Análise e Previsão de Preços de Ações utilizando RandomForest

Sousa Silva, K. G. F.

[1]Curso de Ciência da Computação - Universidade Federal de Roraima (UFRR)
Campus Paricarana, Bloco IV
69304-000 – Boa Vista – RR – Brasil

kaioguilherme444@gmail.com

Resumo. *Este artigo apresenta uma análise abrangente da aplicação do modelo Random Forest Regressor na previsão de cotações de ações com base em dados históricos. Além disso, exploramos a otimização do modelo por meio de algoritmo genético, destacando melhorias significativas em seu desempenho. Avaliamos o modelo em cenários de previsão de curto e longo prazo, incluindo a consideração de múltiplas saídas.*

1. Introdução

No contexto financeiro, a análise e previsão de cotações de ações apresentam desafios significativos, decorrentes da natureza dinâmica e complexa do mercado. Diversos Modelos de Aprendizado de Máquina (AM), incluindo o Random Forest Regressor, podem ser investigados para abordar essa complexidade, buscando oferecer previsões mais refinadas fundamentadas em dados históricos.

Este estudo investiga o desempenho do Random Forest Regressor na previsão de cotações de ações, com ênfase na aplicação de técnicas de otimização, especificamente utilizando algoritmo genético. A análise abrange a validação do modelo, utilizando métricas reconhecidas, e a avaliação da curva de aprendizagem para compreender seu comportamento em relação ao conjunto de treinamento. Os resultados obtidos são discutidos em detalhes, destacando a capacidade do modelo de se adaptar a padrões complexos do mercado de ações.

2. Metodologia

abordamos o desenvolvimento de um modelo preditivo para cotações de ações. Inicialmente, dados da B3 são coletados e sujeitos a uma etapa de limpeza e pré-processamento. A otimização dos hiperparâmetros do modelo Random Forest Regressor é realizada por meio de um algoritmo genético.

A primeira etapa do estudo compreende uma única previsão para o futuro em um intervalo de um ano. Resultados dessa fase orientam ajustes nos hiperparâmetros, com insights do algoritmo genético. Comparativos com modelos base fornecem uma avaliação crítica do desempenho.

Em seguida, o modelo é estendido para previsões ao longo de 12 meses, configurado para intervalos definidos. A avaliação utiliza métricas como erro médio absoluto e curvas de aprendizagem para visualizar o comportamento ao longo do tempo.

Esse processo iterativo de teste, ajuste e avaliação guia o refinamento contínuo do modelo, garantindo sua eficácia em cenários futuros.

2.1. Dados

Nesta seção, apresentaremos o processo de coleta de dados das ações listadas na Bolsa de Valores do Brasil (B3), essenciais para conduzir a análise apresentada neste estudo.

2.1.1. Fonte de Dados

A fonte primária dos dados neste estudo é a plataforma de investimento Investing.com. As informações sobre cotações foram obtidas do Yahoo Finance, enquanto a lista de ações cadastradas na B3 foi extraída do seguinte URL no Investing.com:

```
https://br.investing.com/stock-screener/?sp=country::
  32|sector::a|industry::a|equityType::a|exchange::
                a%3Ceq_market_cap;1
```

Esta plataforma oferece informações abrangentes sobre ações listadas na B3 e permite filtragem com base em critérios como país, setor, indústria e capitalização de mercado.

2.1.2. Coleta de Dados

A coleta de dados foi conduzida meticulosamente para garantir a integridade e precisão das informações. Os dados coletados incluem, mas não se limitam a:

- Ticker
- Setor
- Indústria
- Data
- Preço de Abertura (Open)
- Preço de Fechamento (Price)
- Volume de Negociação (Volume)
- Dividendo

- Dividend Yield
- P/L (Preço/Lucro)
- P/VP (Preço/Valor Patrimonial)
- LPA (Lucro por Ação)
- VPA (Valor Patrimonial por Ação)
- Índice de Graham
- Enterprise Value
- Capitalização de Mercado (Market Cap)
- Valoração de 12 meses (Valuation 12M)
- Preço Mínimo de Valoração (Valuation Price Min)
- Preço Médio de Valoração (Valuation Price Med)
- Preço Máximo de Valoração (Valuation Price Max)
- Datas no formato yyyy-mm-dd, abrangendo o período de 2023-11-05 até 2018-11-05

Toda a coleta destas informações forão efetuadas por meio de um script personalizado que utilizou a biblioteca ”yahooquery”para efetuar requisições ao Yahoo Finanças. Isso assegurou a obtenção dos dados mais recentes disponíveis no momento da solicitação.

2.1.3. Período de Análise

Os dados coletados proporcionam uma série temporal valiosa que se estende desde a data de coleta (2023-11-05) até 5 anos atrás (2018-11-05). Esta extensa série temporal é fundamental para a análise das tendências de preços e desempenho das ações no contexto da Bolsa de Valores do Brasil.

Todas as informações coletadas foram organizadas em uma tabela no formato CSV, onde cada linha representa uma ação específica. Essa tabela serve como a base para as análises subsequentes neste estudo.

A coleta de dados foi conduzida de maneira rigorosa para garantir a qualidade, confiabilidade e atualidade das informações, atendendo aos requisitos de um estudo científico.

2.2. Limpeza de Dados

Nesta seção, descreveremos o processo de limpeza de dados realizado em nosso estudo, com ênfase na exclusão das ações que não continham dados temporais completos dentro do período de 5 anos de histórico.

2.2.1. Exclusão de Ações com Dados Temporais Incompletos

Uma etapa crítica na limpeza de dados foi a exclusão das ações que não possuíam dados temporais completos dentro do período de 5 anos retroativos a partir da data da coleta (2023-11-05 até 2018-11-05). Esta exclusão foi realizada para garantir que todas as ações analisadas tenham um histórico de cotações completo durante o período de interesse.

Considerando que a bolsa de valores B3 não opera nos finais de semana e feriados, os dados temporais diários são, por padrão, contínuos a cada dia útil. Portanto, não foi

necessário realizar exclusões com base nos finais de semana ou feriados, uma vez que a ausência de cotações nesses dias é esperada e não implica em dados incompletos.

O resultado foi um conjunto de ações com dados completos e prontos para análises posteriores.

O processo de limpeza de dados resultou em um conjunto de dados preparado e consistente, adequado para análises estatísticas e modelagem.

2.3. Preprocessamento

2.3.1. Limpeza e Preprocessamento de Dados

Antes de avançarmos para a etapa de modelagem, foi realizado um processo meticuloso de limpeza e preprocessamento nos dados do dataset. Estas ações visam garantir que o conjunto de dados esteja adequadamente tratado para a tarefa de previsão de preços de ações com base em dados temporais.

1. **Filtragem de Ações com Preços Superiores a 500:**
   Foi realizada a filtragem de ações com valores superiores a 500 unidades devido à presença limitada dessas ações no conjunto de dados. A escassez de casos poderia introduzir imprecisões durante a modelagem, uma vez que essas ações podem apresentar comportamentos menos representativos.
2. **Seleção de Preços e Definição da Janela Temporal:**
   A escolha estratégica da janela temporal, centrada na data de $2023-11-01$, juntamente com a exclusão de informações anteriores, foi realizada para preparar o conjunto de dados para a tarefa de previsão temporal. Adotamos uma abordagem de previsão para um ano no futuro, utilizando dados do intervalo de $2022-09-30$ a $2020-04-07$. Essa estratégia busca capturar padrões e tendências que possam antecipar os preços das ações em novembro de 2023.

2.4. Regressão com saida unica

Na etapa de regressão, empregamos um algoritmo genético para otimizar os hiperparâmetros de um modelo de Regressor de Floresta Aleatória (*RandomForestRegressor*). Essa escolha se baseia na robustez e flexibilidade desse modelo.

2.5. Algoritmo Genético

O algoritmo genético foi configurado com os seguintes hiperparâmetros:

- **Tamanho do Cromossomo (Chromosome Size):** 5 (representando os hiperparâmetros do Random Forest).
- **Tamanho da População (Population Size):** 100 indivíduos.
- **Número de Genes (Genes Number):** 30 (representando valores de 0 a 29, correspondendo aos hiperparâmetros do Random Forest).
- **Número de Gerações (Generations):** 50 iterações.
- **Função de Aptidão (Fitness Function):** Coeficiente de Determinação (R^2) entre as previsões e os rótulos de teste.
- **Probabilidade de Mutação (Mutation Probability):** 2.5%.
- **Elitismo (Best):** 50% (preservação dos melhores indivíduos).
- **Número de Elitismos (Num Elites):** 4 melhores indivíduos preservados.
- **Probabilidade de Seleção (Selection Probability):** 0.1%.

2.6. Hiperparâmetros do Random Forest Otimizados

Os seguintes hiperparâmetros do modelo Random Forest foram otimizados:

1. **Número de Estimadores (*n_estimators*):** Varia de 1 a 30.
2. **Profundidade Máxima (*max_depth*):** Varia de 1 a 30.
3. **Número Mínimo de Amostras para Divisão (*min_samples_split*):** Varia de 2 a 32.
4. **Número Mínimo de Amostras em Folhas (*min_samples_leaf*):** Varia de 1 a 31.
5. **Número Máximo de Características (*max_features*):** Varia de 1 a 31.

2.7. Influência dos Hiperparâmetros do Random Forest

1. **Número de Estimadores (*n_estimators*):** - Influência na complexidade do modelo. Valores mais altos podem melhorar o desempenho, mas aumentam o custo computacional.
2. **Profundidade Máxima (*max_depth*):** - Controla a profundidade máxima das árvores. Profundidades maiores podem levar a overfitting.
3. **Número Mínimo de Amostras para Divisão (*min_samples_split*):** - Define o número mínimo de amostras necessárias para realizar uma divisão. Valores mais altos evitam divisões excessivas.
4. **Número Mínimo de Amostras em Folhas (*min_samples_leaf*):** - Especifica o número mínimo de amostras permitidas em uma folha. Valores mais altos evitam folhas com muito poucas amostras.
5. **Número Máximo de Características (*max_features*):** - Indica o número máximo de características consideradas para a divisão de um nó. Valores menores podem reduzir a correlação entre as árvores.

2.8. Regressão com Múltiplas Saídas: Previsão de 12 Meses

Nesta etapa, ampliamos a aplicação do modelo de Regressão com Múltiplas Saídas para realizar previsões dos valores das ações nos próximos 12 meses. Para configurar essa previsão, realizamos ajustes específicos no conjunto de dados e nas configurações de treinamento.

2.8.1. Configuração para Previsão de 12 Meses

No processo de configuração, ajustamos o conjunto de dados para incluir as labels correspondentes aos meses que desejamos prever. Utilizamos o índice dos meses desejados e configuramos o modelo para realizar as previsões em relação a esses meses específicos.

2.8.2. Intervalos Menores e Configurações Adicionais

Além da previsão mensal, é possível ajustar as configurações de treinamento para realizar previsões em intervalos menores, até mesmo diários. Isso pode ser alcançado através da escolha adequada dos índices temporais e ajuste do conjunto de dados.

Dessa forma, as configurações do modelo podem ser adaptadas conforme necessário para atender a diferentes requisitos de previsão temporal.

2.9. Validação

Na fase de validação, é essencial avaliar o desempenho dos modelos de Aprendizado de Máquina (AM) para garantir que suas previsões sejam confiáveis e generalizáveis para novos dados. As técnicas utilizadas para essa validação foram centradas em métricas de avaliação e análise da curva de aprendizagem.

2.9.1. Métricas de Avaliação

A avaliação do modelo incluiu métricas amplamente reconhecidas que fornecem insights sobre diferentes aspectos do desempenho. As principais métricas utilizadas foram:

- **Coeficiente de Determinação (R^2):** Avalia a proporção da variabilidade nos dados de resposta explicada pelo modelo. Valores próximos a 1 indicam previsões precisas.
- **Erro Médio Absoluto (MAE):** Calcula a média das diferenças absolutas entre as previsões e os valores reais. Menor MAE indica maior precisão.
- **Erro Quadrático Médio (MSE):** Calcula a média dos quadrados das diferenças entre previsões e valores reais. Semelhante ao MAE, com ênfase em erros maiores.
- **Score de Variação Explicado (EVS):** Mede a variação explicada pelo modelo em relação à variação total. Próximo de 1 sugere um bom poder explicativo.

Essas métricas fornecem uma visão abrangente do desempenho do modelo em diferentes aspectos, permitindo uma avaliação mais completa de sua eficácia.

2.9.2. Curva de Aprendizagem

A análise da curva de aprendizagem é fundamental para compreender como o modelo se comporta em relação ao tamanho do conjunto de treinamento. A convergência entre as curvas de treinamento e validação indica um modelo bem ajustado, evitando overfitting ou underfitting. A análise dessa curva permite ajustar a complexidade do modelo para obter o equilíbrio certo entre viés e variância.

Essas avaliações oferecem insights essenciais sobre o comportamento do modelo em diferentes cenários, fornecendo confiança em sua capacidade de generalização para novos dados.

O código completo para a análise e visualização está disponível no seguinte repositório do GitHub:

`https://github.com/Kaioguilherme1/Analise-e-Previssao-de-Precos-de-Acoes-utilizando-RandomForest/tree/main`

3. Resultados

3.1. Resultados com Algoritmo Genético

A aplicação do Algoritmo Genético para otimização de hiperparâmetros resultou em melhorias significativas no desempenho do modelo. O melhor resultado obtido após 50 gerações do Algoritmo Genético está destacado na Tabela 1.

Tabela 1. Melhor Resultado do Algoritmo Genético

Número de Gerações	50
Cromossomo	49
Fitness	[[3, 4, 29, 14, 21], 0.904506166262167]

A Figura 1 ilustra o comportamento do Algoritmo Genético ao longo das gerações, destacando que na geração 30 atingiu seu ápice.

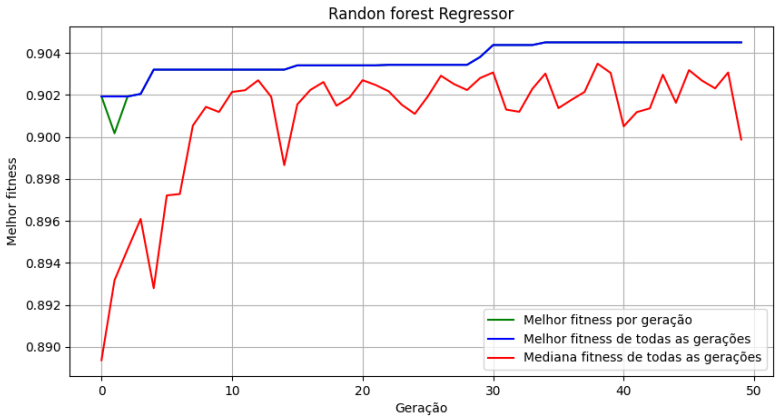

Figura 1. Comportamento do Algoritmo Genético ao longo das gerações.

3.2. Resultados com Saída Única

A análise do modelo para previsões de longo prazo, utilizando como ponto de partida o último dado conhecido (intervalo: 2022-09-30 a 2020-04-07) prevendo os valores para o dia 2023-11-01, resultou nos seguintes indicadores de desempenho no conjunto de teste, onde é possivel notar uma melhora significativa nos resultados com a otimização.

Tabela 2. Comparação dos Indicadores de Desempenho no Conjunto de Teste

Indicador	Sem Otimização	Com Otimização
Número de ações	844	844
Número de dias	618	618
R^2 no conjunto de teste	0.8844	0.9045
Raiz do Erro Médio Quadrático	26.2010	23.8141
Erro Médio Absoluto	15.5249	15.0906
Explained Variance Score	0.8862	0.9057
Quantidade total de itens testados	211	211

A Figura 2 mostra o gráfico de comparação entre os valores reais e previstos para um modelo sem otimização, enquanto a Figura 3 exibe a mesma comparação, porém para um modelo otimizado.

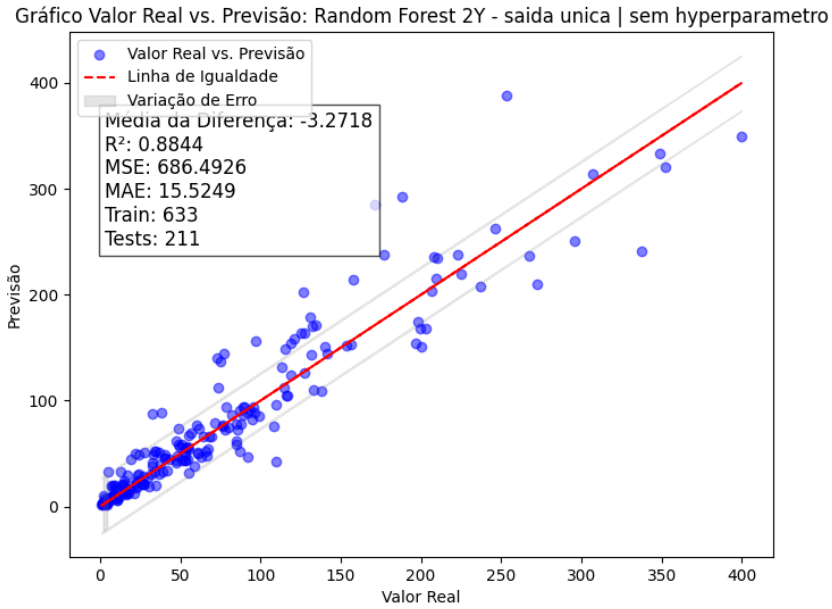

Figura 2. Comparação entre valores reais e previstos com Saída Única sem Otimização.

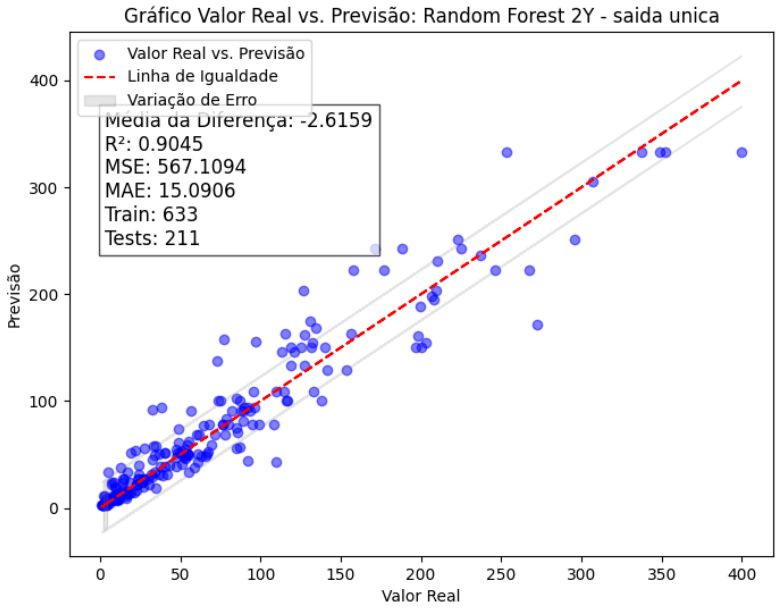

Figura 3. Comparação entre valores reais e previstos com Saída Única Otimização.

Nas figuras a seguir, analisamos o comportamento da taxa de aprendizado em modelos de Saída Única. A Figura 4 retrata a taxa de aprendizado em um modelo sem otimização, oferecendo uma visão do processo de aprendizado padrão.

Já na Figura 5, apresentamos a taxa de aprendizado para um modelo otimizado. A comparação entre essas figuras destaca como a otimização pode influenciar de forma positiva a adaptação do modelo durante o treinamento.

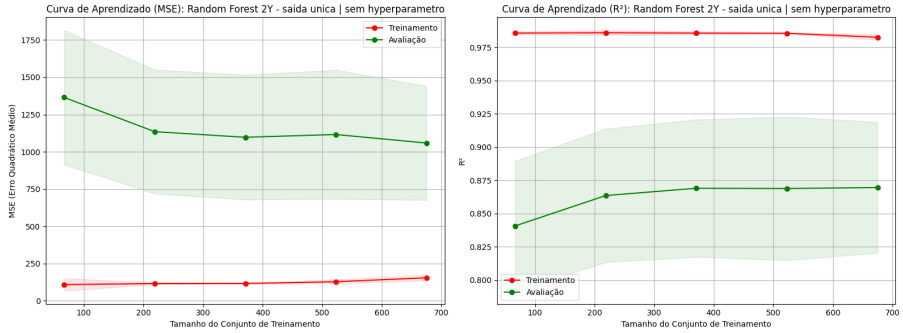

Figura 4. Comportamento da taxa de aprendizado com Saída Única Normal.

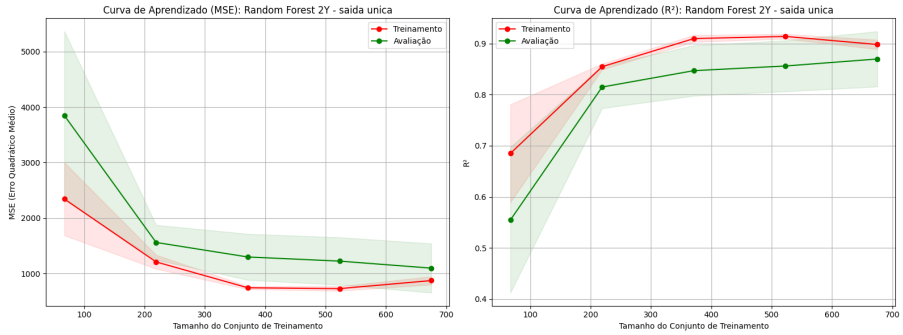

Figura 5. Comportamento da taxa de aprendizado com Saída Única Otimizado.

3.3. Resultados com Múltiplas Saídas

A aplicação do modelo para previsões de longo prazo, gerando um gráfico das cotações a cada dia 1 de cada mês durante 1 ano (mês 11 de 2022 a mês 11 de 2023), resultou nos seguintes indicadores de desempenho no conjunto de teste. É notável que a diferença entre os resultados não foi muito alta, o que era de se esperar, dado que o algoritmo genético foi treinado para otimizar um modelo cujas previsões se limitam a uma única data um ano à frente. Mesmo assim, os resultados obtidos foram satisfatórios, demonstrando a eficácia do modelo mesmo quando utilizando os mesmos hiperparâmetros.

Tabela 3. Comparação dos Indicadores de Desempenho no Conjunto de Teste (Sem e Com Hiperparâmetros)

Indicador	Sem Otimização	Com Otimização
R^2 no conjunto de teste	0.9431	0.9482
Raiz do Erro Médio Quadrático	19.6270	18.7415
Erro Médio Absoluto	10.8472	11.2081
Explained Variance Score	0.9433	0.9489

Nas figuras a seguir, exploramos o desempenho de modelos com Múltiplas Saídas, analisando o gráfico das cotações. onde sera exibido a cotação que obteve o melhor r² score o medio e o menor, A Figura 6 apresenta o resultado para um modelo sem otimização, permitindo-nos observar o comportamento padrão das previsões.

Contrastando com isso, a Figura 7 exibe o gráfico das cotações para um modelo otimizado com Múltiplas Saídas. Ao comparar essas visualizações,

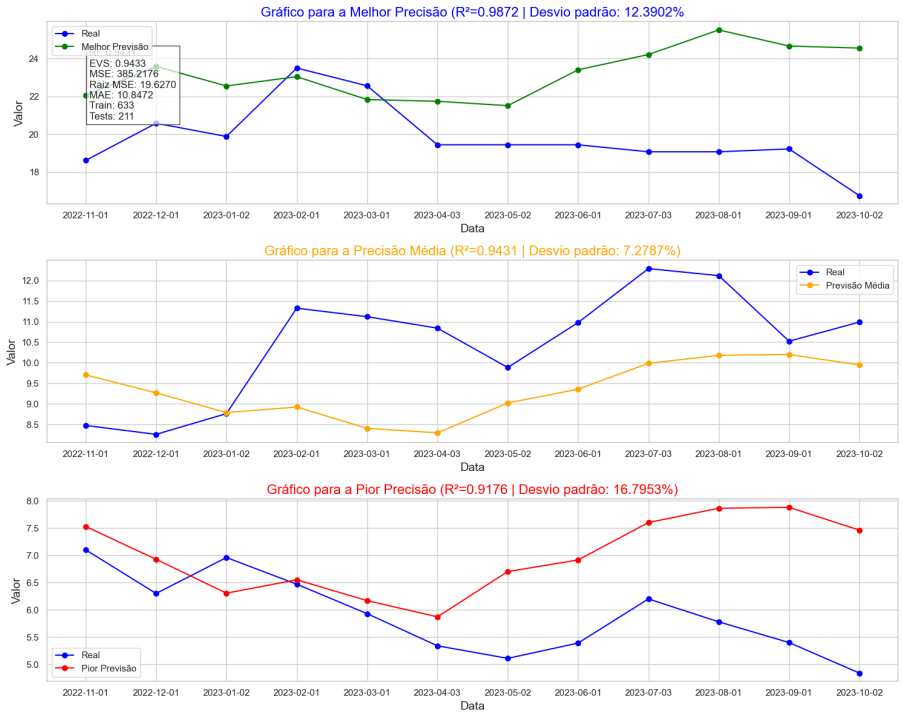

Figura 6. Gráfico das cotações com Múltiplas Saídas Normal.

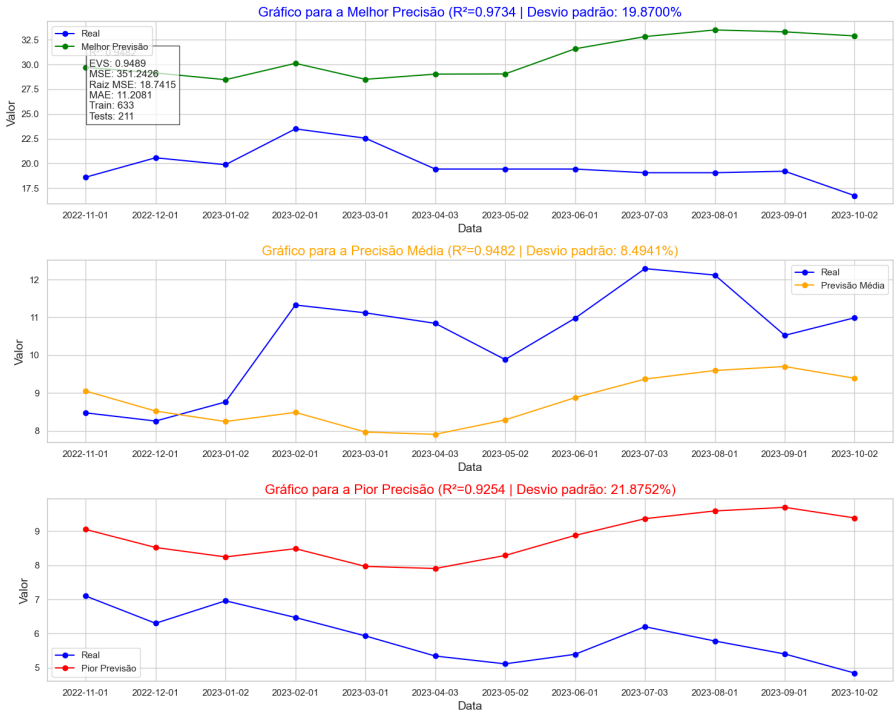

Figura 7. Gráfico das cotações com Múltiplas Saídas Otimizada.

Nas figuras a seguir, examinamos o comportamento da taxa de aprendizado em modelos com Múltiplas Saídas. Na Figura 8, observamos o padrão da taxa de aprendizado para um modelo sem otimização.

Por outro lado, na Figura 9, apresentamos o comportamento da taxa de aprendizado para um modelo otimizado com Múltiplas Saídas. Ao comparar essas visualizações.

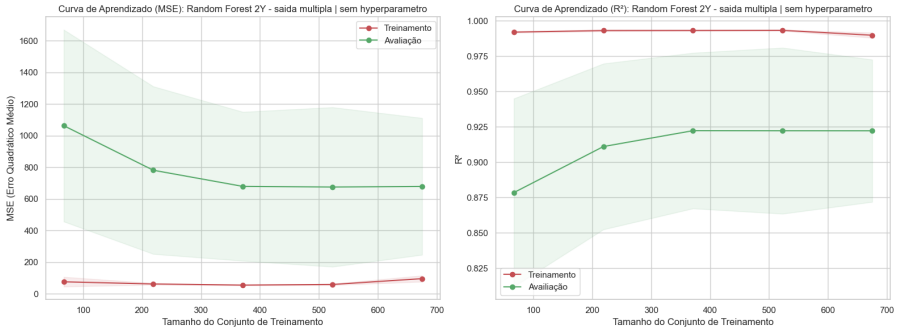

Figura 8. Comportamento da taxa de aprendizado com Múltiplas Saídas Normal.

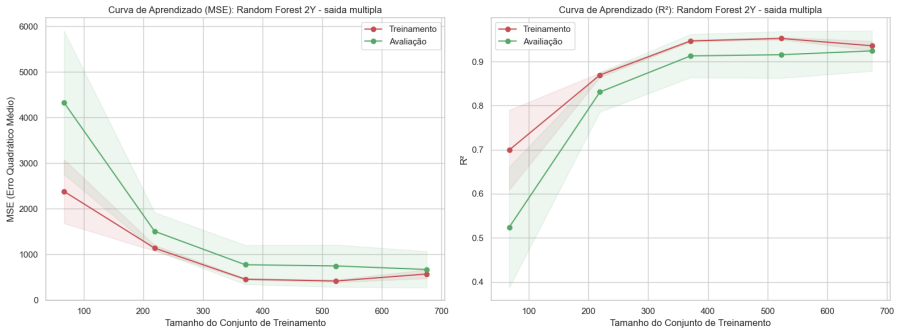

Figura 9. Comportamento da taxa de aprendizado com Múltiplas Saídas Otimizada.

4. Discussão

Os resultados obtidos na aplicação do modelo de Random Forest Regressor para a previsão de cotações de ações revelam um desempenho notável. Os indicadores apresentados nas Tabelas 2 e 3 destacam a eficácia do modelo na captura de padrões e tendências nos dados históricos.

É importante ressaltar que a Random Forest Regressor demonstrou uma capacidade robusta de lidar com a complexidade inerente ao mercado de ações. Mesmo em um ambiente caracterizado por volatilidade e imprevisibilidade, o modelo conseguiu produzir previsões relativamente precisas, como indicado pelos valores de R^2, Raiz do Erro Médio Quadrático e Explained Variance Score.

Ao testar o modelo em previsões de cotações para vários dias separados, observou-se que, como esperado, a precisão aumenta à medida que diminui o intervalo que o modelo pretende prever. Este comportamento está alinhado com a natureza dinâmica do mercado de ações, onde previsões a curto prazo podem ser mais precisas devido à menor variabilidade nos dados.

5. Considerações Finais

Os resultados mostram que o modelo Random Forest Regressor é eficaz na previsão de cotações de ações com base em dados históricos. Sua capacidade de se adaptar a padrões complexos o destaca para previsões financeiras.

A otimização usando algoritmo genético não só melhorou o modelo como também revelou sua grande flexibilidade ao ajustar parâmetros. Isso destaca seu potencial como uma ferramenta versátil e robusta para previsões financeiras.

6. Trabalhos Futuros

Apesar dos resultados promissores, há oportunidades para expandir e aprimorar este estudo no futuro:

- **Inclusão de Novas Features:** Explorar a adição de novas features, como informações setoriais e industriais, para melhorar a precisão das previsões.
- **Consideração de Fatores Externos:** Investigar o impacto de fatores externos, como eventos econômicos ou políticos, no desempenho do modelo.
- **Avaliação de Métricas Alternativas:** Analisar métricas adicionais para avaliar o desempenho do modelo em diferentes aspectos.
- **Ampliação do Conjunto de Dados:** Expandir o conjunto de dados para incluir uma variedade mais ampla de ações e períodos temporais.

Essas sugestões visam enriquecer a capacidade do modelo de Random Forest e explorar áreas ainda não consideradas no escopo deste estudo.

[inv , yah a, yah b, sci].

Referências

Investing.com - stock screener. https://br.investing.com/stock-screener/?sp=country::32|sector::a|industry::a|equityType::a|exchange::a%3Ceq_market_cap;1. Acessado em: 10 de outubro de 2023.

Scikit-learn documentation - randomforestregressor. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html. Acessado em: 20 de outubro de 2023.

Yahoo finance. https://finance.yahoo.com. Acessado em: 10 de outubro de 2023.

Yahooquery documentation. https://yahooquery.dpguthrie.com. Acessado em: 11 de outubro de 2023.