5.33	-6.17
JBSS3	▲3.39	▼1.42	0.1	19.07	-11.45	2.78	-0.04	3.83	-6.49	-1.40
BRKM5	▲29.68	▼-10.78	0.34	5.38	-2.78	7.52	-0.76	5.68	-11	-17.60
PCAR4	▲7.76	▼-4.13	0.13	6.03	-4.08	3.57	-0.12	7.23	-5.38	-7.48
VIVT4	▲2.03	▼-14.04	-0.21	2.05	-4.4	-5.25	-0.18	6.23	-3.73	-9.28
BRFS3	▼-20.84	▲27.79	0.08	4.01	-3.41	1.90	0.47	5.34	-3.34	25.34
CSNA3	▲27.82	▼-7.68	0.13	6.72	-7.02	6.64	-1.47	3.52	-6.38	-13.77
ITUB4	▲15.28	▼3.39	0.3	4.97	-3.29	11.14	-0.13	4.99	-4.12	-7.13
MRFG3	▲24.02	▼-5.47	0.51	9.89	-4.44	10.82	-0.23	5.9	-6.71	-14.71
VALE3	▲80.24	▼12.47	0.85	7.68	-6.28	39.90	-0.48	5.32	-7.39	-20.64
BRAP4	▲107.07	▼-20.52	0.9	7.3	-5.39	29.10	-0.87	7.26	-6.65	-38.51
BRML3	▲16.4	▼-1.86	0.17	8.51	-8.1	7.95	-0.35	2.92	-5.46	-9.14
CYRE3	▲23.14	▼-10.28	0.2	4.55	-4.47	9.18	-0.47	8.09	-8.46	-17.94
LREN3	▲-5.11	▼-21.09	-0.42	2.92	-4.87	-10.92	-0.19	3.57	-6.68	-11.44
EMBR3	▲25.9	▼-16.95	0.17	3.47	-3.77	3.91	-0.42	4.48	-6.85	-20.10
USIM5	▲46.22	▼-41.28	-0.39	6.53	-6.55	-4.45	-0.8	7.28	-10.84	-39.33
PETR3	▲1.99	▼-4.63	0.05	10.6	-5.82	0.67	-0.24	4.32	-4.93	-5.65
BBAS3	▲38.05	▼-32.05	0.03	1.89	-2	0.02	-0.44	6.9	-7.82	-32.07
GOAU4	▲63.28	▼-29.81	1.58	6.12	-2.75	11.11	-0.59	8.71	-8.76	-37.89

Tabela 5.20. Resultado dos valores percentuais (%) de: retorno financeiro do *Buy and Hold* (B&H), retorno financeiro total acumulado (RFT), média dos retornos financeiros (Média), valor máximo do retorno financeiro (Max), valor mínimo do retorno financeiro (Min), e retorno financeiro acumulado (RF) por classe, utilizando o classificador aleatório. Os símbolos (▲) (▼) demonstram a comparação do Retorno Financeiro entre o classificador aleatório e *Buy and Hold*

			Altas				Baixas			
Ações	**B&H**	**RFT**	**Média**	**Max**	**Min**	**RF**	**Média**	**Max**	**Min**	**RF**
MRFG3	▲24.02	▼4.78	0.37	8.5	-5.04	15.24	-0.23	5.11	-7.62	-10.46
BRML3	▲16.4	▼0.01	0.22	6.96	-7.7	9.39	-0.22	4.31	-7.66	-9.38
ELET3	▲9.5	▼-2.22	0.15	6.52	-4.56	5.96	-0.2	4.89	-7.1	-8.18
EMBR3	▲25.9	▼5.32	0.34	5.73	-4.14	15.47	-0.25	4.1	-5.14	-10.15
ENBR3	▲0.29	▼-5.42	-0.03	2.99	-4.12	-1.19	-0.1	4.54	-3.97	-4.23
ITSA4	▲13.31	▼-1.35	0.17	4.18	-4.79	7.16	-0.2	4.35	-4.19	-8.51
NATU3	▼-19.37	▲-7.65	-0.32	4.83	-11.01	-12.89	0.14	5.51	-4.71	5.24
PETR4	▲10.93	▼-11.19	0.03	7.54	-7.79	0.98	-0.28	7.28	-8.5	-12.17
CMIG4	▲7.41	▼-16.35	-0.06	5.63	-7.76	-2.15	-0.32	7.25	-12.06	-14.2
CPFE3	▲6.01	▼0.25	0.08	1.67	-1.26	3.19	-0.06	1.13	-1.54	-2.94
CYRE3	▲23.14	▼-2.03	0.3	5.45	-6.9	12.7	-0.34	7.97	-7.6	-14.73
GOAU4	▲63.28	▼-4.52	0.57	8.63	-8	24.85	-0.73	7.83	-8.09	-29.37
ITUB4	▲15.28	▼-0.1	0.2	4.47	-4.6	8.81	-0.23	4.66	-4.47	-8.91
TIMP3	▲12.41	▼-4.71	0.09	3.79	-5.01	4.25	-0.19	4.4	-4.23	-8.96
BBDC4	▲5.02	▼-1.21	0.11	3.96	-7.13	3.96	-0.13	7.08	-3.73	-5.17
BRAP4	▲107.07	▼-11.92	0.79	7.08	-6.69	30.4	-0.94	6.6	-6.85	-42.32
BRKM5	▲29.68	▼-3.45	0.24	9.54	-5.92	10.77	-0.33	5.2	-8.26	-14.22
CCRO3	▲-4.56	▼-8.21	-0.11	4.43	-6.87	-5.02	-0.08	6.87	-5.48	3.19
FIBR3	▲18.45	▼-8.95	0.12	8.03	-8.21	4.9	-0.34	6.7	-9.41	-13.85
PCAR4	▲7.76	▼1.97	0.15	5.55	-6.52	6.29	-0.11	6.7	-5.61	-4.32
SUZB5	▲23.19	▼-0.42	0.25	10.58	-5.03	10.96	-0.27	4.35	-9.01	-11.38
VIVT4	▲2.03	▼-0.17	0.05	3.42	-5.77	1.82	-0.05	5.27	-3.41	-1.99
CPLE6	▼-4.05	▲-1.02	-0.03	6.45	-8.46	-0.6	0	10.39	-5.66	-0.42
CSAN3	▲10	▼7.4	0.28	5.46	-6.55	11.42	-0.09	6.57	-5.38	-4.02
CSNA3	▲27.82	▼-8.06	0.19	6.28	-6.79	7.97	-0.39	6.51	-6.44	-16.03
LAME4	▼-19.75	▲-16.91	-0.41	5.25	-7.39	-16.8	-0.01	6.6	-6.71	-0.11
LREN3	▲-5.11	▼-14.02	-0.17	3.99	-4.83	-6.96	-0.17	4.39	-6.34	-7.06
PETR3	▲1.99	▼-6.98	0.02	8.09	-5.62	0.38	-0.17	5.66	-8.49	-7.36
USIM5	▲46.22	▼-18.84	0.26	9.95	-6.74	11.91	-0.8	6.8	-10.09	-30.75
ABEV3	▼-10.72	▲-4.44	-0.16	2.94	-3.18	-6.95	0.05	3.13	-3.08	2.51
BRFS3	▼-20.84	▲-8.31	-0.35	3.68	-4.99	-14.67	0.16	4.88	-3.75	6.36
GGBR4	▲41.76	▼9.38	0.61	9.36	-6.07	24.3	-0.33	6.1	-7.78	-14.92
VALE3	▲80.24	▼-8.95	0.55	7.62	-6.12	24.18	-0.77	6.2	-7.42	-33.13
BBAS3	▲38.05	▼-14.15	0.25	7.18	-6.64	10.45	-0.57	6.79	-6.92	-24.6
MULT3	▲-2.43	▼-6.78	-0.05	5.02	-3.89	-2.1	-0.1	4.08	-5.14	-4.68
EGIE3	▲-6.25	▼-6.65	-0.13	3.3	-4.28	-5.73	-0.03	3.63	-2.61	-0.92
WEGE3	▲-9.57	▼-10.5	-0.2	4.43	-5.66	-8.6	-0.05	3.71	-4.49	-1.9
JBSS3	▲3.39	▼-11.37	0.03	12.86	-6.76	-1.85	-0.26	9.5	-12.79	-9.59

que para a classe de Baixas somente 5 ações superaram. Nota-se um melhor retorno financeiro para a classe de Altas em comparação à classe de Baixas. Entretanto é observado que para este período testado, a maioria das ações se encontravam em uma tendência de alta dos preços, podendo ser observado tanto pela porcentagem de altas no conjunto de teste (Tabela 5.13 por exemplo), quanto pelo número de ações que obtiveram retorno positivo ao utilizar a estratégia de *Buy and Hold* (28 ações). Este fato pode ser uma das causas para essa diferença no retorno financeiro encontrado em ambas as classes.

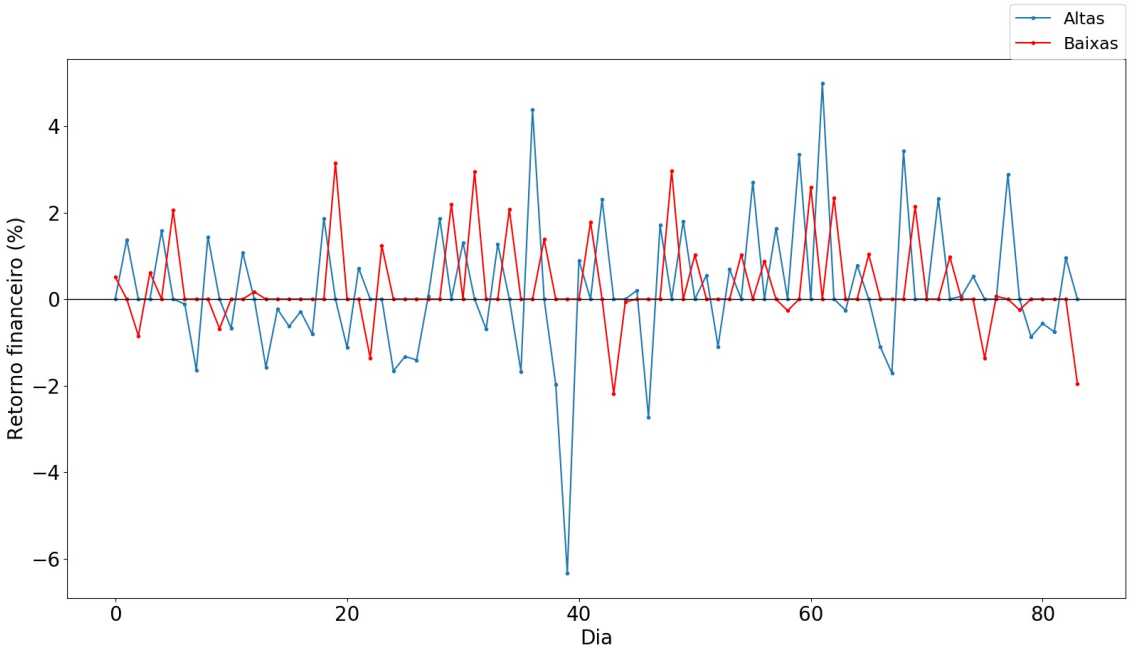

Figura 5.8. Comparação do retorno financeiro entre as classes de Altas e Baixas do conjunto de teste da rede LSTM1 de uma execução utilizando a ação WEGE3

É possível observar um comportamento similar dos resultados ao observar na Tabela 5.19. Das 38 ações analisadas, 13 (34.21%) obtiveram retorno financeiro maior que o *Buy and Hold*. Mas observando-se as 19 primeiras ações, 12 (63.15%) ações obtiveram melhores resultados que o *baseline*, e das 10 melhores 7 (70%) apresentaram retornos financeiros maiores. Analisando as classes separadamente, 15 ações da classe de Altas superaram o *Buy and Hold*, enquanto que na classe de Baixas 10 ações foram superiores. Em relação as taxas SELIC e CDI a rede LSTM2 foi superior a ambas em 12 ações. Estes resultados demonstram que apesar da maioria das ações obterem retornos financeiros menores do que a estratégia de *Buy and Hold*, quando se observa somente os melhores modelos de classificação (19 primeiras ações) é constatado que a

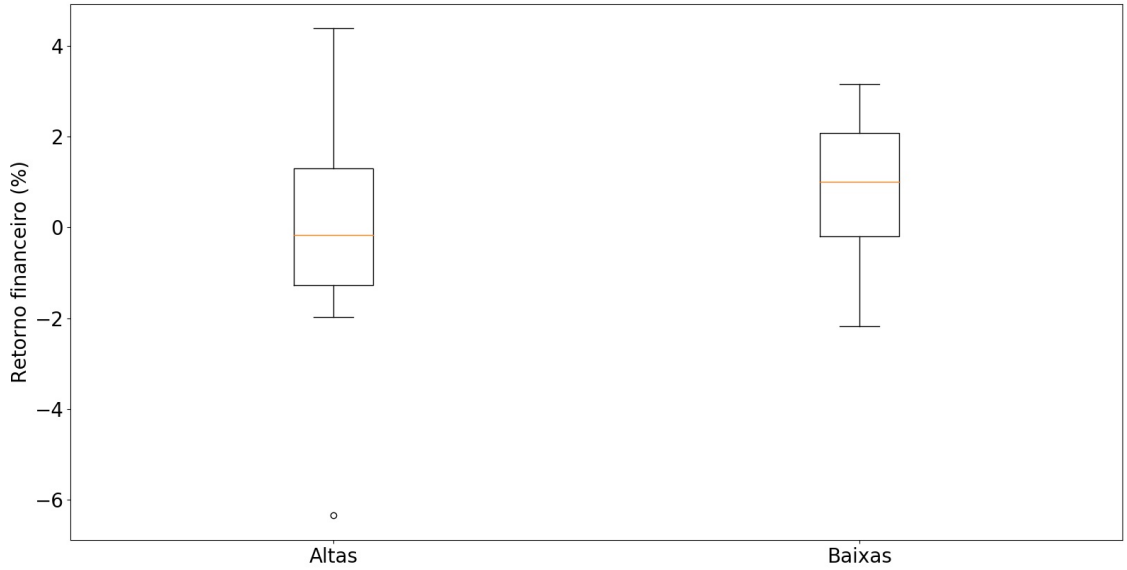

Figura 5.9. Comparação dos gráficos de Boxplot das classes de Altas e Baixas do conjunto de teste da rede LSTM1 de uma execução utilizando a ação WEGE3

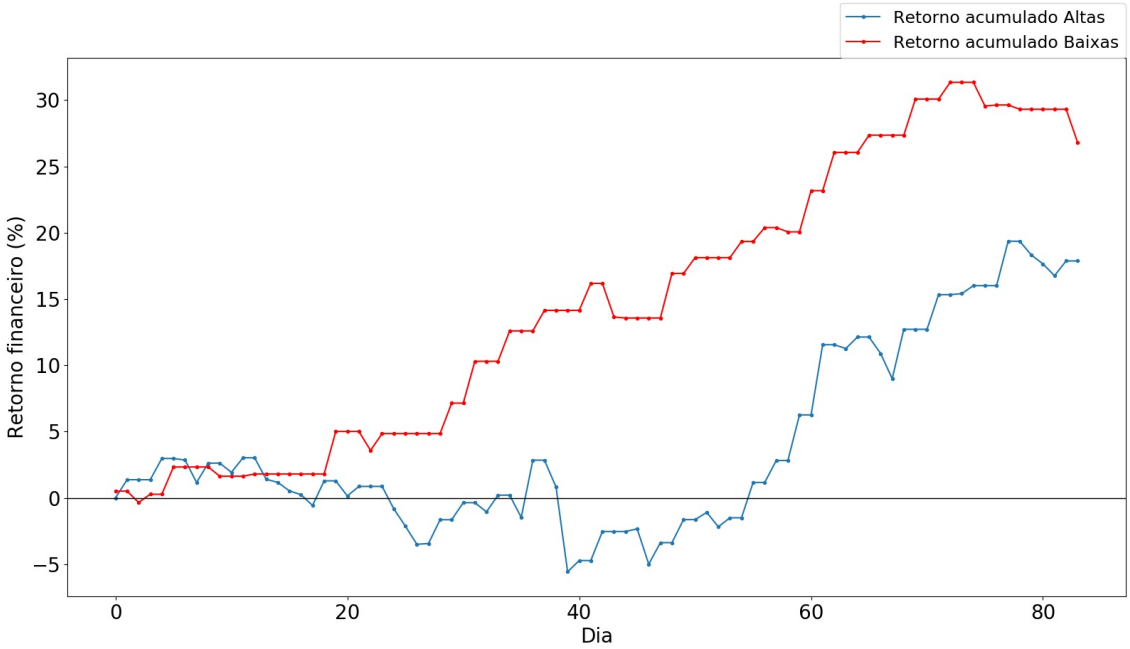

Figura 5.10. Comparação do retorno financeiro acumulado entre as classes de Altas e Baixas do conjunto de teste da rede LSTM1 de uma execução utilizando a ação WEGE3

maioria dos valores obtidos do modelo é superior aos valores dos *baselines*. Observa-se

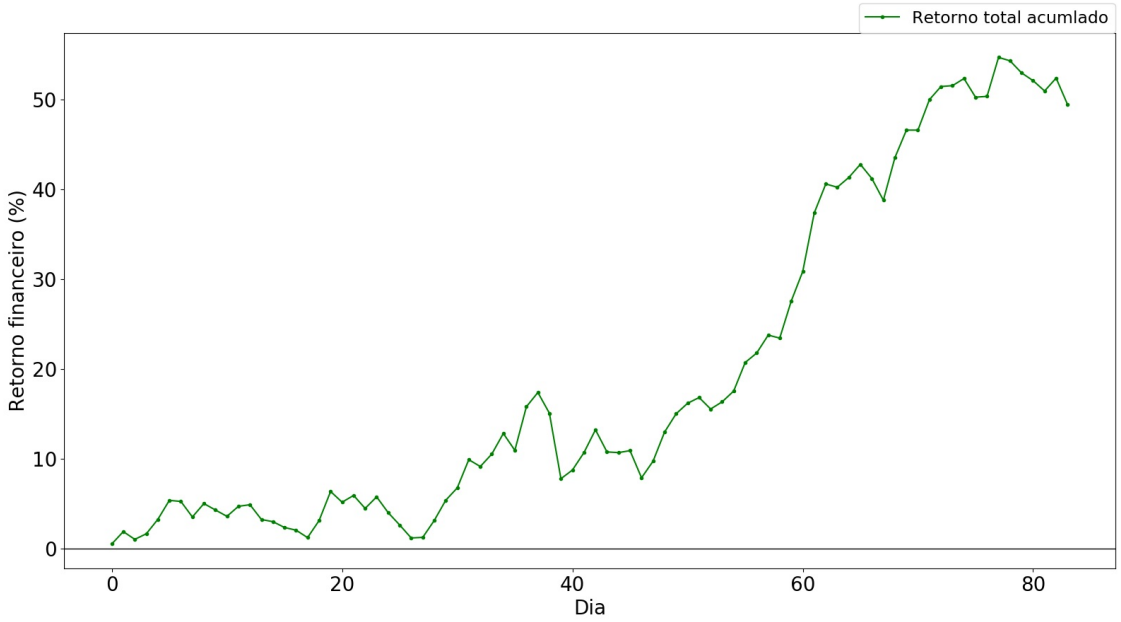

Figura 5.11. Gráfico do retorno financeiro acumulado do conjunto de teste da rede LSTM1 de uma execução utilizando a ação WEGE3

também uma pequena vantagem da LSTM2 em relação LSTM1, uma vez que o número de ações que superou o *Buy and Hold* foi maior.

Utilizando o classificador aleatório nota-se que das 38 ações, apenas 5 (13.15%) foram superiores a estratégia de *Buy and Hold*. Das 19 primeiras ações, somente uma (5.26%) obteve retorno financeiro maior que o *Buy and Hold*. Em relação à taxa SELIC e CDI o previsor aleatório obteve retornos financeiros melhores em 4 ações. Para testar a diferença nos resultados da LSTM1 e LSTM2 em relação ao previsor aleatório foi calculado o teste Z de diferença de proporção. Neste caso a proporção foi calculada como o número de previsões que superou os *baselines Buy and Hold*, taxa SELIC e CDI. A Tabela 5.21 demonstra os resultados dos p-valores obtidos pelos testes. Os p-valores podem ser interpretados como a probabilidade das proporções não serem diferentes. Claramente há uma diferença significante entre LSTM1 e LSTM2 em relação ao previsor aleatório.

É interessante notar também que as 3 Tabelas demonstram uma relação direta entre o valor da média das classes com seu respectivo retorno financeiro, e uma assimetria dos retornos financeiros das classes de Altas e Baixas. Em todos os casos o número de retornos financeiros negativos da classe de Baixas foi superior ao da classe de Altas.

Os resultados demonstram que a maioria dos melhores modelos tanto da LSTM1 quanto da LSTM2 obtiveram retornos financeiros superiores aos *baselines*: classificador

Tabela 5.21. Resultado dos p-valores para o teste Z de diferença de proporção

	LSTM1 X Aleatório	**LSTM2 X Aleatório**
B&H	0.159	0.058
SELIC	0.048	0.048
CDI	0.048	0.048

aleatório, estratégia de *Buy and Hold*, rendimento da taxa SELIC e CDI. Entretanto, o restante das ações não apresentou bons retornos financeiros. Novamente percebe-se uma pequena vantagem ao utilizar a LSTM2 em relação a LSTM1.

As Figuras 5.8, 5.9, 5.10 e 5.11 ilustram respectivamente os resultados dos retornos financeiros de cada operação, da comparação dos gráficos de Boxplot de ambas as classes, da comparação do retorno financeiro acumulado de cada classe e do retorno financeiro total acumulado, para uma única execução da rede LSTM1 utilizando a ação WEGE3. Foram previstas 54 operações da classe de Altas com Precisão de 0.56 e retorno financeiro de 17.87%, e 30 operações da classe de Baixas com Precisão de 0.7 e retorno financeiro de 26.79%. A Acurácia do modelo foi igual a 0.61 e retorno financeiro total igual a 49.45%. A distribuição da classe de altas possui média igual a 0.32%, desvio padrão de 1.9%, valor máximo de 4.98% e mínimo de -6.34%. Já a classe de Baixas apresentou uma média de 0.8%, desvio padrão de 1.46%, valor máximo de 3.15% e mínimo de -2.18%. Nota-se claramente um bom resultado obtido.

Mesmo obtendo um modelo de classificação razoavelmente satisfatório na etapa anterior é necessário avaliar o retorno financeiro do modelo. A modelagem do problema em valores discretos de duas classes (Altas e Baixas) acaba perdendo informação em relação à magnitude dos valores originais. Todos os valores de uma mesma classe são considerados iguais e isso pode gerar alguns problemas. Valores extremos classificados incorretamente podem superar vários valores menores classificados corretamente e ao final pode-se obter um retorno financeiro negativo ou insatisfatório. É interessante observar que se todos os valores fossem iguais este problema não iria ocorrer. Vale ressaltar então que a modelagem do problema é melhor aplicada nas ações que possuem baixa volatilidade. Para exemplificar esta situação foi escolhida a ação PETR4 que na Tabela 5.13 foi a terceira melhor.

Seguem os resultados de uma única execução do modelo de classificação LSTM1 sobre a ação PETR4. Foram previstas 37 operações da classe de Altas com Precisão de 0.65 e retorno financeiro de -2.28%, e 47 operações da classe de Baixas com Precisão de 0.64 e retorno financeiro de -14.86%. Percebe-se o contraste entre um valor alto de Acurácia igual a 0.64 com retorno financeiro total negativo de -16.80%. A distribuição da classe de Altas possui média igual a -0.02%, desvio padrão de 2.86%, valor máximo

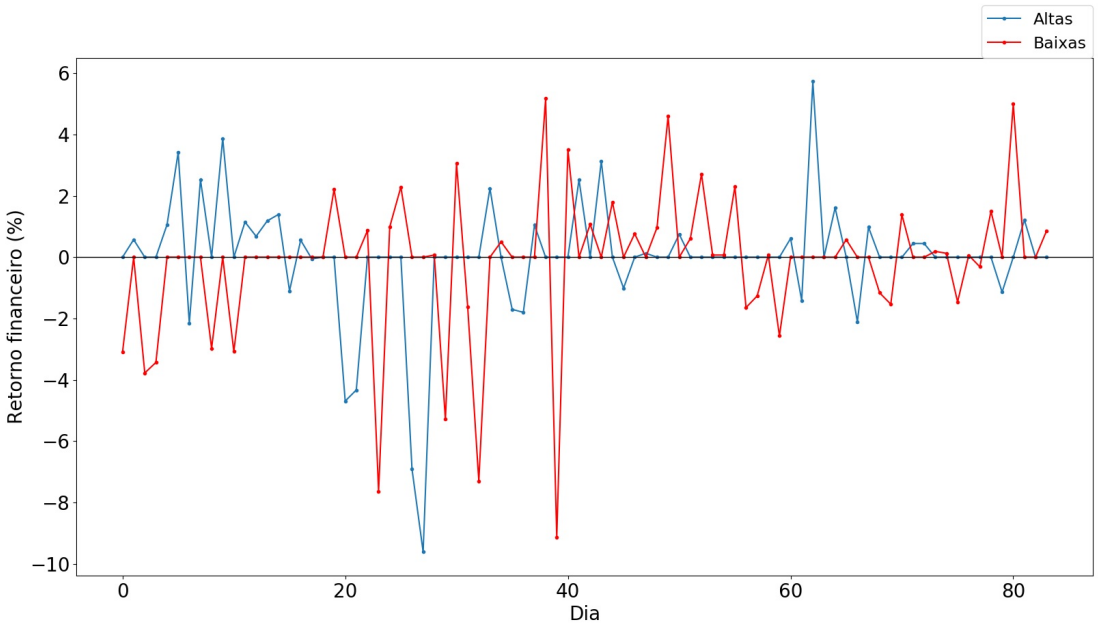

Figura 5.12. Comparação do retorno financeiro entre as classes de Altas e Baixas do conjunto de teste da rede LSTM1 de uma execução utilizando a ação PETR4

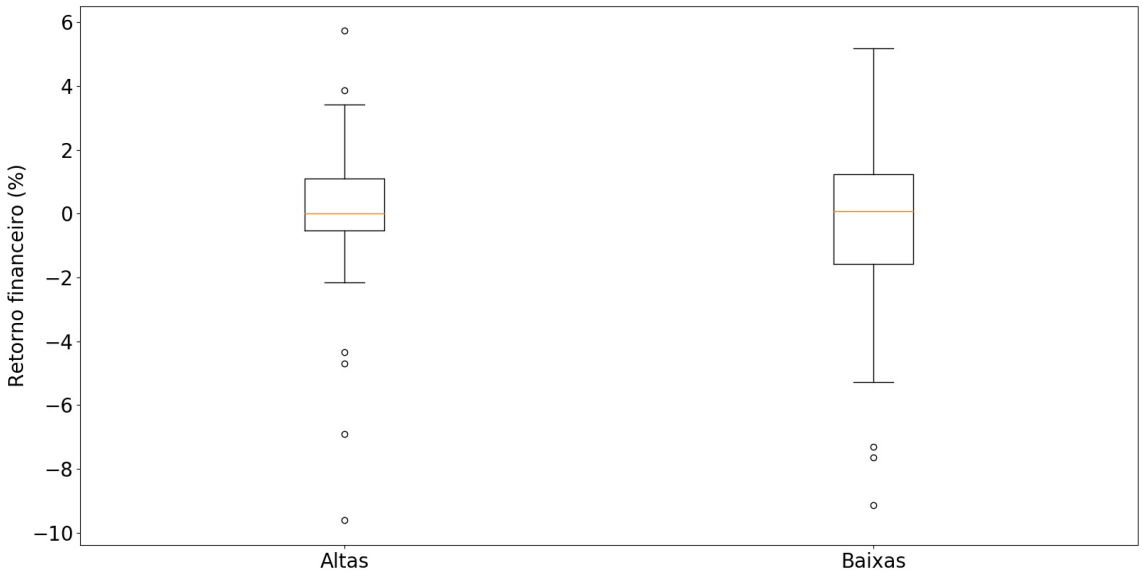

Figura 5.13. Comparação dos gráficos de Boxplot das classes de Altas e Baixas do conjunto de teste da rede LSTM1 de uma execução utilizando a ação PETR4

de 5.73% e mínimo de -9.61% . Já a classe de Baixas apresentou uma média de -0.3%, desvio padrão de 3%, valor máximo de 5.17% e mínimo de -9.14%.

As Figuras 5.12 a 5.15 ilustram os resultados dos retornos financeiros de cada

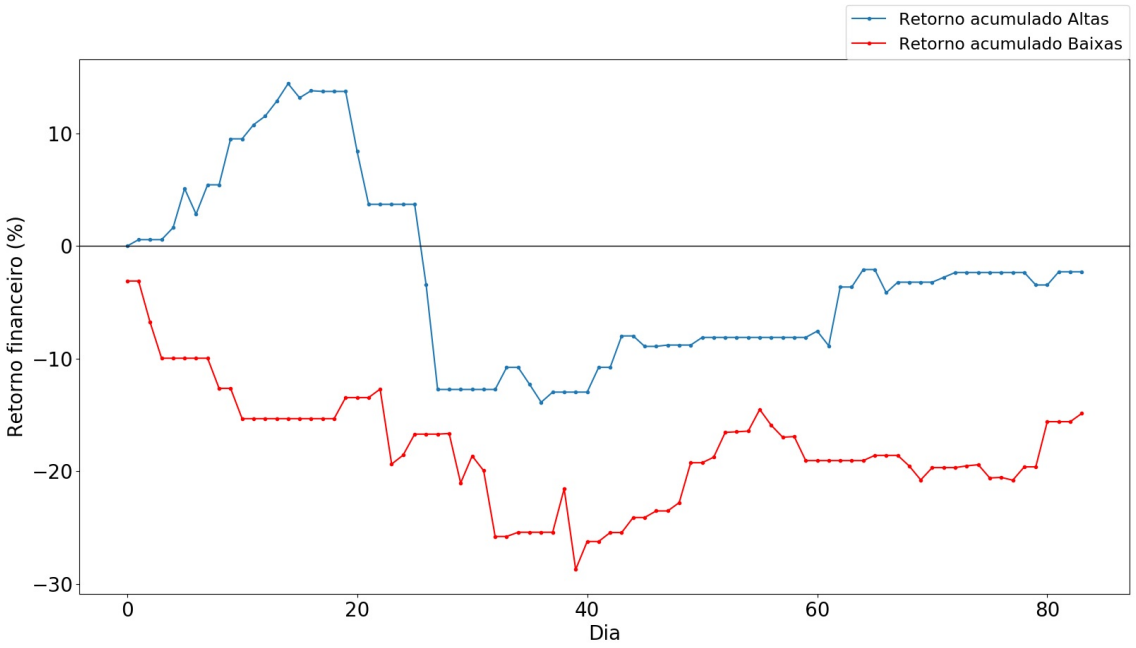

Figura 5.14. Comparação do retorno financeiro acumulado entre as classes de Altas e Baixas do conjunto de teste da rede LSTM1 de uma execução utilizando a ação PETR4

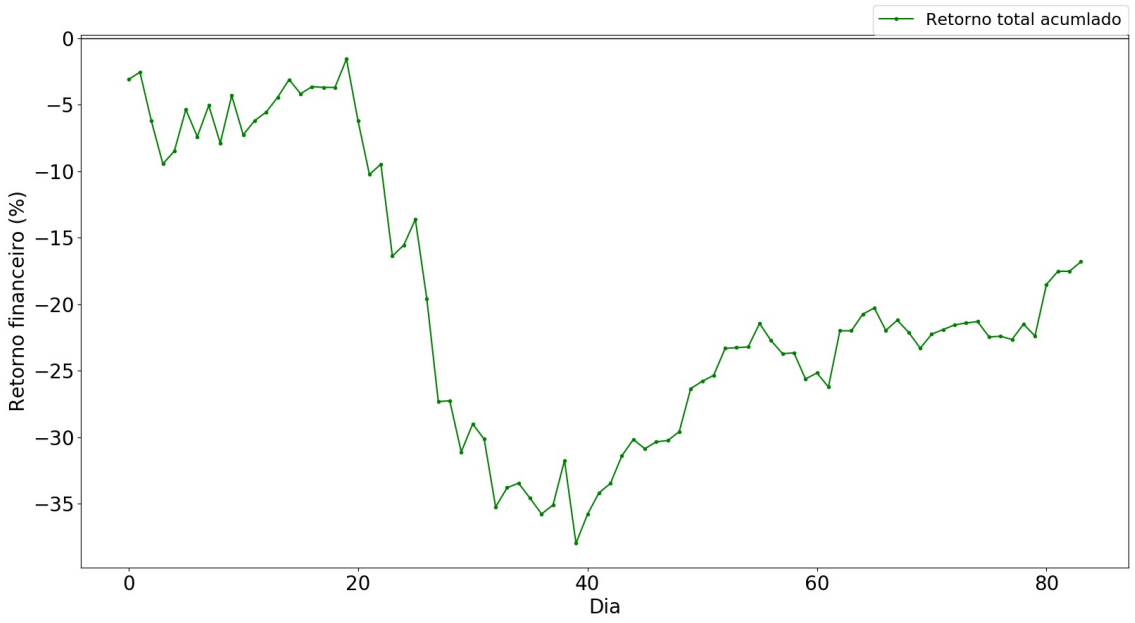

Figura 5.15. Gráfico do retorno financeiro acumulado do conjunto de teste da rede LSTM1 de uma execução utilizando a ação PETR4

operação, da comparação dos gráficos de Boxplot de ambas as classes, da comparação do retorno financeiro acumulado de cada classe e do retorno financeiro acumulado total respectivamente. Percebe-se nitidamente que apesar da maioria das operações terem sido previstas de forma correta, as operações incorretas obtiveram valores muito elevados comparados ao restante.

Os resultados analisados para a ação PETR4 demonstram que obter um modelo de classificação razoavelmente bom não é suficiente para garantir retornos financeiros positivos para o investidor.

5.7.1 Combinação da estratégia de operação com *Stop Loss*

As Figuras 5.16 a 5.19 ilustram os resultados da mesma execução da estratégia para a ação PETR4 mas utilizando-se *Stop Loss* com valor igual -4.00% para classe de Altas e de -4.05% para a classe de Baixas. Houve uma melhora no retorno financeiro de -2.28% para 6.67% para a classe de Altas, e de -14.86% para -7.42% para a classe de Baixas. A distribuição da classe de Altas teve um aumento na média para 0.2%, desvio padrão de 2.3%, o valor máximo se manteve em 5.73% e uma diminuição do valor mínimo para -4.8%. Já a distribuição da classe de Baixas apresentou uma diminuição também na média para -0.13%, diminuição no desvio para 2.65%, o valor máximo se manteve igual em 5.17% e uma diminuição no valor mínimo para -7.98%. Percebe-se que neste caso o uso do *Stop Loss* ajudou a estratégia a obter perdas menores contribuindo para aumentar o retorno financeiro acumulado de ambas as classes.

Vale ressaltar que escolher os valores limites de *Stop Loss* não é uma tarefa fácil. Nem sempre esta estratégia funciona como o esperado. Caso o valor seja muito elevado, pode acontecer dos preços nunca atingirem este limite e a estratégia funciona como se não houvesse o *Stop Loss*. Se o valor for muito baixo, o limite será alcançado todas as vezes e a estratégia deixará de ganhar nas vezes que não era necessário o *Stop Loss*. Mesmo um valor intermediário razoável poderá apresentar erros. Por exemplo, há casos em que ao longo do dia o preço está em uma tendência de queda (ou subida), e a partir de um momento sofre uma reversão de tendência.

Para a implementação da estratégia com *Stop Loss*, foram utilizados os valores intra-diários do preço de fechamento das ações, com periodicidade de 15 minutos. Os valores de *Stop Loss* para cada ação foram definidos a partir dos resultados do modelo já treinado aplicado ao conjunto de treino. Apesar de ser contra intuitivo, o conjunto de treino representa os únicos dados disponíveis antes de qualquer operação de previsão. Para cada ação, a rede neural foi treinada e realizou a previsão do conjunto de treino 10 vezes. A partir das previsões foi calculada a média dos resultados para estimar as

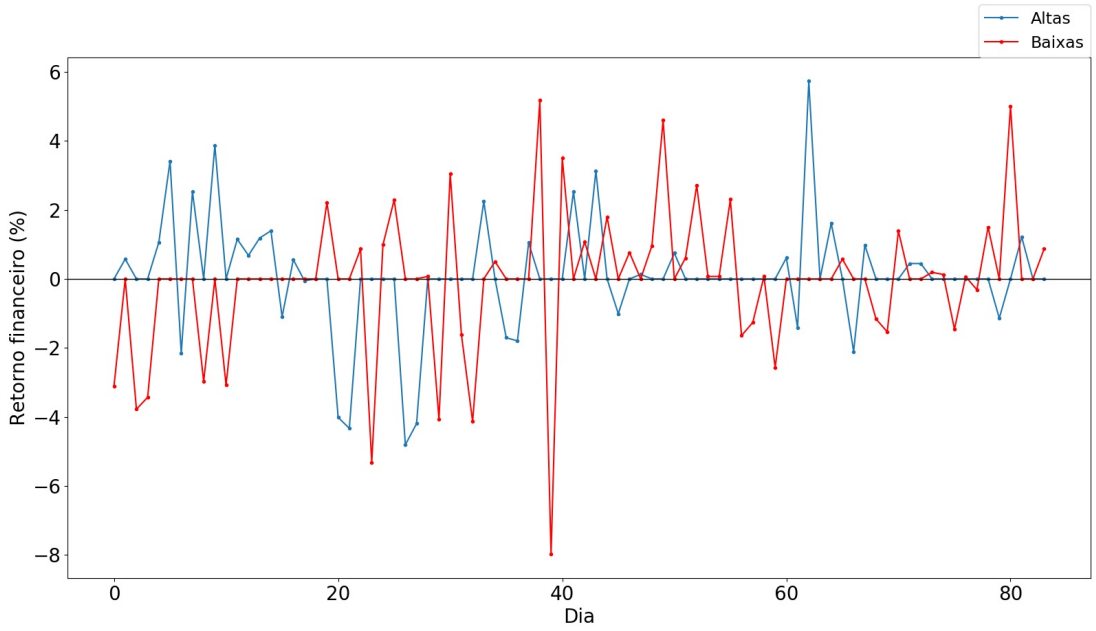

Figura 5.16. Comparação do retorno financeiro entre as classes de Altas e Baixas do conjunto de teste da rede LSTM1 de uma execução utilizando a ação PETR4 em conjunto com *Stop Loss*

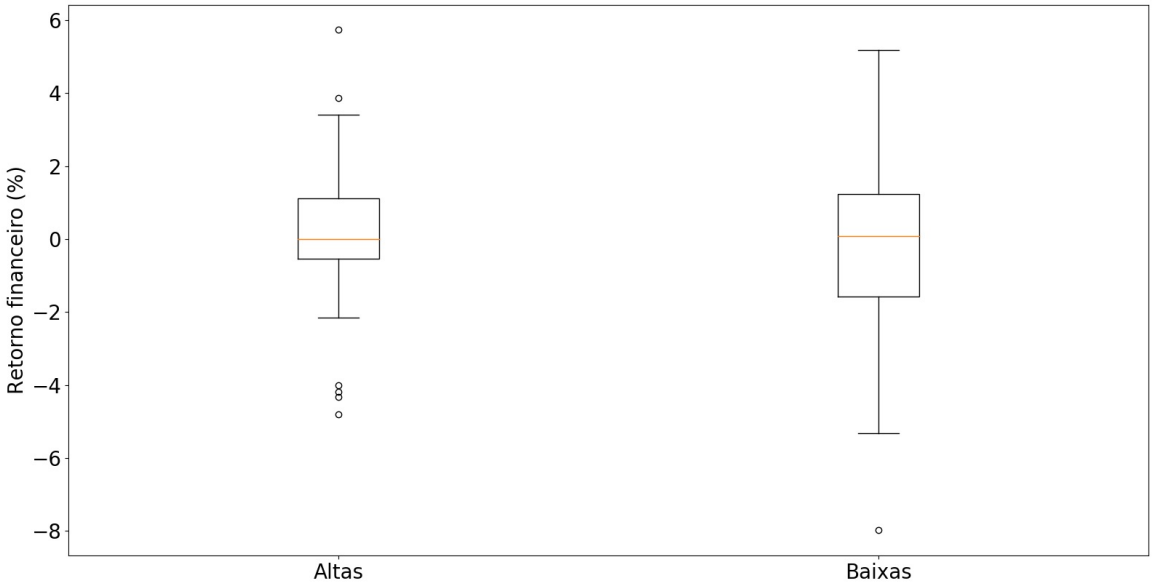

Figura 5.17. Comparação dos gráficos de Boxplot das classes de Altas e Baixas do conjunto de teste da rede LSTM1 de uma execução utilizando a ação PETR4 em conjunto com *Stop Loss*

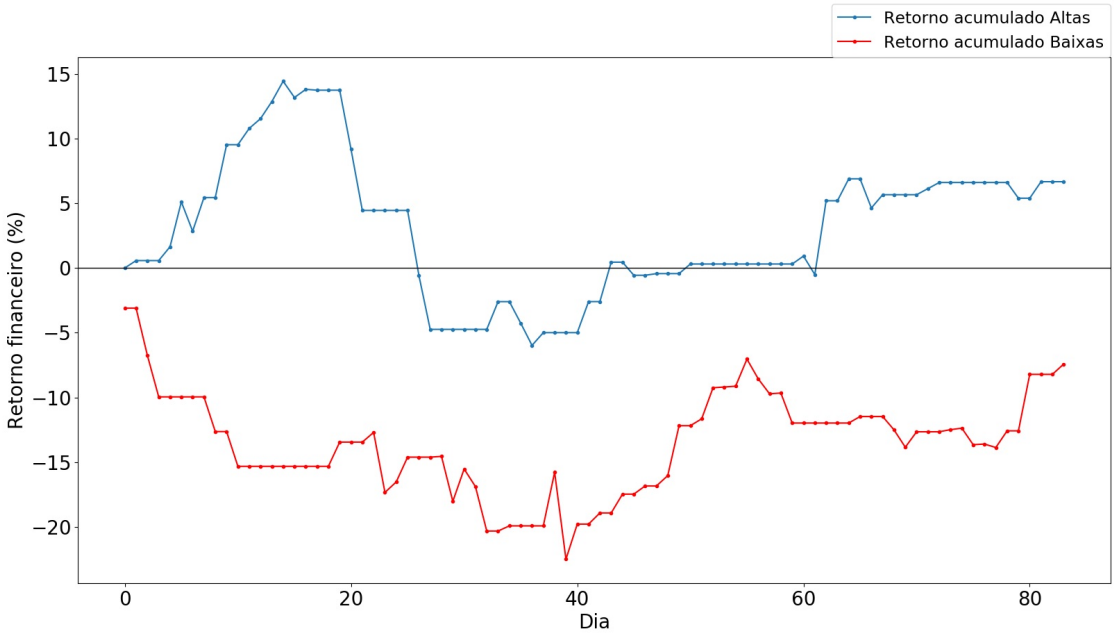

Figura 5.18. Comparação do retorno financeiro acumulado entre as classes de Altas e Baixas do conjunto de teste da rede LSTM1 de uma execução utilizando a ação PETR4 em conjunto com *Stop Loss*

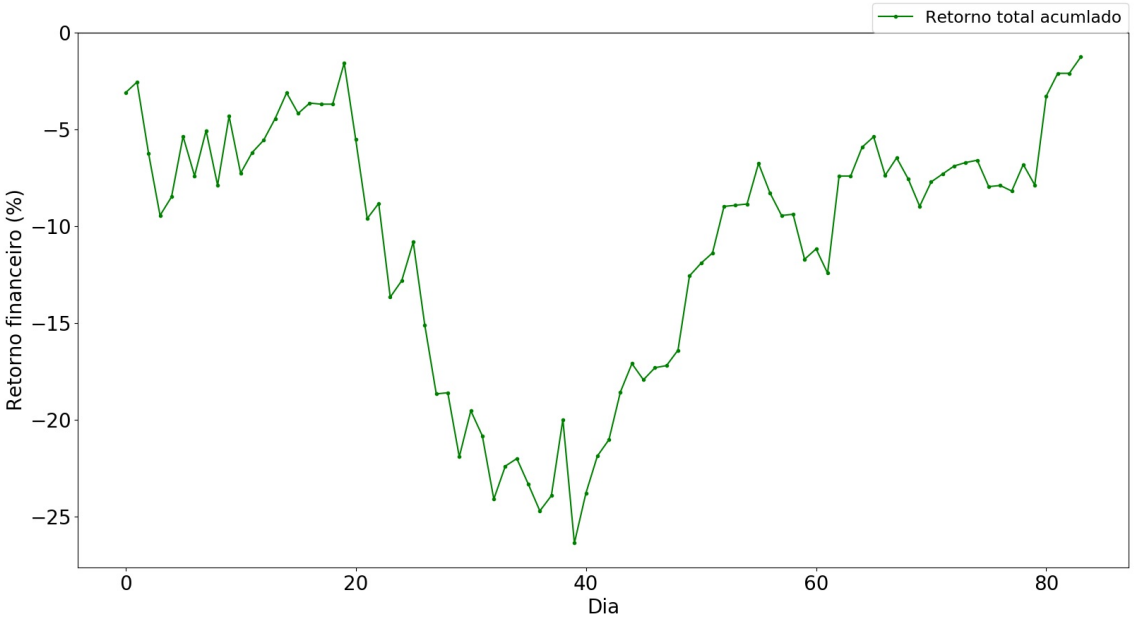

Figura 5.19. Gráfico do retorno financeiro acumulado do conjunto de teste da rede LSTM1 de uma execução utilizando a ação PETR4 em conjunto com *Stop Loss*

distribuições de perdas de ambas as classes. Dessa maneira, esperando que o modelo já treinado tenha um desempenho de certa forma similar ao utilizar o conjunto de teste e treino, é analisada a distribuição das perdas do modelo, tanto para a classe de Altas quanto para a classe de Baixas. A Tabela 5.22 demonstra os valores médios obtidos após 10 execuções para cada ação. Foram calculados os valores da média e desvio padrão das distribuições de perdas, para ambas as classes em relação ao conjunto de treino.

Foram escolhidos os seguintes valores de *Stop Loss* com base nestes resultados: média, média menos um quarto do desvio padrão, média menos três quartos do desvio padrão, média menos um desvio padrão, média menos dois desvios padrões, média menos três desvios padrões. Estes valores foram escolhidos com base na regra de Chebyshev que afirma que para qualquer distribuição de probabilidades pelo menos 0% dos valores estarão no intervalo de [média + desvio, média - desvio], 75% estarão no intervalo de [média + 2 * desvios, média - 2 * desvios], e 88.89% dos valores estarão no intervalo de [média + 3 * desvios, média - 3 * desvios]. Os valores foram definidos da seguinte forma:

- $v1$: média,
- $v2$: média - (0.25 * desvio padrão),
- $v3$: média - (0.75 * desvio padrão),
- $v4$: média - (1 * desvio padrão),
- $v5$: média - (2 * desvio padrão),
- $v6$: média - (3 * desvio padrão).

Em seguida, com os valores definidos por meio do conjunto de treino, foram comparados os retornos financeiros da estratégia original com a combinação de *Stop Loss* no conjunto de teste, ao se variar os valores de parada. Caso ao longo dos dias do conjunto de teste o retorno do preço de fechamento do *candle* intra-diário de 15 em 15 minutos, fosse maior que os valores de *Stop Loss*, a ordem era executada automaticamente não sendo necessário esperar até o final do dia para operar. As Tabelas 5.23 e 5.24 demonstram os resultados obtidos para a média de 10 execuções do retorno financeiro acumulado, ao se utilizar os valores de *Stop Loss* para cada uma das 38 ações. São demonstrados os resultados separados tanto para a classe de Altas quanto para a classe Baixas ao utilizar a LSTM1. As Tabelas demonstram os valores

Tabela 5.22. Valores da média e desvio padrão (em percentual (%)) das distribuições de perdas obtidos pela aplicação dos modelos já treinados utilizando o conjunto de treinamento. São demonstradas na Tabela a média da classe de Altas (M_A), desvio padrão da classe de Altas (SD_A), média da classe de Baixas (M_B) e desvio padrão da classe de Baixas (SD_B), para ambas LSTM1 e LSTM2.

	LSTM1				**LSTM2**			
Ações	**M_A**	**SD_A**	**M_B**	**SD_B**	**M_A**	**SD_A**	**M_B**	**SD_B**
ABEV3	-1.01	0.71	-0.99	0.81	-0.94	0.63	-0.92	0.79
BBAS3	-2.84	3.79	-2.86	2.77	-2.57	2.2	-2.52	2.39
BBDC4	-1.61	1.31	-2.21	2.4	-1.43	1.32	-4.71	3.77
BRAP4	-2.87	1.97	-3.76	3.06	-2.86	1.95	-3.8	3.1
BRFS3	-1.01	1.08	-1.41	1.24	-1.31	1.6	-1.34	1.08
BRKM5	-1.79	2.17	-2.53	2.85	-1.71	1.89	-1.93	1.68
BRML3	-2.2	1.64	-2.32	1.5	-2.18	1.55	-2.04	1.48
CCRO3	-1.7	1.41	-1.98	1.81	-1.47	1.2	-1.69	1.18
CMIG4	-2.39	1.79	-3.27	2.48	-2.05	1.57	-3.12	2.58
CPFE3	-1.15	1.1	-1.75	1.9	-1.24	1.17	-1.49	1.61
CPLE6	-2	1.5	-2.64	1.9	-1.71	1.3	-2.7	1.78
CSAN3	-1.37	1.01	-1.59	1.27	-1.48	1.17	-1.59	1.15
CSNA3	-3.88	3.33	-4.98	4.47	-3.84	3.11	-5.82	5.78
CYRE3	-1.49	1.19	-2.27	2.2	-1.53	1.14	-2.41	2.33
EGIE3	-1.11	0.8	-1.2	0.84	-1.05	0.8	-1.12	0.92
ELET3	-2.3	1.65	-3.26	2.95	-2.3	1.67	-3.12	2.87
EMBR3	-2.32	2.7	-1.63	1.02	-2.33	2.94	-1.6	1.07
ENBR3	-1.47	1.11	-1.46	0.95	-1.29	0.87	-1.11	0.77
FIBR3	-2.32	2.43	-2.56	2.45	-2.25	2.4	-2.57	2.52
GGBR4	-3.39	2.56	-4.25	3.21	-3.08	2.36	-3.49	2.78
GOAU4	-3.39	2.65	-4.7	3.88	-3.32	2.23	-4.78	3.72
ITSA4	-1.76	1.18	-1.96	1.77	-1.61	1.23	-1.92	1.79
ITUB4	-1.5	1.24	-2.04	1.99	-1.43	1.17	-2.03	2.17
JBSS3	-2.42	1.91	-2.45	2.91	-2.41	1.92	-2.16	2.03
LAME4	-2.16	1.61	-2.46	1.52	-2.11	1.63	-2.32	1.53
LREN3	-1.79	1.27	-1.88	1.15	-1.75	1.25	-1.82	1.09
MRFG3	-1.08	0.97	-1.75	1.67	-1.08	1	-1.72	1.57
MULT3	-1.26	0.82	-1.64	1.35	-0.27	0.12	-1.62	1.38
NATU3	-2.06	1.28	-2.66	2.33	-2.05	1.54	-2.61	2.04
PCAR4	-1.53	1.43	-2.04	1.53	-1.67	1.37	-1.84	1.41
PETR3	-3	2.37	-3.58	2.68	-2.96	2.39	-3.4	2.76
PETR4	-3.41	2.38	-3.34	2.84	-2.86	1.77	-3.12	2.63
SUZB5	-2.09	1.92	-2.17	2.3	-2.09	1.62	-1.65	1.7
TIMP3	-2.02	1.39	-1.83	1.65	-2	1.44	-1.76	1.58
USIM5	-4.61	3.58	-5.53	5.29	-5.87	4.15	-5.13	5.62
VALE3	-3.11	2.46	-3.4	2.5	-2.96	2.31	-2.8	2.53
VIVT4	-1.42	1.2	-1.8	1.21	-1.47	1.35	-1.71	1.22
WEGE3	-1.69	2.83	-1.94	1.4	-1.71	3.12	-1.84	1.35

originais da estratégia, o ganho máximo e perda máxima ao se utilizar *Stop Loss*, e os valores de retorno financeiro obtidos.

Ao se analisar os resultados da Tabela 5.23, pode-se verificar que em 25 ações foi possível aumentar o retorno financeiro da estratégia. Entretanto, em 32 ações a utilização incorreta da escolha do valor de *Stop Loss* prejudicou o retorno financeiro. O ganho máximo foi igual a 13.54%, obtido na ação CCRO3 ao utilizar o valor $v2$, e a perda máxima foi de 18.18% para a ação ELET3, ao se utilizar o valor $v1$. Os resultados demonstram que há ações em que os valores de *Stop Loss* foram bem escolhidos já que sempre houve uma melhora no retorno financeiro (CCRO3 e CPLE6), enquanto que em outros casos os valores escolhidos só prejudicaram os retornos financeiros da estratégia (ABEV3 e NATU3 por exemplo). Pode-se notar que dentre os 6 valores escolhidos aqueles que obtiveram os maiores número de ganhos financeiros foram $v3$ e $v4$ com 17 ações apresentando uma melhora no retorno financeiro. Já o valor $v6$ foi o que apresentou o menor número de ganhos com apenas 8 ações.

Os resultados da Tabela 5.24 demonstram um comportamento similar em que 22 ações apresentaram ganhos, enquanto que 29 apresentaram perdas com a utilização incorreta dos valores de *Stop Loss*. O maior ganho registrado foi de 10.77% para a ação FIBR3, enquanto que a maior perda foi de 14.34% para a ação TIMP3. Novamente percebe-se casos em que os valores escolhidos sempre apresentaram melhoras (BRML3 e EMBR3) e casos em que todos os valores escolhidos apresentaram perdas (ABEV3, MRFG3, PCAR4). O valor que obteve o maior número de ganhos foi $v3$ com 17 ações, enquanto que $v6$ foi o menor com apenas 6 ações.

As Tabelas 5.25 e 5.26 demonstram resultados similares ao se utilizar a rede LSTM2 para a classe de Altas e Baixas respectivamente. Ao analisar a classe de Altas, em 24 ações foi possível aumentar o retorno financeiro enquanto que em 28 ações a escolha incorreta dos valores de *Stop Loss* prejudicou os retornos financeiros. O maior ganho obtido foi de 14.72% para a ação CPLE6, e a maior perda de 14.71% para a ação WEGE3. Novamente o valor que obteve o maior número de ganhos foi $v3$ com 18 vezes, enquanto que $v6$ foi o valor que obteve os menores números, somente 10. Já para a classe de Baixas, 23 ações melhoraram os resultados enquanto que em 26 ações a escolha incorreta dos valores de *Stop Loss* diminuiu os valores. O maior ganho foi de 12.39% para FIBR3 e a maior perda de 18.40% para TIMP3. Os valores $v2$ e $v3$ foram os responsáveis pelos maiores números de ganhos com 17 vezes enquanto que $v6$ foi o menor com apenas 7 vezes.

A análise dos resultados das Tabelas 5.23 a 5.26 demonstra que é possível diminuir as perdas do modelo de predição, entretanto não é trivial a escolha dos valores corretos de *Stop Loss* para que isso aconteça. Um mesmo valor pode diminuir algumas perdas

Tabela 5.23. Resultados dos Retornos Financeiros (em percentual (%)) ao se combinar a estratégia do modelo de classificação com a utilização de *Stop Loss* para a classe das Altas. São demonstrados os valores da estratégia sem a utilização de *Stop Loss* (Sem SL), o valor do ganho máximo ao utilizar *Stop Loss* (GMax), o valor da perda máxima (PMax) e os respectivos valores de retorno financeiro ao se variar o valor do *Stop Loss*. Os símbolos (▲) (▼) demonstram ganho e queda do Retorno Financeiro respectivamente.

				Retorno financeiro com *Stop Loss*					
Ações	**Sem SL**	**GMax**	**PMax**	**v1**	**v2**	**v3**	**v4**	**v5**	**v6**
ABEV3	8.41	-0.51	8.15	▼2.16	▼0.26	▼2.41	▼1.28	▼3.37	▼7.9
BBAS3	6.57	9.7	0.5	▲16.27	▲13.92	▲8.38	▼6.07	6.57	6.57
BBDC4	7.73	3.96	0.17	▼7.56	▲10.72	▲10.76	▲9.43	▲11.69	▲10.59
BRAP4	34.69	3.64	3.96	▲38.33	▲38.21	▼33.18	▼30.73	34.69	34.69
BRFS3	1.73	1.43	2.4	▼-0.67	▼1.61	▲3.16	▲2.44	▲1.83	▲1.96
BRKM5	9.63	0.82	8.39	▼8.67	▼5.49	▼5.41	▼1.24	▲10.45	9.63
BRML3	3.38	2.78	2.72	▼0.66	▲3.94	▲6.16	▲3.91	▲5.89	▲4.03
CCRO3	-6.21	13.54	-2.4	▲7.09	▲7.33	▲4.51	▲2.5	▲0.44	▲-3.81
CMIG4	6.12	5.36	1.67	▼5.56	▲9.78	▲11.48	▲9.72	▲7.72	▼4.45
CPFE3	4.67	0	3.74	▼2.69	▼2.16	▼0.93	▼2.47	4.67	4.67
CPLE6	1.04	12.28	-7.04	▲11.01	▲10.82	▲13.32	▲13.32	▲9.89	▲8.08
CSAN3	13.02	-0.21	2.36	▼11.22	▼12.81	▼11.19	▼10.66	▼11.62	▼12.6
CSNA3	-3.25	0	6.77	▼-10.02	▼-9.69	▼-3.74	▼-4.46	-3.25	-3.25
CYRE3	0.41	-1.42	7.16	▼-6.75	▼-5.35	▼-3.23	▼-3.78	▼-1.01	▼-1.82
EGIE3	3.55	0.42	7.3	▼-3.75	▼-1.38	▼2.6	▼1.99	▲3.9	▲3.97
ELET3	5.83	0	18.18	▼-12.35	▼-12.04	▼-1.12	▼-2.22	5.83	5.83
EMBR3	5.8	0	3.6	▼2.2	▼5.14	▼4.97	▼4.25	5.8	5.8
ENBR3	-2	2.27	1	▼-3	▼-3	▼-2.15	▲0.27	▼-2.35	▲-1.86
FIBR3	-8.47	8.85	0	▲-1.88	▲0.38	▲-0.17	▲-0.7	▲-6.78	-8.47
GGBR4	15.62	1.75	6.24	▼9.38	▲17.37	▼12.49	▼15.14	15.62	15.62
GOAU4	22.01	2.17	0	▲23.49	▲23.07	▲24.18	▲23.68	22.01	22.01
ITSA4	15.63	1.3	0.12	▲16.04	▲16.14	▲16.09	▲16.93	▲15.95	▼15.51
ITUB4	3.86	3.44	1.54	▼2.32	▼3.66	▲7.3	▲6.42	▲4.51	▼3.22
JBSS3	8.46	0	12.75	▼-0.23	▼-4.29	8.46	▼7.88	▼6.25	▼4.16
LAME4	-9.15	2.34	5.71	▼-14.86	▼-10.14	▲-7.85	▲-9.08	▲-6.81	▼-9.34
LREN3	-5.23	4.96	0	▲-1.45	▲-0.27	▲-3.38	▲-3.98	▲-4.82	-5.23
MRFG3	15.66	3.95	11.16	▼4.53	▼4.5	▼15.31	▲19.61	▲17.74	15.66
MULT3	-1.11	-0.43	9.02	▼-8.59	▼-8.04	▼-10.13	▼-8.01	▼-5.07	▼-1.54
NATU3	-13.47	-1.96	6.29	▼-17.56	▼-18.19	▼-19.76	▼-16.47	▼-15.88	▼-15.43
PCAR4	13.91	3.86	3.97	▲17.77	▲15.8	▼12.35	▼9.94	▲14.64	▲14.22
PETR3	-5.24	2.14	0.01	▲-3.1	▲-3.35	▲-4.03	▼-5.25	-5.24	-5.24
PETR4	3.49	6.05	0.01	▲7.52	▲9.54	▲7.39	▲7.23	▲5.03	▼3.48
SUZB5	18.73	0	0	18.73	18.73	18.73	18.73	18.73	18.73
TIMP3	9.18	0	4.85	▼6.62	▼8.78	▼4.37	▼4.33	▼8.32	9.18
USIM5	-1.37	0.9	0.57	▲-0.47	▼-1.94	-1.37	-1.37	-1.37	-1.37
VALE3	43.76	9.45	1.04	▲53.21	▲47.52	▲47.28	▲44.42	▼42.72	43.76
VIVT4	-1.73	-0.16	5.51	▼-7.24	▼-6.48	▼-5.92	▼-6.77	▼-1.89	▼-3.2
WEGE3	17.06	2.83	17.02	▼0.035	▼9.98	▲19.89	▲18.27	17.06	17.06

Tabela 5.24. Resultados dos Retornos Financeiros (em percentual %) ao se combinar a estratégia do modelo de classificação com a utilização de *Stop Loss* para a classe das Baixas. São demonstrados os valores da estratégia sem a utilização de *Stop Loss* (Sem SL), o valor do ganho máximo ao utilizar *Stop Loss* (GMax), o valor da perda máxima (PMax) e os respectivos valores de retorno financeiro ao se variar o valor do *Stop Loss*. Os símbolos (▲) (▼) demonstram ganho e queda do Retorno Financeiro respectivamente.

				Retorno financeiro com *Stop Loss*					
Ações	**Sem SL**	**GMax**	**PMax**	**v1**	**v2**	**v3**	**v4**	**v5**	**v6**
ABEV3	19.28	-0.02	10.51	▼9.98	▼8.77	▼16.64	▼15.95	▼19.26	▼19.23
BBAS3	-25	7.16	0	▲-17.84	▲-21.08	▲-24.17	▲-24.2	-25	-25
BBDC4	-1.09	3.13	0	▲2.04	▲-0.1	▲-0.95	-1.09	-1.09	-1.09
BRAP4	-35.14	0	3.32	▼-35.66	▼-38.46	▼-36.3	▼-37.29	-35.14	-35.14
BRFS3	25.06	0	12.15	▼12.91	▼18.94	▼22.14	▼21.8	▼24.56	25.06
BRKM5	-14.3	0.47	3.41	▼-17.71	▼-17.6	▲-13.83	▼-14.82	▲-13.95	▲-14.2
BRML3	-13.61	2.76	-0.32	▲-12.59	▲-10.85	▲-11.76	▲-11.62	▲-13.29	▲-13.29
CCRO3	-3.25	0	0.61	▼-3.49	▼-3.51	▼-3.36	▼-3.86	▼-3.31	-3.25
CMIG4	-6.25	3.45	0.34	▲-5.19	▲-2.8	▲-5.23	▼-6.59	▲-3.85	▲-5.51
CPFE3	-1.53	0	0	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53
CPLE6	2.03	2.09	7.38	▼-5.35	▼-5.23	▲4.12	▲3.32	▼0.62	▼2.02
CSAN3	-3.41	3.1	1.58	▲-0.81	▲-1.99	▲-0.31	▲-1.44	▼-4.99	▼-3.57
CSNA3	-25.16	0	2.63	▼-25.24	▼-27.79	-25.16	-25.16	-25.16	-25.16
CYRE3	-24.7	3	0	▲-21.7	▲-24.48	▲-22.41	▲-22.76	▲-23.26	-24.7
EGIE3	8.94	0.09	5.66	▼3.28	▼3.68	▼8	▼8.56	▲9.03	▼8.91
ELET3	-7.87	1.92	1.33	▼-9.2	▲-6.21	▲-5.95	▲-6.5	-7.87	-7.87
EMBR3	-18.93	5.49	-0.19	▲-13.44	▲-16.15	▲-17.29	▲-17.92	▲-18.46	▲-18.74
ENBR3	-5.12	-0.22	8.38	▼-10.25	▼-13.05	▼-13.5	▼-10.31	▼-6.87	▼-5.34
FIBR3	-24.05	10.77	-0.94	▲-13.28	▲-14.1	▲-19.42	▲-19.76	▲-21.22	▲-23.11
GGBR4	-21.48	0	3.43	▼-23.2	▼-23.01	▼-24.61	▼-24.91	-21.48	-21.48
GOAU4	-30.3	0.05	5.99	▼-31.49	▼-30.79	▼-36.29	▲-30.25	-30.3	-30.3
ITSA4	-1.05	0.07	2.39	▼-3.44	▼-1.54	▼-1.59	▲-0.98	-1.05	-1.05
ITUB4	-13.16	1.67	0.75	▲-11.49	▼-13.64	▼-13.3	▼-13.91	-13.16	-13.16
JBSS3	3.99	2.33	1.4	▼2.87	▼2.59	▲6.32	▲5.51	3.99	3.99
LAME4	7.47	0	11.05	▼-3.58	▼-1.05	▼5.66	▼7.24	▼6.58	7.47
LREN3	-5.73	1.32	2.39	▼-8.12	▼-7.53	▼-5.76	▲-4.58	▼-6.07	▲-4.41
MRFG3	-10.95	-0.34	7.89	▼-12.22	▼-13.97	▼-14.88	▼-18.84	▼-11.29	▼-11.29
MULT3	-3.54	1.23	1.87	▼-4.91	▼-5.41	▲-2.31	▲-2.9	▲-2.97	▼-3.58
NATU3	3.25	0	1.89	▼1.36	▼1.62	▼2.67	▼3.01	3.25	3.25
PCAR4	1.8	-0.12	5.25	▼1.68	▼-0.96	▼0.5	▼1.19	▼-3.45	▼1.5
PETR3	-11.06	0	0.94	▼-12	-11.06	-11.06	-11.06	-11.06	-11.06
PETR4	-8.53	5.14	0.89	▼-9.42	▲-3.39	▲-6.9	▲-7.72	▼-8.65	▼-8.53
SUZB5	-4.38	0	0	-4.38	-4.38	-4.38	-4.38	-4.38	-4.38
TIMP3	-4.39	-7.13	14.34	▼-18.73	▼-16.74	▼-16.19	▼-11.52	▼-11.63	▼-11.63
USIM5	-37.9	6.32	0	▲-31.58	▲-33.29	▲-36.89	-37.9	-37.9	-37.9
VALE3	-18.37	0.31	1.93	▲-18.13	▼-20.3	▲-18.06	▼-19.43	-18.37	-18.37
VIVT4	-6.07	2.25	1.15	▲-4.94	▲-3.82	▲-5.72	▼-7.22	▼-6.64	-6.07
WEGE3	25.83	0	7.83	▼18	▼22.04	▼22.54	▼22.94	25.83	25.83

Tabela 5.25. Resultados dos Retornos Financeiros (em percentual (%)) ao se combinar a estratégia do modelo LSTM2 de classificação com a utilização de *Stop Loss* para a classe das Altas. São demonstrados os valores da estratégia sem a utilização de *Stop Loss* (Sem SL), o valor do ganho máximo ao utilizar *Stop Loss* (GMax), o valor da perda máxima (PMax) e os respectivos valores de retorno financeiro ao se variar o valor do *Stop Loss*. Os símbolos (▲) (▼) demonstram ganho e queda do Retorno Financeiro respectivamente.

				Retorno Financeiro com *Stop Loss*					
Ações	**Sem SL**	**GMax**	**PMax**	**v1**	**v2**	**v3**	**v4**	**v5**	**v6**
ABEV3	14.41	-1.35	10.01	▼4.89	▼4.4	▼4.45	▼4.84	▼9.11	▼13.06
BBAS3	0.02	0	0.03	▼-0.01	▼-0.01	0.02	0.02	0.02	0.02
BBDC4	14.35	4.69	0.47	▼13.88	▲17.81	▲19.04	▲16.04	▲19.03	▲18.73
BRAP4	29.1	0	6.76	▼28.97	▼27.65	▼24.51	▼22.34	29.1	29.1
BRFS3	1.9	3.07	0	▲2.47	▲4.97	▲2.98	▲2.39	1.9	1.9
BRKM5	7.52	0	6.39	▼4.88	▼3.16	▼1.84	▼1.13	▼6.76	7.52
BRML3	7.95	2.63	2.29	▼5.66	▲8.97	▲10.4	▲8.62	▲10.58	▲9.12
CCRO3	0.25	12.5	-4.24	▲12.13	▲12.59	▲12.75	▲11.88	▲6.7	▲4.49
CMIG4	-1.24	6.91	-1.49	▲4.18	▲4.59	▲5.67	▲5.26	▲2.88	▲0.25
CPFE3	2.91	0	3.34	▼0.76	▼-0.43	▼0.75	▼2.91	2.91	2.91
CPLE6	8.2	14.72	-5.38	▲13.58	▲16.06	▲22.92	▲21.04	▲19.41	▲16.57
CSAN3	8.56	6.57	-1.06	▲13.9	▲15.13	▲14.06	▲14.44	▲12.03	▲9.62
CSNA3	6.64	0	7.2	▼-0.37	▼-0.56	▼5.79	▼4.78	6.64	6.64
CYRE3	9.18	1.12	12.11	▼-2.93	▼2.3	▼3.85	▼4.67	▲10.3	9.18
EGIE3	1.69	0.72	8.58	▼-6.89	▼-3.58	▼0.5	▼0.58	▲2.41	▲2.1
ELET3	12.7	0	6.03	▼8.9	▼6.67	▼9.28	▼9.02	12.7	12.7
EMBR3	3.91	0	7.4	▼-3.49	▼3.1	▼2.83	3.91	3.91	3.91
ENBR3	2.81	0.3	1.96	▼1.66	▼0.96	▲3.11	▼0.85	▲2.94	▼1.58
FIBR3	-7.91	9.99	0	▲-0.57	▲1.04	▲2.08	▲0.57	▲-6.2	-7.91
GGBR4	33.03	0	10.09	▼25.59	▼22.94	▼30.38	▼29.79	33.03	33.03
GOAU4	11.11	1.81	0	▲12.92	▲11.88	▲12.25	▲12.61	▲11.14	11.11
ITSA4	14.17	0	2.92	▼11.25	▼13.11	▼12.98	▼12.98	14.17	14.17
ITUB4	11.14	1.55	2.89	▼8.25	▼10.4	▲12.69	▲11.95	▲11.17	11.14
JBSS3	2.78	0.01	11.86	▼-4.93	▼-9.08	▲2.79	▼2.23	▼0.69	▼-1.28
LAME4	-11.53	1.65	2.91	▼-14.44	▲-11.21	▲-9.88	▼-12.12	▼-12.36	▼-14.09
LREN3	-10.92	4.66	0	▲-8.55	▲-6.26	▲-9.94	▲-10.73	▲-10.54	-10.92
MRFG3	10.82	5.26	7.68	▼3.9	▼3.14	▲11.38	▲16.08	▲13.17	10.82
MULT3	1.27	-11.59	13.19	▼-11.17	▼-11.7	▼-11.92	▼-10.32	▼-11.47	▼-11.79
NATU3	-11.47	0	8.01	▼-16.81	▼-19.48	▼-14.22	▼-12.61	▼-15.54	-11.47
PCAR4	3.57	2.26	4.25	▲5.83	▼3.48	▼1.6	▼-0.68	3.57	3.57
PETR3	0.67	2.31	2.11	▲2.98	▲1.35	▼-0.04	▼-1.44	▼-1.3	▼-1.3
PETR4	4.62	1.3	0.16	▲5.92	▲5.1	▼4.46	▲5.26	▲5.27	▲4.89
SUZB5	34.1	0	0	34.1	34.1	34.1	34.1	34.1	34.1
TIMP3	15.33	0	5.9	▼9.53	▼14.05	▼9.43	▼9.43	▼13.39	15.33
USIM5	-4.45	0.32	0.03	▼-4.48	▲-4.13	▲-4.37	▲-4.37	▲-4.37	▲-4.37
VALE3	39.9	9.08	0	▲48.98	▲45.06	▲44.68	▲40.85	39.9	39.9
VIVT4	-5.25	0.21	4.86	▼-10.11	▼-9.95	▼-8.72	▼-7.62	▼-5.96	▲-5.04
WEGE3	11.51	2.7	14.71	▼-3.2	▼6.12	▲14.21	▲12.67	11.51	11.51

Tabela 5.26. Resultados dos Retornos Financeiros (em percentual %) ao se combinar a estratégia do modelo LSTM2 de classificação com a utilização de *Stop Loss* para a classe das Baixas. São demonstrados os valores da estratégia sem a utilização de *Stop Loss* (Sem SL), o valor do ganho máximo ao utilizar *Stop Loss* (GMax), o valor da perda máxima (PMax) e os respectivos valores de retorno financeiro ao se variar o valor do *Stop Loss*. Os símbolos (▲) (▼) demonstram ganho e queda do Retorno Financeiro respectivamente.

				Retorno Financeiro com Stop Loss					
Ações	**Sem SL**	**GMax**	**PMax**	**v1**	**v2**	**v3**	**v4**	**v5**	**v6**
ABEV3	25.61	0.82	11.33	▼14.28	▼14.29	▼18.53	▼23.19	▲26.43	▼25.44
BBAS3	-32.07	10.18	0.63	▲-23.68	▲-21.89	▲-25.8	▲-28.48	-32.7	-32.07
BBDC4	5.79	0	0	5.79	5.79	5.79	5.79	5.79	5.79
BRAP4	-38.51	0	6.03	▼-38.68	▼-44.54	▼-39.61	▼-40.55	-38.51	-38.51
BRFS3	25.34	0	14.27	▼11.07	▼15.85	▼20.85	▼20.89	▼25.25	25.34
BRKM5	-17.6	3.82	0.47	▲-15.75	▼-18.07	▲-17.21	▼-17.92	▲-13.78	▲-16.16
BRML3	-9.14	2.16	0	▲-8.67	▲-6.98	▲-8.17	▲-8.65	▲-8.9	-9.14
CCRO3	3.37	0	4.34	▼0.13	▼-0.97	▼2.36	▼1.84	▼0.53	3.37
CMIG4	-13.71	3.7	-0.84	▲-12.87	▲-10.39	▲-11.32	▲-12.17	▲-10.01	▲-12.56
CPFE3	-3.14	0	0	-3.14	-3.14	-3.14	-3.14	-3.14	-3.14
CPLE6	9.72	1.38	7.99	▼4.7	▼1.73	▲11.1	▲10.56	▼8.29	▼9.71
CSAN3	-6.17	5.71	0.16	▲-1.34	▲-0.46	▲-1.01	▲-1.87	▲-4.94	▼-6.33
CSNA3	-13.77	0	0.17	▼-13.94	-13.77	-13.77	-13.77	-13.77	-13.77
CYRE3	-17.94	4.42	0	▲-13.52	▲-17.6	▲-14.89	▲-16.11	▲-16.89	-17.94
EGIE3	6.92	0	8.37	▼-1.45	▼-1.09	▼3.17	▼4.15	6.92	6.92
ELET3	-3.48	2.48	4.75	▼-8.23	▲-3.12	▲-1	▲-2.18	-3.48	-3.48
EMBR3	-20.1	6.47	-0.09	▲-13.63	▲-15.7	▲-18.21	▲-18.71	▲-19.78	-20.01
ENBR3	0.08	-0.13	7.25	▼-4.56	▼-4.84	▼-7.17	▼-6.62	▼-1.07	▼-0.05
FIBR3	-23.6	12.39	-0.84	▲-11.21	▲-13.44	▲-18.61	▲-18.61	▲-20.76	▲-22.76
GGBR4	-9.29	0	4.95	▼-13	▼-12.96	▼-11.73	▼-12.9	▼-14.24	-9.29
GOAU4	-37.89	1.8	6.28	▲-36.09	▲-37.23	▼-44.17	▲-37.79	-37.89	-37.89
ITSA4	-2.51	0	2.81	▼-5.32	▼-3.88	▼-4.48	-2.51	-2.51	-2.51
ITUB4	-7.13	3.55	0.16	▲-3.58	▲-6.78	▼-7.29	▲-7.11	-7.13	-7.13
JBSS3	-1.4	1.27	0	▲-0.52	▲-0.13	▲-0.91	▲-1.12	▲-1.1	-1.4
LAME4	5.15	0.82	2.86	▼2.9	▲5.97	▼2.29	▼4.89	▲5.25	▲5.17
LREN3	-11.44	1.92	1.07	▼-12.51	▼-12.37	▲-10.4	▲-9.52	▼-11.88	▲-10.03
MRFG3	-14.71	1.76	7.72	▼-17.6	▼-18.13	▼-20.02	▼-22.43	▲-12.95	▼-15.03
MULT3	-1.13	2.17	0.49	▲0.09	▲1.04	▲-0.76	▼-1.62	▲-1.03	▼-1.14
NATU3	6.7	1.17	0	▲7.48	▲6.86	▲7.87	▲7.38	6.7	6.7
PCAR4	-7.48	-0.7	4.12	▼-8.65	▼-8.18	▼-9.33	▼-9.41	▼-11.16	▼-11.6
PETR3	-5.65	-1.73	4.17	▼-9.82	▼-7.4	▼-7.38	▼-7.38	▼-7.38	▼-7.38
PETR4	-10.67	3.85	-0.01	▲-6.82	▲-8.01	▲-7.13	▲-8.28	▲-10.34	▲-10.66
SUZB5	9.16	0	0	9.16	9.16	9.16	9.16	9.16	9.16
TIMP3	1.05	-8.4	18.4	▼-15.56	▼-17.35	▼-11.71	▼-7.35	▼-7.47	▼-7.47
USIM5	-39.33	6.76	-0.13	▲-32.57	▲-34.58	▲-38.21	▲-38.95	▲-39.2	▲-39.2
VALE3	-20.64	0.86	3.1	▲-19.78	▼-23.52	▼-21.01	▼-23.74	▼-20.87	-20.64
VIVT4	-9.28	1.76	1.24	▼-9.98	▲-7.52	▲-8.23	▼-10.52	▼-9.84	-9.28
WEGE3	19.52	0	8.62	▼10.9	▼15.89	▼15.92	▼15.83	▼19.26	19.52

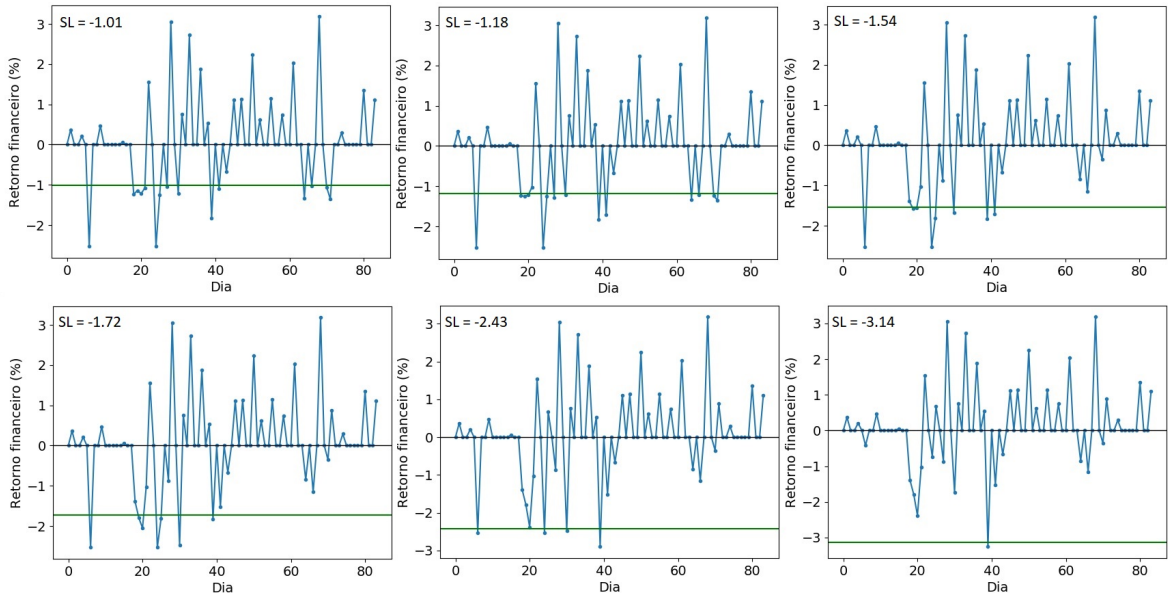

Figura 5.20. Comparação das perdas obtidas pelo modelo de predição ao se variar o valor de *Stop Loss*. Percebe-se que um mesmo valor pode diminuir algumas perdas mas aumentar outras. Na Figura é utilizado uma execução da LSTM1 para a ação ABEV3

mas aumentar outras. A Figura 5.20 demonstra esse comportamento ao se variar o valor *Stop Loss* para a ação ABEV3.

Nota-se que para cada ação é necessária uma calibração deste valor e que não houve um valor que sempre obteve ganhos. É constatado que a escolha dos valores por meio da distribuição das perdas do conjunto de treinamento pode ser uma boa escolha caso seja similar com a distribuição das perdas do conjunto de teste.

5.8 Análise dos custos de operação

Mesmo obtendo um retorno financeiro positivo é necessário avaliar o custo operacional de cada negociação, uma vez que a estratégia adotada no trabalho utiliza operações diárias. Esta análise se faz necessária já que os custos de operação podem ultrapassar os lucros obtidos pelos investidores dependendo do tamanho do lote negociado.

A Tabela 5.27 apresenta os valores relativos a média de 10 execuções dos custos de operação para cada ação, ao se variar o tamanho do lote negociado, utilizando as predições encontradas pela rede LSTM1 para as 12 melhores ações em termos de Acurácia. Os valores de custo apresentados correspondem ao custo total de todas as operações realizadas. Para cada operação foi calculado o custo a partir dos valores

Tabela 5.27. Resultados do custo de operação por ação ao se variar o tamanho do lote (L) de negociação. Para comparação é demonstrado o valor Preço Inicial (PI) de cada ação em Reais (R$) e também o valor do Retorno Financeiro em Reais (R$) para lote de uma única ação (RF, L = 1)

			Custo Operacional				
Ações	**PI**	**RF, L = 1**	**L=1**	**L=100**	**L=1000**	**L=10000**	**L=100000**
WEGE3	17,56	8,31	472,97	728,01	3046,58	26242,17	258004,40
ABEV3	19,22	5,68	472,40	670,72	2473,65	20498,77	200754,05
PETR4	13,54	-0,64	473,08	738,54	3151,82	27371,45	268613,15
CPFE3	23,96	0,73	471,12	543,21	1198,54	7575,03	71516,80
EGIE3	38,53	4,95	473,79	810,36	3870,02	34208,80	70527,45
ENBR3	13,91	-0,96	471,71	602,39	1790,40	13684,77	132471,39
ITSA4	8,04	1,17	471,43	573,78	1518,96	10813,65	103853,00
SUZB5	10,52	1,54	472,55	685,51	2621,52	21981,70	215583,40
BBDC4	29,69	2,04	474,69	899,53	4534,63	41215,36	406893,60
NATU3	32,21	-3,36	474,03	834,37	4110,12	36869,45	364442,55
CPLE6	33,86	1,18	475,27	958,00	5346,44	49443,28	488074,60
JBSS3	12,09	1,57	472,38	669,08	2457,27	20348,08	199157,99

definidos na Seção 4.7.3 do capítulo da Metodologia. O que se destaca é o fato de que para lotes de até 100 ações, a estratégia de negociação adotada não é interessante já que os valores do lucro bruto são inferiores ao custo operacional, como pode-se observar a partir dos valores de retorno financeiro (exceto para WEGE3). Como a estratégia utiliza duas operações por dia, o custo fixo destas operações acaba se sobrepondo ao valor do lucro obtido para pequenos lotes. A estratégia começa a se tornar viável com lotes de 1000 ações, como pode-se perceber ao se verificar os valores do retorno financeiro e custo operacional das ações ABEV3 e EGIE3 por exemplo. Entretanto mesmo utilizando lotes de 1000 ações o custo operacional se sobrepõe a alguns lucros brutos, como por exemplo CPFE3 e ITSA4. Fica evidente a necessidade de se utilizar um número de lotes muito grande, como dez mil, cem mil ou mais. Mesmo utilizando lotes de cem mil há casos de ações que obtiveram retornos financeiros positivos mas que ainda são insuficientes para superar os custos de operação, por exemplo CPLE6 e BBDC4.

Foi realizado um experimento buscando simular a aplicação das estratégias utilizando um lote com 10 mil ações e considerando os custos operacionais. Os resultados foram comparados com os *baselines Buy and Hold*, taxas SELIC e CDI. Foram calculados os valores do lucro bruto, custo e lucro líquido. A Tabela 5.28 demonstra os valores do retorno financeiro em Reais das taxas SELIC e CDI. Já as Tabelas 5.29 e 5.30 demonstram respectivamente a comparação dos resultados das redes LSTM1 e

LSTM2 com a estratégia de *Buy and Hold*.

Os resultados da Tabela 5.29 demonstram que 4 ações obtiveram lucro líquido positivo e em 5 de 9 ações, os custos de operação ultrapassaram o lucro bruto positivo. Utilizando a estratégia combinada com os melhores valores correspondentes de *Stop Loss*, há um melhora nos resultados e o número de ações com retorno financeiro positivo aumenta para 6, revertendo os resultados das ações CPLE6 e BBDC4. No entanto, há 4 ações (PETR4, CPFE3, SUBZ5, JBSS3) que os custos de operação continuam ultrapassando o lucro bruto. Ao comparar os resultados da LSTM1 combinada com *Stop Loss*, com os resultados dos retornos financeiros dos *baselines*, somente em três casos (WEGE3, ABE3 e EGIE3) a estratégia de operação conseguiu superar todos os *baselines*. Vale ressaltar que há casos em que a estratégia superou os resultados do *Buy and Hold* mas não das taxas SELIC e CDI, como por exemplo CPLE6.

Já os resultados da Tabela 5.30 demonstraram que ao utilizar a LSTM2 houve uma melhora em relação a LSTM1. Das 9 ações que obtiveram um lucro bruto positivo, somente em 3 ações (ENBR3, ITSA4 e JBSS3) os custos de operação foram maiores, obtendo-se ao final um lucro líquido negativo. Ao utilizar *Stop Loss* houve uma melhora em todos os casos, com destaque para a ação CPLE6. Do total de 12 ações, 5 ações (WEGE3, ABEV3, SUBZ5, BBDC4, CPLE6) obtiveram lucros líquidos melhores que todos os *baselines*. Novamente há casos de ações que foram melhores que o *Buy and Hold* mas não superaram as taxas SELIC e CDI, como por exemplo EGIE3 e NATU3.

Este experimento demonstra a complexidade em se desenvolver estratégias financeiras eficientes. Mesmo utilizando as 12 ações que obtiveram os melhores resultados em termos de classificação (para LSTM1) combinadas com os melhores valores de *Stop Loss*, mais da metade delas não conseguiu superar todos os *baselines*. Isso se deve ao fato do custo operacional ser muito elevado para a estratégia, juntamente com o valor alto das perdas encontradas pelo modelo de predição.

Tabela 5.28. Valores do retorno financeiro em Reais (R$) da taxa SELIC e CDI ao se aplicar um capital equivalente a um lote de 10000 para as ações correspondentes

Ações	Preço Inicial	Rendimento taxa SELIC	Rendimento CDI
WEGE3	17,56	7621,04	7533,24
ABEV3	19,22	8341,48	8245,38
PETR4	13,54	5876,36	5808,66
CPFE3	23,96	10398,64	10278,84
EGIE3	38,53	16722,02	16529,37
ENBR3	13,91	6036,94	5967,39
ITSA4	8,04	3489,36	3449,16
SUZB5	10,52	4565,68	4513,08
BBDC4	29,69	12885,46	12737,01
NATU3	32,21	13979,14	13818,09
CPLE6	33,86	14695,24	14525,94
JBSS3	12,09	5247,06	5186,61

Tabela 5.29. Comparação dos valores de Lucro Bruto (LB), Custo e Lucro Líquido (LL) ao se utilizar a estratégia da LSTM1, LSTM1 com *StopLoss* e o *baseline Buy and Hold*. Também é demonstrado o valor do Preço Inicial (PI) no início do investimento. Todos os valores estão em Reais (R$) e são utilizados lotes de 10000 ações. Os símbolos (▲) (▼) demonstram Lucro Líquido maior que todos os *baselines* e Lucro Líquido menor do que pelo menos um dos *baselines*, respectivamente.

		LSTM1			**LSTM1 + SL**			**B&H**		
Ações	**PI**	**LB**	**Custo**	**LL**	**LB**	**Custo**	**LL**	**LB**	**Custo**	**LL**
WEGE3	17,56	83111,48	26223,81	▲56887,67	89292,60	26242,17	▲63050,43	-16804,92	-2109,81	-18914,73
ABEV3	19,22	56852,76	20498,77	▲36353,99	56852,76	20498,77	▲36353,99	-20603,84	-2585,63	-23189,47
PETR4	13,54	-6445,04	27284,67	▼-33729,71	7880,28	27371,45	▼-19491,17	14799,22	4818,44	9980,77
CPFE3	23,96	7379,68	7575,03	▼-195,35	7379,68	7575,03	▼-195,35	14399,96	4688,58	9711,37
EGIE3	38,53	49511,05	34201,41	▼15309,64	51437,55	34208,80	▲17228,75	-24081,25	-3021,17	-27102,42
ENBR3	13,91	-9667,45	13670,51	▼-23337,96	-6760,26	13684,77	▼-20445,03	403,39	136,20	267,18
ITSA4	8,04	11730,36	10808,66	▼921,70	12687,12	10813,65	▼1873,47	10701,24	3485,57	7215,66
SUZB5	10,52	15411,80	21981,70	▼-6569,90	15411,80	21981,70	▼-6569,90	24395,88	7939,75	16456,12
BBDC4	29,69	20426,72	41112,72	▼-20686,00	41447,24	41215,36	▼231,88	14904,38	4852,64	10051,73
NATU3	32,21	-33627,24	36867,61	▼-70494,85	-33627,24	36869,45	▼-70496,69	-62390,77	-7819,44	-70210,21
CPLE6	33,86	11817,14	49230,81	▼-37413,67	60880,28	49443,28	▼11437,00	-13713,30	-1722,59	-15435,89
JBSS3	12,09	15741,18	20339,15	▼-4597,97	18509,79	20348,08	▼-1838,29	4098,51	1338,040	2760,46

Tabela 5.30. Comparação dos valores de Lucro Bruto (LB), Custo e Lucro Líquido (LL) ao se utilizar a estratégia da LSTM2, LSTM2 com *StopLoss* e o *baseline Buy and Hold*. Também é demonstrado o valor do Preço Inicial (PI) no início do investimento. Todos os valores estão em Reais (R$) e são utilizados lotes de 10000 ações. Os símbolos (▲) (▼) demonstram Lucro Líquido maior que todos os *baselines* e Lucro Líquido menor do que pelo menos um dos *baselines*, respectivamente.

		LSTM2			**LSTM2 + SL**			**B&H**		
Ações	**PI**	**LB**	**Custo**	**LL**	**LB**	**Custo**	**LL**	**LB**	**Custo**	**LL**
WEGE3	17,56	58808,44	24762,93	▲34045,51	64094,00	24781,30	▲39312,70	-16804,92	-2109,81	-18914,73
ABEV3	19,22	84010,62	22376,43	▲61634,19	85798,08	22381,94	▲63416,14	-20603,84	-2585,63	-23189,47
PETR4	13,54	-8083,38	27846,71	▼-35930,09	-1760,20	26271,02	▼-28031,22	14799,22	4818,44	9980,77
CPFE3	23,96	-766,72	6737,25	▼-7503,97	-766,72	6737,25	▼-7503,97	14399,96	4688,58	9711,37
EGIE3	38,53	33829,34	32862,62	▼966,72	36564,97	32874,80	▼3690,17	-24081,25	-3021,17	-27102,42
ENBR3	13,91	4089,54	14981,91	▼-10892,37	5466,63	14986,90	▼-9520,27	403,39	136,20	267,18
ITSA4	8,04	9085,20	10541,32	▼-1456,12	9085,20	10541,32	▼-1456,12	10701,24	3485,57	7215,66
SUZB5	10,52	48949,56	24878,18	▲24071,38	48949,56	24878,18	▲24071,38	24395,88	7939,75	16456,12
BBDC4	29,69	62378,69	44790,16	▲17588,53	76986,17	44852,21	▲32133,96	14904,38	4852,64	10051,73
NATU3	32,21	-17779,92	38075,88	▼-55855,80	-14494,50	38064,86	▼-52559,36	-62390,77	-7819,44	-70210,21
CPLE6	33,86	63656,80	53735,52	▼9921,28	123792,16	53956,49	▲69835,67	-13713,30	-1722,59	-15435,89
JBSS3	12,09	1716,78	19200,35	▼-17483,57	3203,85	19204,55	▼-16000,70	4098,51	1338,040	2760,46

5.9 Considerações finais

Neste Capítulo foram apresentados os resultados dos experimentos realizados e suas respectivas análises. Destacam-se vários pontos interessantes alcançados pelo trabalho:

- É constatado que todas as séries analisadas possuem dois fatos estilizados conhecidos na literatura: as distribuições dos log-retornos financeiros não são normais e a correlação linear presente nas séries transformadas de log-retorno finaceiro. Entretanto, ao utilizar os testes estatísticos de Ljung–Box e razão da variância, foi possível perceber que a maioria das séries de log-retornos financeiros rejeita as hipóteses i.i.d. e de passeio aleatório, indicando uma possível correlação linear que não pode ser desprezível. Ao utilizar a medida de correlação de distância foi possível demonstrar que existe uma dependência não linear significativa ao longo de todas as séries de log-retornos financeiros. Este ponto alcança o objetivo de demonstrar a dependência temporal nas séries. Entretanto os resultados obtidos também sugerem que esta dependência não é forte, o que continua sendo um desafio poder realizar uma predição nas séries de log-retornos financeiros.

- Com o intuito de explorar a dependência das séries foi proposto um método de criação de um atributo de entrada para o modelo de predição a partir dos atributos originais. O método se baseia em maximizar a correlação de distância entre a entrada do modelo e a série de valores futuros do conjunto de treino. Ao se utilizar o algoritmo *differential evolution* foi possível criar atributos mais dependentes para todas as séries de log-retornos financeiros.

- Os resultados de classificação obtidos pela rede neural LSTM tanto utilizando como entrada os atributos originais, quanto o novo atributo, demonstraram ser superiores ao previsor aleatório. Em alguns casos ao se comparar a Acurácia dos modelos, a rede neural obteve valores até 15% maiores em relação ao classificador aleatório. O desempenho da rede neural foi bom comparado a outros trabalhos na literatura. Foi possível constatar também que apesar das predições terem sido um pouco melhores ao utilizar o novo atributo em comparação com os dados originais, a diferença não é significativa.

- Os resultados da rede neural em termos de retornos financeiros foram comparados com o previsor aleatório, com a estratégia *Buy and Hold* e com as taxas de rendimentos SELIC e CDI. A rede neural novamente se manteve superior ao previsor aleatório, obtendo retornos financeiros maiores na maioria das vezes. Ao

se comparar com a estratégia de *Buy and Hold*, o *baseline* obteve retornos financeiros maiores na maioria das vezes. Entretanto vale ressaltar que esse cenário se altera ao se comparar com os 19 melhores modelos da rede neural. Fica evidente que os melhores modelos de predição da rede neural são superiores aos *baselines* tanto em termos de classificação quanto em retornos financeiros.

- A análise dos resultados financeiros demonstrou que, mesmo obtendo modelos com altos valores de Acurácia, é possível obter retornos financeiros negativos. Mesmo acertando a maioria das predições houve casos em que as perdas do classificador se sobrepuseram a maioria dos acertos. Por este motivo foi implementada juntamente com a estratégia a utilização de *StopLoss* com o objetivo de diminuir a magnitude das perdas dos modelos. A utilização destes valores de parada demonstrou ser possível diminuir as perdas, e consequentemente aumentar o retorno financeiro. Entretanto é necessário ressaltar que a escolha correta para estes valores não é uma tarefa trivial podendo em vários casos prejudicar a estratégia original. Os resultados demonstraram que não há um valor padrão que funciona para todos os casos e que para cada ação é necessária uma calibragem para se obter os melhores benefícios da utilização de *StopLoss*.

- Foi demonstrado que é necessário levar em consideração os custos de operação da estratégia. Mesmo obtendo taxas de Acurácia elevadas e retornos financeiros positivos, houve vários casos em que o custo de operação se sobrepôs ao lucro bruto obtido pela estratégia. A estratégia financeira do trabalho demonstrou não ser eficiente para lotes pequenos, sendo necessário aplicá-la em lotes com pelo menos mais de mil ações, para os melhores modelos de predição.

Capítulo 6

Conclusão

Neste trabalho foi utilizado como objeto de estudo séries reais financeiras de 38 ativos da B3. O objetivo do trabalho foi criar estratégias de investimento inteligentes e autônomas, baseadas em um modelo de predição utilizando uma rede neural LSTM. A pesquisa também tinha como objetivo caracterizar e analisar estas séries para tentar compreender suas características e propriedades específicas, de forma a evidenciar uma dependência temporal nas mesmas. A demonstração de uma dependência temporal é uma etapa imprescindível do trabalho uma vez que corrobora a utilização de dados históricos como base para modelos de predição. Caso fosse constatado que não há dependências nas séries, não seria razoável a utilização de valores passados para a predição.

Foi realizado um levantamento de trabalhos similares na literatura envolvendo três temas principais: fatos estilizados nas séries e testes de hipóteses nas mesmas, aplicação da medida de correlação de distância e função de auto correlação de distância em séries temporais, e utilização de redes neurais LSTM como modelo de predição em séries financeiras. Vários destes trabalhos serviram como fundamentação para pesquisa, incluindo técnicas, modelagem de problemas e experimentos utilizados pelos autores.

A metodologia do trabalho demonstra todas as etapas necessárias para a elaboração das estratégias desde a coleta dos dados até a validação dos resultados, tanto em termos de predição quanto em retornos financeiros.

Por meio dos experimentos e das análises dos resultados foram alcançados os objetivos propostos no trabalho: caracterização das séries demonstrando suas propriedades em comum e fatos estilizados, foi constatada a presença de dependência não linear temporal em todas as séries, buscou-se explorar esta dependência ao elaborar uma metologia de criação de um atributo mais dependente, utilizou-se um modelo de aprendizado de máquina para realizar predições com dados históricos, foi desenvolvida

uma estratégia de investimento com base nas predições, e foi possível obter resultados tanto em termos de classificação quanto em termos financeiros superiores aos *baselines*, para os melhores casos.

Ainda é um desafio entender o porquê de algumas séries apresentarem resultados de classificação tão superiores as outras, sendo que a metodologia aplicada foi a mesma. O ideal é buscar um modelo de predição que tenha um desempenho bom para qualquer série. É necessário também buscar melhorias tanto no modelo de predição para diminuir as perdas, quanto na estratégia de operação para diminuir os custos, de forma a ser possível obter estratégias financeiras viáveis e eficientes.

6.1 Trabalhos futuros

A pesquisa realizada neste trabalho pode ser estendida em alguns aspectos. Em primeiro lugar pode-se citar a utilização da rede neural LSTM em outros períodos de tempo além do ano de 2016, para avaliar se os resultados se mantêm os mesmos. O segundo ponto seria testar a rede com um conjunto de treinamento maior, além de variar os métodos de treino e teste, aplicando técnicas de validação cruzada e janelas deslizantes. No sentido de aprendizado de máquina pode-se testar também outros algoritmos de predição, como por exemplo SVM e outros tipos e arquiteturas de redes neurais. Também é interessante comparar os resultados obtidos com outros *baselines*, como por exemplo, utilização de médias móveis, tendência do dia anterior, outras implementações de algoritmos de aprendizado de máquina.

Outra abordagem interessante de se realizar é a análise estatística no conjunto de treino do algoritmo de aprendizagem. Aplicar as mesmas análises da correlação e correlação de distância, combinadas com testes estatísticos de passeio aleatório, para poder garantir que o conjunto de treinamento apresenta dependências suficientes para ser explorado. Vale ressaltar que a literatura apresenta inúmeros testes de hipóteses sobre as séries temporais, e que seria interessante um estudo sobre a relação entre estes testes e o desempenho da predição nas séries.

Pode-se estender o trabalho buscando modelar o problema de classificação em mais de duas classes, por exemplo classe de Altas, Baixas e Movimentos Laterais, ou em um problema de regressão ao invés de classificação. Essa outra forma de modelar pode ajudar a diminuir a magnitude das perdas encontradas no modelo de classes binárias. Em relação a estratégia de negociação pode-se implementar alguma que mantenha posições iguais, a fim de reduzir os custos de operação. Também é interessante combinar a estratégia com modelos de previsão de volatilidade. Dessa forma, ao operar em

períodos de baixa volatilidade espera-se evitar perdas muito significativas.

Outro ponto analisado no trabalho e que pode ser promissor como trabalhos futuros é metodologia para criação de um atributo mais dependente. Neste trabalho o novo atributo é gerado por meio de uma combinação linear dos atributos originais que busca maximizar a correlação de distância com a série de valores futuros. Pode-se pensar em outras formas de combinar os atributos originais, por meio de funções não lineares, buscando alcançar cada vez mais valores maiores de correlação de distância.

Por fim, nota-se que a área de predições em séries financeiras é bem abrangente e pode-se combinar a metodologia desenvolvida neste trabalho junto com outras estratégias. Por exemplo, pode-se realizar um estudo buscando indicadores técnicos e outros atributos (valores de volume, máximo e mínimo do dia, entre outros) para se obter melhores resultados de predição. Pode-se combinar o modelo de predição com a análise de sentimentos de notícias e opiniões de investidores buscando ter sempre melhores previsões. Também é possível estudar o comportamento de dependência inter-séries, buscando encontrar relações entre diferentes ações.

Referências Bibliográficas

Allen, F. & Karjalainen, R. (1999). Using genetic algorithms to find technical trading rules. *Journal of Financial Economics*, 51(2):245 – 271. ISSN 0304-405X.

Bachelier, L. (1900). Theory of speculation. *reprinted in P. Cootner(ed.), 1964, The Character of Stock Market Prices*, pp. 17–78.

BancoCentral (2019). <https://www.bcb.gov.br/controleinflacao/historicotaxasjuros>, acessado em 25 de abril de 2019.

Basheer, I. & Hajmeer, M. (2001). Artificial neural networks: Fundamentals, computing, design, and application. 43:3–31.

Bebarta, D. K.; Biswal, B. & Dash, P. K. (2015). Polynomial based functional link artificial recurrent neural network adaptive system for predicting indian stocks. *International Journal of Computational Intelligence Systems*, 8(6):1004–1016.

Biondo, A. E.; Pluchino, A.; Rapisarda, A. & Helbing, D. (2013). Are random trading strategies more successful than technical ones? *PloS one*, 8(7):e68344.

Cargill, T. F. & Rausser, G. C. (1975). Temporal price behavior in commodity futures markets*. *The Journal of Finance*, 30(4):1043--1053. ISSN 1540-6261.

Carvalho, A., C. (2010). Análise de viabilidade da aplicação de redes neurais no mercado financeiro. *Engenharia de Computação em Revista*, 1.

Chen, K.; Zhou, Y. & Dai, F. (2015). A lstm-based method for stock returns prediction: A case study of china stock market. Em *2015 IEEE International Conference on Big Data (Big Data)*, pp. 2823–2824. ISSN .

Chow, K. V., D. K. C. (1993). A simple multiple variance ratio test. *Journal of Econometrics*, p. 385–401.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223--236. ISSN 1469-7688.

Cooper, J. C. B. (1982). World stock markets: some random walk tests. *Applied Economics*, 14(5):515–531.

Cornell, W. B. & Dietrich, J. K. (1978). The efficiency of the market for foreign exchange under floating exchange rates. *The Review of Economics and Statistics*, 60(1):111–120. ISSN 00346535, 15309142.

Cunningham, S. W. (1973). The predictability of british stock market prices. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 22(3):315–331. ISSN 00359254, 14679876.

Das, S. & Suganthan, P. (2011). Differential evolution: A survey of the state-of-the-art. *Evolutionary Computation, IEEE Transactions on*, 15:4 – 31.

de Pádua Braga, A. (2007). *Redes neurais artificiais: teoria e aplicações*. LTC Editora. ISBN 9788521615644.

Dryden, M. M. (1970). A statistical study of u.k. share prices. *Scottish Journal of Political Economy*, 17(3):369--389. ISSN 1467-9485.

Dusak, K. (1973). Futures trading and investor returns: An investigation of commodity market risk premiums. *Journal of Political Economy*, 81(6):1387–1406.

Fama, E. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383--417.

Fama, E. F. (1965). The behavior of stock-market prices. *Journal of Business*, 38(1):34--105.

Fama, E. F. & Blume, M. E. (1966). Filter rules and stock-market trading. *The Journal of Business*, 39(1):226–241. ISSN 00219398, 15375374.

Fama, E. F. & French, K. R. (1988). Permanent and temporary components of stock prices. *Journal of Political Economy*, 96(2):246–273.

Fokianos, K. & Pitsillou, M. (2017). Consistent testing for pairwise dependence in time series. *Technometrics*, 59(2):262–270.

Fortuna, E. (2007). *Mercado Financeiro: produtos e serviços*. QualityMark.

Google (2019). <https://www.google.com.br/>, acessado em 25 de abril de 2019.

Granger, C. & Morgenstern, O. (1970). *Predictability of stock market prices*. Lexington books. Heath, Lexington/Mass.

Greff, K.; Srivastava, R. K.; Koutník, J.; Steunebrink, B. R. & Schmidhuber, J. (2015). Lstm: A search space odyssey. *CoRR*, abs/1503.04069.

Gunasekarage, A. & Power, D. M. (2001). The profitability of moving average trading rules in south asian stock markets. *Emerging Markets Review*, 2(1):17 – 33. ISSN 1566-0141.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735--1780.

Hong, Y. (1999). Hypothesis testing in time series via the empirical characteristic function: A generalized spectral density approach. *Journal of the American Statistical Association*, 94(448):1201–1220.

Hoque, H. A.; Kim, J. H. & Pyun, C. S. (2007). A comparison of variance ratio tests of random walk: A case of asian emerging stock markets. *International Review of Economics Finance*, pp. 488–502.

Huang, B.; Ding, Q.; Sun, G. & Li, H. (2018). Stock prediction based on bayesian-lstm. Em *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, ICMLC 2018, pp. 128--133, New York, NY, USA. ACM.

Huang, W.; Nakamori, Y. & Wang, S.-Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers Operations Research*, 32(10):2513 – 2522. ISSN 0305-0548. Applications of Neural Networks.

Huo, X. & Székely, G. J. (2016). Fast computing for distance covariance. *Technometrics*, 58(4):435–447.

j. Kim, K. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, 55(1–2):307 – 319. ISSN 0925-2312. Support Vector Machines.

Jegadeesh, n. (1991). Seasonality in stock price mean reversion: Evidence from the u.s. and the u.k. *The Journal of Finance*, 46(4):1427–1444.

Jennergren, L. P. & Korsvold, P. E. (1974). Price formation in the norwegian and swedish stock markets: Some random walk tests. *The Swedish Journal of Economics*, 76(2):171–185. ISSN 00397318.

Kamijo, K. & Tanigawa, T. (1990). Stock price pattern recognition-a recurrent neural network approach. Em *Neural Networks, 1990., 1990 IJCNN International Joint Conference on*, pp. 215–221 vol.1.

Kara, Y.; Boyacioglu, M. A. & Ömer Kaan Baykan (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the istanbul stock exchange. *Expert Systems with Applications*, 38(5):5311 – 5319. ISSN 0957-4174.

Karmiani, D.; Kazi, R.; Nambisan, A.; Shah, A. & Kamble, V. (2019). Comparison of predictive algorithms: Backpropagation, svm, lstm and kalman filter for stock market. Em *2019 Amity International Conference on Artificial Intelligence (AICAI)*, pp. 228–234. ISSN .

Kendall, M. G. & Hill, A. B. (1953). The analysis of economic time-series-part i: Prices. *Journal of the Royal Statistical Society. Series A (General)*, 116(1):11–34. ISSN 00359238.

Kim, J. H. & Shamsuddin, A. (2008). Are asian stock markets efficient? evidence from new multiple variance ratio tests. *Journal of Empirical Finance*, pp. 518–532.

Kirkpatrick, C. & Dahlquist, J. (2006). *Technical Analysis: The Complete Resource for Financial Market Technicians*. FT Press, first edição. ISBN 0131531131.

Levich, R. M. (1979). Analyzing the accuracy of foreign exchange advisory services: Theory andevidence. Working Paper 336, National Bureau of Economic Research.

Levine, D.; BERENSON, M. & STEPHAN, D. (2008). *Estatística: teoria e aplicações : usando o Microsoft Excel em português*. Livros Técnicos e Científicos. ISBN 9788521616344.

Lin, M. & Chen, C. (2018). Short-term prediction of stock market price based on ga optimization lstm neurons. Em *Proceedings of the 2018 2Nd International Conference on Deep Learning Technologies*, ICDLT '18, pp. 66--70, New York, NY, USA. ACM.

Lo, A. W. (2004). The adaptive markets hypothesis. *The Journal of Portfolio Management*, 30(5):15--29. ISSN 0095-4918.

Lo, A. W. & MacKinlay, A. C. (1988). Stock market prices do not follow random walks: Evidence from a simple specification test. Working Paper 2168, National Bureau of Economic Research.

Lo, A. W. & MacKinlay, A. C. (2011). *A non-random walk down Wall Street*. Princeton University Press.

Malkiel, B. G. (1999). *A random walk down Wall Street: including a life-cycle guide to personal investing*. WW Norton & Company.

Mantegna, R. N. & Stanley, H. E. (2000). *Introduction to Econophysics: correlations and complexity in Finance*. Cambridge University press.

Mcculloch, W. & Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:127--147.

Melo, B. (2012). Considerações cognitivas nas técnicas de previsão no mercado financeiro. *Universidade Estadual de Campinas*.

Mood, A. M. (1940). The distribution theory of runs. *Ann. Math. Statist.*, 11(4):367--392.

Morettin, A. P.; Toloi, C. M. (2006). *Análise de Séries Temporais*. Blucher.

Nakano, M.; Takahashi, A. & Takahashi, S. (2017). Generalized exponential moving average (ema) model with particle filtering and anomaly detection. *Expert Systems with Applications*, 73:187 – 200. ISSN 0957-4174.

Nelson, D. M. Q.; Pereira, A. C. M. & de Oliveira, R. A. (2017). Stock market's price movement prediction with lstm neural networks. Em *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 1419–1426. ISSN 2161-4407.

Neto, M. C. A.; Tavares, G.; Alves, V. M. O.; Cavalcanti, G. D. C. & Ren, T. I. (2010). Improving financial time series prediction using exogenous series and neural networks committees. Em *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. ISSN 2161-4407.

Oliveira, A. B. & Ziegelmann, F. A. (2010). Um estudo comparativo de redes neurais e modelos garch para previsão da volatilidade de séries temporais financeiras. *Simpósio Nacional de Probabilidade e Estatística*.

Pawar, K.; Jalem, R. S. & Tiwari, V. (2019). Stock market price prediction using lstm rnn. Em Rathore, V. S.; Worring, M.; Mishra, D. K.; Joshi, A. & Maheshwari, S., editores, *Emerging Trends in Expert Applications and Security*, pp. 493--503, Singapore. Springer Singapore.

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240--242.

Perfeito, P. M. (2011). Análise do comportamento do mercado de acções utilizando técnicas de data mining. Mestrado em sistemas integrados de apoio à decisão, Instituto Universitário de Lisboa, Lisboa.

Pommeranzenbaum, I. R. (2014). Redes neurais artificias na predição das principais séries do índice ibovespa e suas aplicações em sistemas automatizados de negociação. Dissertação de mestrado, Universidade Federal do Rio de Janeiro.

Praetz, P. D. (1969). Australian share prices and the random walk hypothesis. *Australian Journal of Statistics*, 11(3):123--139. ISSN 1467-842X.

Praetz, P. D. (1979). Testing for a flat spectrum on efficient market price data. *The Journal of Finance*, 34(3):645--658. ISSN 00221082, 15406261.

Rather, A. M.; Agarwal, A. & Sastry, V. (2015). Recurrent neural network and a hybrid model for prediction of stock returns. *Expert Syst. Appl.*, 42(6):3234--3241. ISSN 0957-4174.

Reis, C. E. dos; Triches, D. (2007). Seleção e composição de uma carteira de ações com base na técnica grafista. *Perspectiva Econômica*, pp. 1–26.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, pp. 365--386.

Roy, S. S.; Mittal, D.; Basu, A. & Abraham, A. (2015). Stock market forecasting using lasso linear regression model. Em Abraham, A.; Krömer, P. & Snasel, V., editores, *Afro-European Conference for Industrial Advancement*, pp. 371--381, Cham. Springer International Publishing.

Samarawickrama, A. J. P. & Fernando, T. G. I. (2017). A recurrent neural network approach in predicting daily stock prices an application to the sri lankan stock market. Em *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, pp. 1–6. ISSN .

Selvin, S.; Vinayakumar, R.; Gopalakrishnan, E. A.; Menon, V. K. & Soman, K. P. (2017). Stock price prediction using lstm, rnn and cnn-sliding window model. Em *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1643–1647. ISSN .

Shao, X.; Ma, D.; Liu, Y. & Yin, Q. (2017). Short-term forecast of stock price of multi-branch lstm based on k-means. Em *2017 4th International Conference on Systems and Informatics (ICSAI)*, pp. 1546–1551. ISSN .

Storn, R. & Price, K. (1997). Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341--359. ISSN 1573-2916.

SzéKely, G. J. & Rizzo, M. L. (2013). The distance correlation t-test of independence in high dimension. *J. Multivar. Anal.*, 117:193--213. ISSN 0047-259X.

Székely, G. J.; Rizzo, M. L. & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.*, 35(6):2769--2794.

Taylor, S. J. (1982). Tests of the random walk hypothesis against a price-trend hypothesis. *Journal of Financial and Quantitative Analysis*, 17(1):37–61.

Taylor, S. J. (2007). *Asset Price Dynamics, Volatility, and Prediction*. Princeton University Press. ISBN 978-0-691-13479-6.

Wang, J.-H. & Leu, J.-Y. (1996). Stock market trend prediction using arima-based neural networks. Em *Neural Networks, 1996., IEEE International Conference on*, volume 4, pp. 2160–2165 vol.4.

Whang, Y. J., K. J. (2003). A multiple variance ratio test using subsampling. *Economics Letters*, p. 225–230.

Working, H. (1934). A random-difference series for use in the analysis of time series. *Journal of the American Statistical Association*, 29(185):11–24. ISSN 01621459.

Wright, J. H. (2000). Alternative variance-ratio tests using ranks and signs. *Journal of Business Economic Statistics*, pp. 1–9.

Yao, S.; Luo, L. & Peng, H. (2018). High-frequency stock trend forecast using lstm model. Em *2018 13th International Conference on Computer Science Education (ICCSE)*, pp. 1–4. ISSN 2473-9464.

Zhao, Z.; Rao, R.; Tu, S. & Shi, J. (2017). Time-weighted lstm model with redefined labeling for stock trend prediction. Em *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 1210–1217. ISSN 2375-0197.

Zhou, Z. (2012). Measuring nonlinear dependence in time-series, a distance correlation approach. *Journal of Time Series Analysis*, 33(3):438–457.