Comparação entre modelos de machine learning na modelagem da volatilidade de séries financeiras

Lucas Amorim Mendes[1*]; Thiago Gentil Ramires[2]

[1] Maricá, Rio de Janeiro, Brasil, Bacharelado em Estatística.
[2] Universidade Tecnológica Federal do Paraná. Doutor em Ciências. R. Marcílio Dias, 635 - Jardim Paraíso, 86812-460, Apucarana, Paraná, Brasil.

Comparação entre modelos de machine learning na modelagem da volatilidade de séries financeiras

Resumo

As ações de empresas constituem uma classe de investimentos marcada por um elevado grau de incerteza, o que torna fundamental a mensuração dos riscos associados. No contexto da análise de risco em séries temporais financeiras, o cálculo da volatilidade se configura como uma métrica essencial. Nesse cenário, torna-se necessária a adoção de modelos específicos para estimar essa variável. A utilização de modelos de *machine learning* oferece uma alternativa eficaz, permitindo a análise desses dados sem a exigência de pré-requisitos tradicionais, como os exigidos pelos modelos estatísticos convencionais.

Este trabalho tem por objetivo comparar a volatilidade dos retornos do preço de fechamento de ações da bolsa brasileira (B3) calculada por diferentes modelos de *machine learning* com a volatilidade calculada por um modelo estatístico. Os modelos de *machine learning* utilizados foram: *Random Forest*, *XGBoost*, Regressão Multivariada e, para comparação, o modelo estatístico *Generalized Autoregressive Conditional Heteroskedasticity* (GARCH). Os modelos de *machine learning* foram treinados com os 5 últimos preços de fechamento das séries financeiras. Como método de comparação, foram utilizadas as métricas de erro absoluto percentual médio e R-Quadrado.

Palavras-chave: Random Forest; XGBoost; Regressão Multivariada; Aprendizado de Máquina; Séries Financeiras.

Comparison of Machine Learning Models in Modeling the Volatility of Financial Series

Abstract

Corporate stocks represent a class of investments characterized by a high degree of uncertainty, making the measurement of associated risks essential. In the context of risk analysis in financial time series, the calculation of volatility stands out as a crucial metric. In this scenario, the adoption of specific models to estimate this variable becomes necessary. The use of machine learning models offers an effective alternative, enabling the analysis of these data without the need for traditional prerequisites, such as those required by conventional statistical models.

This survey aims to compare the volatility of the closing price returns of stocks on the Brazilian stock exchange (B3) calculated by different machine learning models with the volatility calculated by a statistical model. The machine learning models used were: Random Forest, XGBoost, Multivariate Regression, and for comparison, the statistical model Generalized Autoregressive Conditional Heteroskedasticity (GARCH). The machine learning models were trained using the last 5 closing prices of the financial series. As comparison metrics, the Mean Absolute Percentage Error (MAPE) and R-squared were employed.

Keywords: Random Forest; XGBoost; Multivariate Regression; Machine Learning; Financial Series.

Introdução

Investimento é um tema popular por aqueles que procuram proteger e multiplicar o patrimônio. Bodie (2014) define investimento como sendo: “Investimento é o comprometimento de dinheiro ou de outros recursos no presente com a expectativa de colher benefícios futuros”. Já em da Silva (2017), é estudado a educação financeira de servidores públicos e afirma a importância da educação financeira. Em sua conclusão é afirmada a importância de ter conhecimento em como investir, e é concluído que o baixo nível de educação financeira influencia em não perceber riscos nas decisões financeiras e também nos riscos na compra de produtos no contexto de operações financeiras. Ainda no mesmo autor, é comentado sobre o aumento da expectativa de vida e como isso aumenta a importância de investir, principalmente com foco na aposentadoria.

Segundo Carminati (2013), a globalização influencia para o aumento do investimento estrangeiro no país, com isso estuda o impacto do investimento estrangeiro na economia do Brasil e conclui que um dos fatores que atrapalham este tipo de investimento é a baixa infraestrutura do país. No mesmo arquivo, o autor lembra que a infraestrutura é um dos principais pilares da economia e é apontado como um dos principais custos de investir no Brasil. Em Tavares (2006) foi mencionado que o investimento direto no exterior (IDE) aumentou consideravelmente por parte das empresas brasileiras. Isso se dá ao fato do IDE ser de grande importância para alguns fatores que afetam as empresas brasileiras, sendo alguns deles a melhoria de operações e o desenvolvimento tecnológico. Contudo, Tavares (2006) comenta que as empresas brasileiras possuem relativamente poucos investimentos no exterior, porém estes investimentos têm crescido consideravelmente desde o início e com foco nos meados dos anos noventa. O autor argumenta que uma motivação para este tipo de investimento se dá pelas deficiências do ambiente de negócios no Brasil e pelas disparidades de entrada ao mercado. Ainda no mesmo estudo, conclui-se que minimizar esses fatores poderão garantir uma melhora nos benefícios do investimento direto no exterior para o país. Podendo focar em uma política integrada de apoio internacional, porém tais políticas devem focar nos benefícios que o país terá com tal internacionalização.

Antunes (2003) disponibiliza indícios da influência das decisões de investimento das empresas nos preços das ações no mercado de capitais. Antunes (2003) argumenta que o mercado reage às expectativas das empresas para os resultados futuros do fluxo de caixa e isto é um indício que o mercado reagiu aos sinais das decisões de investimento emitidos pelas empresas. O autor explica que uma alta do valor da empresa no mercado, aponta para escolhas focadas em investimentos em projetos com Valor Presente Líquido (VPL) positivo. Como também uma queda no valor da empresa, aponta para escolhas focadas em

investimentos em projetos com VPL negativo. Ainda no mesmo estudo, o autor finaliza mostrando que a divulgação das demonstrações financeiras influencia nos preços das ações, porém existem outros fatores que possam influenciar no mercado, como por exemplo o lucro.

Em Funchal (2016) é sugerido que alguns tipos de investidores sofisticados estão correlacionados a procurar uma diminuição dos riscos em fundos. Funchal (2016) indica que este fato pode levar a problemas de agência, onde os gestores tendem a se submeter a mais riscos em fundos com foco no público geral.

Por fim, cada investimento possui o seu risco e um dos métodos de mensurar esse risco é o de modelar a volatilidade de um certo ativo. Com isso, vem a importância de se utilizar métodos que modelam e prevejam a volatilidade.

Diante ao exposto, o objetivo dessa pesquisa foi comparar a volatilidade dos retornos do preço de fechamento de ações da bolsa brasileira (B3) calculada por diferentes modelos com a volatilidade calculada por um modelo estatístico.

Material e Métodos

Foi obtida a série financeira das ações da AMBEV com código ABEV3.SA, utilizando o site Yahoo Finanças pelo pacote *Quantmod* do *software R* no período de 01 de Janeiro de 2020 até 01 de Janeiro de 2024, os dados se referem apenas a dias úteis, totalizando ao todo 993 valores. Foi montada uma base de dados com a série dos fechamentos em função dos fechamentos dos cinco dias anteriores.

A base de dados foi construída a partir das séries financeiras das ações da AMBEV (ABEV3.SA), Itaú (ITSA4.SA), Vale (VALE3.SA), Petrobrás (PETR4.SA) e Magazine Luiza (MGLU3.SA). Os dados foram obtidos por meio do site Yahoo Finanças pelo pacote *Quantmod* do *software R*. O período de coleta abrange de 1º de janeiro de 2020 a 1º de janeiro de 2024, totalizando 993 valores, correspondendo apenas a dias úteis, ou seja, excluindo fins de semana e feriados.

Os dados coletados consistem nos preços de fechamento das ações, que refletem o valor final das transações no mercado ao final de cada dia útil. Também foram incluídos os preços de fechamento dos cinco dias anteriores. Essa abordagem, conhecida como 'lag', é uma técnica comum em análises financeiras que permite observar como os preços passados podem influenciar o preço atual. Assim, a base foi estruturada para incluir colunas que representam os fechamentos dos preços em cada um dos cinco dias anteriores ao

fechamento do dia atual. Para modelar os dados, os valores foram normalizados para ficarem entre 1 e 101.

Com isso, a base de dados é dividida em duas partes, sendo elas: treino e teste, com os respectivos objetivos: treinar os modelos e comparar a volatilidade dos modelos treinados com a volatilidade histórica e estatística. O modelo utilizado para modelar a volatilidade estatística foi o modelo Generalized Autoregressive Conditional Heteroskedasticity (GARCH). Os modelos foram treinados pela biblioteca Keras do python 3, com o tensorflow como backend, por meio da plataforma Google Colab. Os modelos treinados foram: Random Forest, XGBoost e Regressão Multivariada.

A Random Forest é um método ensemble que combina várias árvores de decisão com o intuito de melhorar a precisão e evitar o overfitting. Cada árvore é treinada em um subconjunto aleatório dos dados e por fim é utilizado uma média dessas árvores para realizar a previsão. Neste trabalho foi utilizado uma profundidade máxima da árvore igual a 100 e 1000 árvores aleatórias

O XGBoost é um método ensemble que utiliza árvores de decisão com boosting, ou seja, constrói árvores em sequência. Cada árvore é construída com o objetivo de diminuir o erro da anterior. Neste trabalho o modelo XGBoost foi treinado utilizando a GPUs CUDA da NVIDIA e o método de histograma para construção de árvores.

A Regressão Múltipla é um modelo estatístico que divide as variáveis entre dependentes e independentes. Modelando a variável dependente em função de um intercepto, um erro aleatório e das variáveis independentes, neste trabalho será considerado o intercepto.

As métricas de comparação utilizadas são o erro absoluto percentual e o R-Quadrado. Os plots analisados serão o plot das importâncias das features e o de comparação entre o retorno e as volatilidades modeladas.

Resultados e Discussão

Random Forest

A Random Forest foi treinada utilizando como parâmetros: a profundidade máxima da árvore igual a 100 e 1000 árvores aleatórias. Analisando as features, percebe-se que o último dia útil (Figura 1) é o mais importante para a tomada de decisão do modelo. O erro absoluto percentual médio para a base de treino foi igual a 4,03% e o R-Quadrado igual a 99%. Já na base de teste o erro absoluto percentual médio foi igual a 10,57% e o R-Quadrado igual a 95%. O cálculo da volatilidade histórica foi realizado considerando um período de 6 dias.

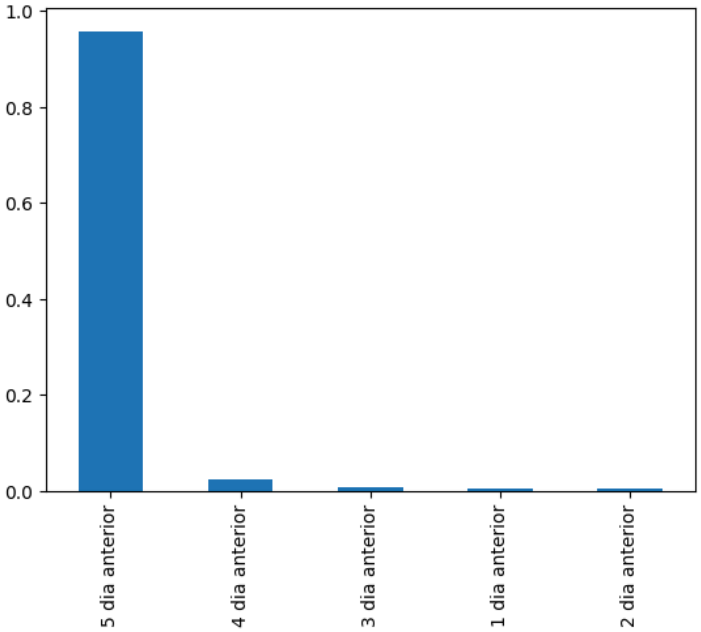

Figura 1. Importância das features - AMBEV

Fonte: Autor

Ao comparar a volatilidade histórica dos valores ajustados com as demais volatilidades modeladas (Figura 2). Nota-se que a volatilidade histórica da série ajustada consegue captar as tendências dos retornos e comparada com a volatilidade estatística possui um comportamento semelhante ao se tratar de altas variações.

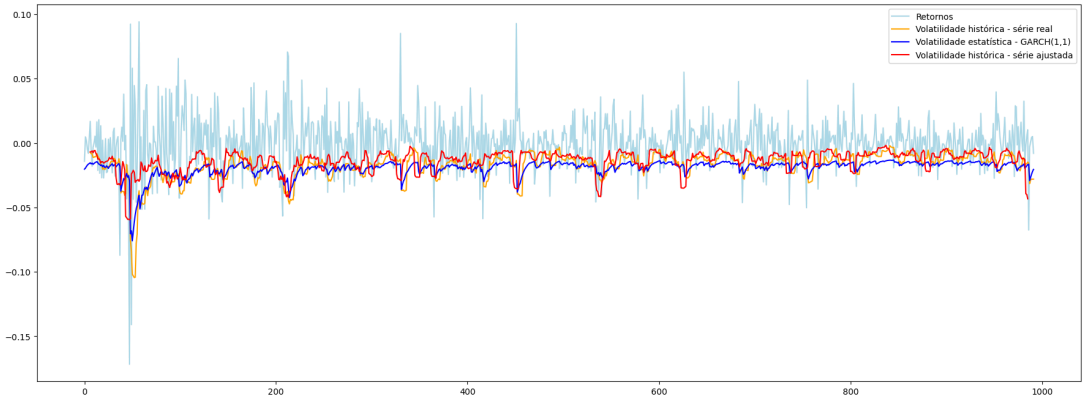

Figura 2. Comparação entre Retorno e Volatilidades modeladas - AMBEV

Fonte: Autor

XGBoost

O XGBoost foi treinado utilizando a GPUs CUDA da NVIDIA e o método de histograma para construção de árvores. Analisando as features, percebe-se que o último dia útil (Figura 3) é o mais importante para a tomada de decisão do modelo.

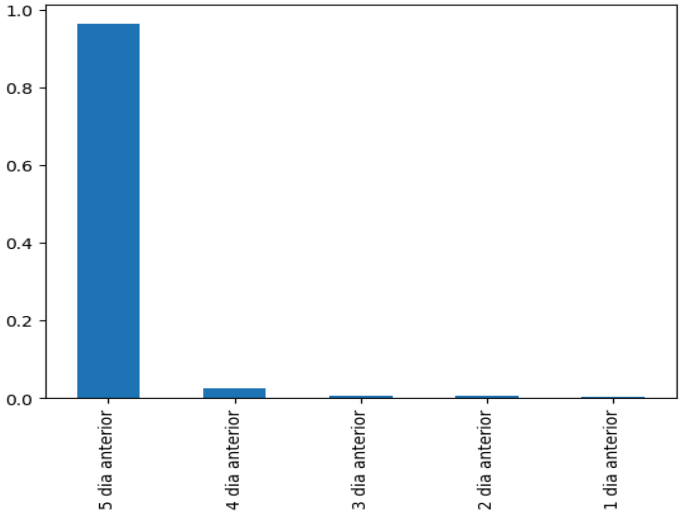

Figura 3. Importância das features - AMBEV

Fonte: Autor

O erro absoluto percentual médio para a base de treino foi igual a 0,97% e o R-Quadrado igual a 99%. Já na base de teste o erro absoluto percentual médio para a base de teste foi igual a 1,14% e o R-Quadrado igual a 99%. O cálculo da volatilidade histórica foi realizado considerando um período de 6 dias.

Ao comparar a volatilidade histórica dos valores ajustados com as demais volatilidades modeladas (Figura 4). Nota-se que a volatilidade histórica da série ajustada consegue captar as tendências dos retornos e comparada com a volatilidade estatística é mais influenciada pelas altas variações.

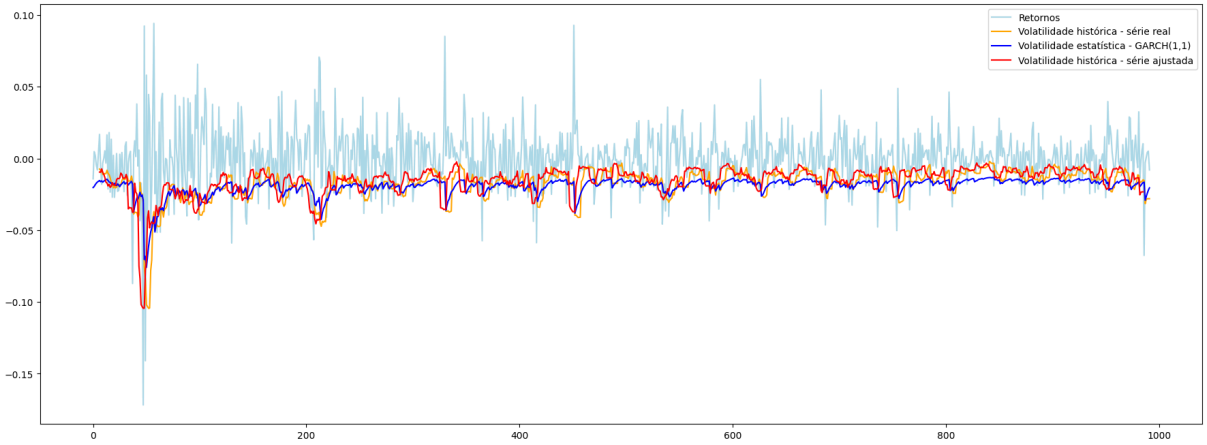

Figura 4. Comparação entre Retorno e Volatilidades modeladas - AMBEV

Fonte: Autor

Regressão Múltipla

A regressão múltipla foi modelada considerando o intercepto. Os coeficientes calculados são: 0,93 (fechamento do dia anterior), 0,14 (fechamento de dois dias anteriores), -0,10 (fechamento de três dias anteriores), -0,02 (fechamento de quatro dias anteriores), 0,02 (fechamento de cinco dias anteriores) e 1,00 (intercepto). Analisando os coeficientes, percebe-se que o último dia útil é o que mais influencia para a variação do valor ajustado do modelo, sendo que um aumento nos valores do fechamento do terceiro dia anterior ou do quarto dia anterior gera uma diminuição no valor ajustado

O erro absoluto percentual médio para a base de treino foi igual a 9,83% e o R-Quadrado igual a 97%. Já na base de teste o erro absoluto percentual médio foi igual a 9,69% e o R-Quadrado igual a 96%. O cálculo da volatilidade histórica foi realizado considerando um período de 6 dias.

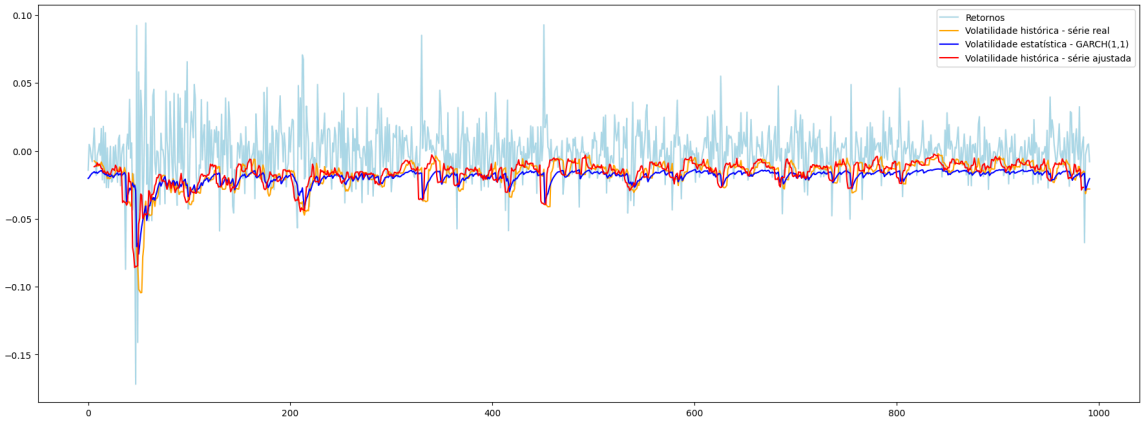

Figura 5. Comparação entre Retorno e Volatilidades modeladas - AMBEV
Fonte: Autor

Ao comparar a volatilidade histórica dos valores ajustados com as demais volatilidades modeladas (Figura 5). Nota-se que a volatilidade histórica da série ajustada consegue captar as tendências dos retornos e comparada com a volatilidade estatística é mais influenciada pelas altas variações.

Comparando os modelos com outras séries

Ao utilizar os mesmos modelos para outras séries com o mesmo período de tempo, percebe-se que as métricas seguem o mesmo padrão, como mostrado nas Tabelas 1, Tabelas 2, Tabelas 3 e Tabelas 4. O XGBoost se adequa melhor tanto na base de teste quanto na do treino e considerando apenas a base de teste a Random Forest é o modelo que possui o pior desempenho em relação às métricas.

ITSA4.SA	**MAPE - Teste**	**R-Squared - Teste**	**MAPE - Treino**	**R-Squared - Treino**
Random Forest	7,11%	95,00%	2,99%	99,00%
XGBoost	0,81%	100,00%	0,79%	100,00%
Regressão Multivariada	6,25%	96,17%	7,07%	96,39%

Tabela 1. Comparação de métricas dos modelos para a série ITSA4.SA

Fonte: Autor

VALE3.SA	**MAPE - Teste**	**R-Squared - Teste**	**MAPE - Treino**	**R-Squared - Treino**
Random Forest	4,50%	99,00%	2,05%	100,00%
XGBoost	0,67%	100,00%	0,71%	100,00%
Regressão Multivariada	4,16%	98,90%	4,62%	99,05%

Tabela 2. Comparação de métricas dos modelos para a série VALE3.SA

Fonte: Autor

PETR4.SA	**MAPE - Teste**	**R-Squared - Teste**	**MAPE - Treino**	**R-Squared - Treino**
Random Forest	5,06%	97,00%	2,35%	100,00%
XGBoost	0,60%	100,00%	0,65%	100,00%
Regressão Multivariada	4,48%	96,76%	5,08%	97,62%

Tabela 3. Comparação de métricas dos modelos para a série PETR4.SA

Fonte: Autor

MGLU3.SA	**MAPE - Teste**	**R-Squared - Teste**	**MAPE - Treino**	**R-Squared - Treino**
Random Forest	5,40%	100,00%	1,97%	100,00%
XGBoost	1,30%	100,00%	1,42%	100,00%
Regressão Multivariada	4,40%	99,72%	4,47%	99,73%

Tabela 4. Comparação de métricas dos modelos para a série MGLU3.SA

Fonte: Autor

Conclusão

Ao comparar os modelos para o cálculo da volatilidade histórica, utilizando o erro absoluto percentual médio e o R-Quadrado para a base de teste, o modelo que melhor se ajustou foi o XGBoost e que menos se ajustou foi a Random Forest. Contudo a Random Forest, para a base de treino, obteve estatísticas de erro absoluto percentual médio e R-Quadrado melhores quando comparada com as métricas da regressão multivariada. Para ambos os modelos o último dia útil é o que mais influência para a variação do valor ajustado dos modelos.

Referências

BODIE, Zvi; KANE, Alex; MARCUS, Alan. Fundamentos de investimentos. AMGH Editora, 2014.

DA SILVA, Jucyara Gomes; NETO, Odilon Saturnino Silva; DA CUNHA ARAÚJO, Rebeca Cordeiro. Educação financeira de servidores públicos: hábitos de consumo, investimento e percepção de risco. Revista Evidenciação Contábil & Finanças, v. 5, n. 2, p. 104-120, 2017.

DE OLIVEIRA CARMINATI, João Guilherme; FERNANDES, Elaine Aparecida. O impacto do investimento direto estrangeiro no crescimento da economia brasileira. Planejamento e políticas públicas, n. 41, 2013.

TAVARES, Márcia. Investimento brasileiro no exterior: panorama e considerações sobre políticas públicas. CEPAL, 2006.

ANTUNES, Marco Aurélio; PROCIANOY, Jairo Laser. Os efeitos das decisões de investimentos das empresas sobre os preços de suas ações no mercado de capitais. Revista de Administra&ccdeil; ão da Universidade de São Paulo, v. 38, n. 1, 2003.

FUNCHAL, Bruno; LOURENÇO, Diogo; MOTOKI, Fabio Yoshio Suguri. Sofisticação dos investidores, liberdade de movimentação e risco: um estudo do mercado brasileiro de fundos de investimento em ações. Revista de Contabilidade e Organizações, v. 10, n. 28, p. 45-57, 2016.