

FUNDAÇÃO GETULIO VARGAS
ESCOLA DE ECONOMIA DE SÃO PAULO

ANDRÉ FIDELIS FIGUEIREDO DE ABREU

**APLICAÇÃO DE *MACHINE LEARNING* NA
PRÉ-SELEÇÃO DE ATIVOS PARA PORTFÓLIOS DE
INVESTIMENTO**

SÃO PAULO

2021

ANDRÉ FIDELIS FIGUEIREDO DE ABREU

**APLICAÇÃO DE *MACHINE LEARNING* NA
PRÉ-SELEÇÃO DE ATIVOS PARA PORTFÓLIOS DE
INVESTIMENTO**

Dissertação apresentada ao Programa de
Mestrado Profissional da Escola de Economia
de São Paulo da Fundação Getúlio Vargas,
como requisito para a obtenção do título de
Mestre em Economia.

Área de concentração:
Engenharia Financeira.

Orientador:
Prof.^a Dra. Élia Yathie Mastumoto

SÃO PAULO

2021

Abreu, André Fidelis Figueiredo de.

Aplicação de *Machine Learning* na pré-seleção de ativos para portfólios de investimento / André Fidelis Figueiredo de Abreu. - 2021

44 f.

Orientador: Élia Yathie Mastumoto.

Dissertação (mestrado profissional MPFE) – Fundação Getulio Vargas, Escola de Economia de São Paulo.

1. Engenharia financeira. 2. Aprendizado do computador. 3. Investimentos - Análise. 4. Mercado de opções. 5. Engenharia financeira. I. Mastumoto, Élia Yathie. II. Dissertação (mestrado profissional MPFE) – Escola de Economia de São Paulo. III. Fundação Getulio Vargas. IV. Título.

CDU 336.764.2

Ficha Catalográfica elaborada por: Raphael Figueiredo Xavier CRB SP-009987/O

Biblioteca Karl A. Boedecker da Fundação Getulio Vargas - SP

ANDRÉ FIDELIS FIGUEIREDO DE ABREU

APLICAÇÃO DE *MACHINE LEARNING* NA PRÉ-SELEÇÃO DE ATIVOS PARA PORTFÓLIOS DE INVESTIMENTO

Dissertação apresentada ao Programa de Mestrado Profissional da Escola de Economia de São Paulo da Fundação Getulio Vargas, como requisito para a obtenção do título de Mestre em Economia.

Área de concentração:
Engenharia Financeira.

Data da Aprovação: 29/10/2021

Banca Examinadora:

Prof.^a Dra. Élia Yathie Mastumoto
(Orientador)
EESP-FGV

**Prof. Dr. Luiz Henrique Moraes da
Silva**
(Co-Orientador)
EESP-FGV

Prof. Dr. Flávio Cipparrone
USP

AGRADECIMENTOS

Agradeço à minha família, em especial à minha esposa Claudia, pelo incentivo e apoio durante todo este processo. Agradeço também os colegas e professores do Mestrado Profissional em Engenharia Financeira por toda a valiosa cooperação, além do Banco Bradesco pelo suporte financeiro que possibilitou a realização deste curso. Por último agradeço à Fundação Getúlio Vargas e aos professores doutores Élia Yathie Mastumoto e Luiz Henrique Moraes da Silva pelo zelosa orientação durante a confecção deste trabalho.

RESUMO

Este trabalho buscou avaliar o impacto de técnicas de *Machine Learning* junto a estratégias de momento na pré-seleção de ativos financeiros para portfólios de investimentos no mercado brasileiro. Foram utilizados modelos de *Random Forest* para ranquear os ativos presentes no IBOVESPA e aplicar esta pré-seleção em portfólios otimizados pelo método da média-variância. Observou-se uma significativa vantagem na utilização desta implementação nos retornos acumulados a longo prazo, como esperado de estratégias baseadas em momento. O portfólio construído a partir ativos pré-selecionados apresentou melhor retorno acumulado e menor sensibilidade à *drawdowns* sistêmicos, o que indica uma maior eficiência uma vez que o custo de seu gerenciamento é consequentemente reduzido.

Palavras-chave: Pré-seleção de ativos. *Machine Learning*. Momento. Portfólio de investimentos. Mercado brasileiro.

ABSTRACT

This work aimed to evaluate the impact of *Machine Learning* techniques alongside momentum strategies in the pre-selection of financial assets for investment portfolios on the brazilian market. *Random Forest* models were used to rank IBOVESPA assets and apply this pre-selection on portfolios optimized by the mean-variance method. Significant advantage on the accumulated returns in the long run was found employing this implementation, as expected from momentum-based strategies. The Portfolio built with the pre-selected assets presented the best accumulated returns and a lower sensibility to systemic drawdowns, which indicates a greater efficiency as management costs are consequently reduced.

Keywords: Pré-selection of assets. *Machine Learning*. Momentum. Investment Portfolio. Brazilian Market.

LISTA DE ILUSTRAÇÕES

| | |
|---|----|
| Figura 1 – Esquematização do processo de otimização de portfólios de investimentos com pré-seletor de ativos | 16 |
| Figura 2 – Exemplificação da separação de períodos na base de validação | 17 |
| Figura 3 – Esquematização do processo de construção de variáveis para as bases de modelagem | 22 |
| Figura 4 – Exemplo de resultados com e sem ordenação dos decis de probabilidade. | 27 |
| Figura 5 – Resultados do Modelo Baseline | 30 |
| Figura 6 – Ordenação por Decis do modelo Baseline | 31 |
| Figura 7 – Resultados do Modelo Randomized Search CV | 32 |
| Figura 8 – Ordenação por Decis do modelo Randomized Search CV | 32 |
| Figura 9 – Resultados do Modelo Grid Search CV | 33 |
| Figura 10 – Ordenação por Decis do modelo Grid Search CV | 33 |
| Figura 11 – Log-Retorno acumulado obtido ao longo de todos os períodos | 34 |
| Figura 12 – Sharpe obtido de acordo com o número de ativos disponibilizados ao otimizador de portfólios | 41 |
| Figura 13 – Sharpe obtido de acordo com o número de ativos disponibilizados ao otimizador de portfólios com eixo normalizado | 42 |
| Figura 14 – Log-Retorno obtido de acordo com o número de ativos disponibilizados ao otimizador de portfólios | 43 |
| Figura 15 – Log-Retorno obtido de acordo com o número de ativos disponibilizados ao otimizador de portfólios com eixo normalizado | 44 |

LISTA DE TABELAS

| | | |
|-----------|--|----|
| Tabela 1 | – Bibliotecas utilizadas durante o desenvolvimento | 18 |
| Tabela 2 | – Descrição de todas as variáveis utilizadas no modelo de <i>Machine Learning</i> para pré-seleção de ativos financeiros para portfólios de investimentos. | 19 |
| Tabela 3 | – Separação dos períodos históricos e número total de registros | 21 |
| Tabela 4 | – Ativos excluídos das bases de Treinamento e Teste na construção da base de modelagem | 23 |
| Tabela 5 | – Períodos de Validação | 25 |
| Tabela 6 | – Bases de Modelagem Separadas | 25 |
| Tabela 7 | – Hiper-Parâmetros considerados na otimização do <i>Random Forest</i> para pré-seleção de ativos. | 26 |
| Tabela 8 | – Lista pré-definida de valores para os hiper-parâmetros considerada no <i>Randomized Search</i> | 26 |
| Tabela 9 | – Lista pré-definida de valores para os hiper-parâmetros considerada no <i>Grid Search</i> | 26 |
| Tabela 10 | – Lista dos valores de hiper-parâmetro de cada modelo criado. | 27 |

SUMÁRIO

| | | |
|----------|--|-----------|
| 1 | INTRODUÇÃO | 10 |
| 1.1 | Objetivos | 10 |
| 2 | REVISÃO BIBLIOGRÁFICA | 11 |
| 3 | FUNDAMENTAÇÃO TEÓRICA | 12 |
| 3.1 | Momento | 12 |
| 3.2 | Teoria Moderna do Portfólio | 13 |
| 3.3 | Random Forest | 14 |
| 4 | METODOLOGIA | 16 |
| 5 | EXPERIMENTO | 18 |
| 5.1 | Preparação dos Dados | 18 |
| 5.1.1 | Preço e Volume Comercializado Diários | 19 |
| 5.1.2 | Setor | 20 |
| 5.2 | Construção das Variáveis | 21 |
| 5.2.1 | Variável Resposta | 21 |
| 5.2.2 | Variáveis Explicativas | 22 |
| 5.3 | Base de Validação | 24 |
| 5.4 | Classificação das Ações | 24 |
| 5.5 | Otimização dos Portfólios | 28 |
| 5.6 | Avaliação e comparação de Portfólios | 28 |
| 6 | RESULTADOS | 30 |
| 6.1 | Comparação de Modelos de Pré-Seleção | 30 |
| 6.2 | Comparação dos Portfólios | 32 |
| 7 | CONCLUSÃO | 35 |
| | REFERÊNCIAS | 37 |
| | ANEXOS | 39 |
| | ANEXO A RESULTADOS POR PERÍODO DE VALIDAÇÃO | 40 |

1 INTRODUÇÃO

A pesquisa voltada para aplicações de *Machine Learning* em problemas de previsão de retornos de ativos financeiros é um tema recorrente na literatura, porém os ativos utilizados já são normalmente pré-definidos. Desta forma, existem poucos trabalhos voltados para a etapa de pré-seleção de ativos para fundos e portfólios de investimentos, principalmente no mercado brasileiro. A proposta deste trabalho é buscar contribuir para o preenchimento desta atual lacuna, procurando obter um portfólio com resultados similares com um número de ativos à sua disposição reduzido com base nos resultados do pré-seletor desenvolvido.

Serão utilizados modelos de *Machine Learning* supervisionados (*Random Forest*) para a classificação de ações do mercado brasileiro, buscando distinguir, dentre os ativos presentes no IBOVESPA (devido à sua liquidez), os que terão um retorno futuro acumulado positivo nos próximos 20 dias. Para isso serão utilizadas principalmente informações de momento das ações, assim como volume comercializado e setor ao qual o ativo pertence.

A probabilidade obtida a partir dos modelos será utilizada para ranquear e pré-selecionar os ativos financeiros, definindo como os melhores os que obtiverem maior probabilidade de possuir um retorno futuro acumulado positivo.

Em seguida se verificará o impacto desta estratégia no resultado de portfólios de investimentos construídos com e sem a influência do pré-seletor. Isto será feito construindo carteiras de investimento com menos ativos à disposição do otimizador de portfólios e comparando seu retorno com uma carteira construída com todos os ativos à sua disposição.

1.1 Objetivos

- Construir modelos de classificação que, a partir de sinais de momento, estimem a probabilidade de uma ação ter um retorno acumulado positivo nos próximos 20 (buscando assemelhar-se à um mês) dias úteis.
- Utilizar a probabilidade estimada para ordenar os ativos de melhor (maiores probabilidades) para pior (menores probabilidades).
- Submeter, na ordem obtida, os ativos à um otimizador de portfólios (elaborado a partir do método da média variância).
- Comparar os resultados dos portfólios à medida que mais ativos vão sendo oferecidos ao processo de otimização.

2 REVISÃO BIBLIOGRÁFICA

Avanços recentes na área de *Machine Learning* estão encontrando aplicações comerciais em diversos setores, incluindo a financeira (EMERSON et al., 2019). Dentre as aplicações na área de investimentos, como sugere (HUANG, 2012), podemos destacar: modelos para predição de séries temporais; otimizações multi-objetivo do retorno com redução de risco e a seleção de ativos para gestão de portfólios.

A seleção de ativos financeiros para a construção de portfólios foi há muito tempo definida como desafiadora e importante. (HUANG, 2012) propõe que avanços nas tecnologias de *Machine Learning* e *Data Mining* estão originando novas oportunidades nesta área. O trabalho de (RASEKHSCHAFFE; JONES, 2019) indica que modelos quantitativos populares não são capazes de utilizar a grande quantidade de dados que existem atualmente, portanto modelos capazes de dinamicamente aprender com dados históricos vem sendo mais utilizados.

Existem várias maneiras de atacar esta classe de problema, buscando categorizar os ativos financeiros. (LEE, 2009) utiliza a tendência dos ativos para classificar entre bons e ruins, já (FU et al., 2018) os classifica com base na razão de retorno esperado e volatilidade. Uma outra alternativa seria realizar a predição das séries temporais de retorno dos ativos com métodos de regressão, como (HUANG, 2012) faz em seu trabalho.

O trabalho de (FU et al., 2018) buscou modelar a classificação de ações do mercado chinês utilizando a razão de excesso de retorno por volatilidade (índice Sharpe). Foram construídas 244 variáveis técnicas e de análises fundamentalistas para caracterizar os ativos e foram utilizados modelos estatísticos (regressão logística), *Machine Learning* (*Random Forest*, *Deep Neural Network* e *Stacking*) e algoritmos genéticos para a seleção de variáveis explicativas dos modelos.

Existem poucos estudos aplicados ao mercado brasileiro voltados à este tema, dentre eles (PORDEUS, 2018), que realiza a seleção de ativos utilizando *Gradient Boosting*. Nesse caso, o trabalho utilizou outros tipos de modelos baseados em informações da microestrutura do mercado brasileiro para a geração de inputs.

3 FUNDAMENTAÇÃO TEÓRICA

3.1 Momento

Momento é um conjunto de estratégias de investimentos na qual retornos passados podem prever retornos futuros (JEGADEESH, 1993). Foi inicialmente conceitualizado por Robert Levy em 1967, originalmente com o nome de força relativa (*relative strength*) (LEVY, 1967). Porém, devido ao surgimento e soberania da hipótese do mercado eficiente poucos acadêmicos buscaram dar continuidade a este trabalho (GRAY, 2016).

Após décadas de inatividade, esta linha de pensamento voltou a ser estudada no início da década de noventa por Narasimhan Jegadeesh e Sheridan Titman (JEGADEESH, 1993); porém o termo *momento* só foi cunhado por Mark Carhart em 1997 (CARHART, 1997).

Recentemente o *momento* tem sido utilizado em conjunto com o modelo de 5 fatores de Fama-French (FAMA, 2015), junto ao tamanho das empresas, valor book-to-market, excesso de retorno no mercado, lucratividade e investimento interno da empresa. O modelo de 5 fatores é uma variante do modelo de 3 fatores (FAMA, 1993).

Existe comumente uma confusão entre investimentos de momento e crescimento (*growth*), uma vez que ambas estratégias costumam buscar ativos com preços recentes altos. A diferença principal está relacionada à análises fundamentalistas. Enquanto a estratégia de crescimento busca ativos com preço acima do valor intrínseco da empresa, a de momento busca ativos com performance relativamente forte *independentemente* da análise fundamentalista (GRAY, 2016).

Por outro lado, existem mais semelhanças entre as estratégias de momento e as de valor (*value*). As duas funcionam, em seus cerne, de maneira semelhante, pois atuam como sinais de um erro sistemático de expectativa do mercado que, em média, promovem resultados à favor do investidor (ASNESS TOBIAS J. MOSKOWITZ, 2013).

A principal diferença se encontra no viés comportamental que concebe este erro. Enquanto as estratégias de valor dependem de um exagero do mercado com relação às novas informações (variações nos preços causadas por investidores ao reagir às notícias são mais intensas do que as variações nos fundamentos), enquanto as de momento baseiam-se em uma reação atenuada dos investidores (ou seja, desmerecem o impacto de determinada nova informação e reagem menos do que deveriam) (BARBERIS ANDREI SHLEIFER, 1998).

Portfólios de momento, como analisado no mercado americano, costumam superar

os retornos de estratégias de crescimento e valor, porém é uma estratégia de longo prazo. Em curto e médio prazos pode ter extensos períodos de baixa performance, além de ser muito suscetível à *drawdowns* sistêmicos de mercado. Ademais, as estratégias de valor e momento não são correlacionadas, portanto existe uma oportunidade de utilização de ambas para diversificação de uma carteira de investimentos (GRAY, 2016).

A estratégia, no geral, costuma ter melhor performance com uma frequência maior de recalibração e um número reduzido de ativos na carteira (GRAY, 2016). Existe uma preocupação com relação ao primeiro ponto, uma vez que uma grande frequência de recalibrações pode originar um custo alto de transação, o que poderia impossibilitar a estratégia; porém foi provado que os lucros são robustos com relação a estes custos (FRAZZINI RONEN ISRAEL, 2014).

A efetividade do sinal de momento depende do histórico utilizado. Foi identificado que os momentos de curto (até 1 mês) e longo (5 anos pra frente) prazo costumam ocasionar reversões, ou seja, ganhadores passados seriam perdedores futuros e vice-versa. Sinais de médio prazo (entre 1 e 12 meses) não aparentam ter esta característica e portanto são mais utilizados (GRAY, 2016).

Outra qualidade deste sinal é a sua dependência com o caminho do preço. Em (DA UNIT G. GURUN, 2014) pode-se observar que o momento é mais efetivo quando contínuo, ou seja, sem grandes variações rápidas (*jumps*) de preço. Zhi Da e Umit Gurun consideram que isto ocorre devido à atenção limitada dos investidores, ocorrendo mais fortemente o viés de reação atenuada quando as mudanças são graduais (DA UNIT G. GURUN, 2014).

Este conceito pode ser metrificado a partir da equação 3.1 a seguir.

$$ID = Sinal(RetornoPassado) * [\%negativos - \%positivos] \quad (3.1)$$

O índice de Momento acima (ID) é composto pelo produto entre a diferença dos percentuais de dias com retornos positivos e negativos dentro da janela de tempo considerada e do sinal do retorno acumulado na mesma janela. Por exemplo: se uma ação teve um retorno de 60% em uma janela dos últimos 252 dias, sendo que 70% foram retornos negativos e 30% foram positivos, o índice calculado teria um valor de $ID = 1 * [0,7 - 0,3] = 0,4$.

3.2 Teoria Moderna do Portfólio

A teoria moderna do portfólio (TMP), em termos simples, provê uma estrutura para a seleção e construção de portfólios de investimentos baseada no retorno esperado dos mesmos e do apetite de risco do investidor (FABOZZI FRANCIS GUPTA, 2002). Foi fundamentada a partir dos estudos de Harry Markowitz publicados em 1952 no *The*

Journal of Finance (MARKOWITZ, 1952), que foram posteriormente elaborados em seu livro *Portfolio Selection: Efficient Diversification* (MARKOWITZ, 1959), e expandida pelo trabalho de William Sharpe em 1964.

A ideia de que o investidor apenas busca (ou deve buscar) o retorno máximo é falha, segundo Markowitz. Ele sugere que o correto seria considerar que o investidor julga desejável o retorno e indesejável a variância do retorno (MARKOWITZ, 1952). Devido à essa consideração, a TMP também é conhecido como análise de média-variância.

Um dos principais conceitos do TMP é a diversificação de ativos com o propósito de reduzir o risco não-sistêmico (FABOZZI FRANCIS GUPTA, 2002). Isto ocorre devido à estrutura de correlação entre os ativos. No contexto de um portfólio de investimentos, movimentos de depreciação de determinados ativos podem ser compensados por apreciação em outros, reduzindo o risco total.

A análise de média-variância propõe que existe para cada nível de risco um retorno esperado máximo e, da mesma forma, para cada retorno esperado desejado um valor mínimo de risco necessário para obtê-lo.

Uma maneira de obter-se o melhor portfólio, sendo agnóstico ao apetite de risco, é a seleção do portfólio que fornece o maior índice Sharpe. Esta proposição, desenvolvida por William F. Sharpe em 1966 (SHARPE, 1966) indica a razão entre o excesso de retorno esperado (retorno subtraído da taxa livre de risco) por unidade de risco associado (risco visto como variância) (SHARPE, 1994).

Apesar de sua monumental importância histórica, o TMP possui diversas críticas devido à suas premissas, sendo algumas das principais: investidor não ser racional; investidor não possuir completa informação a respeito do mercado; os mercados não serem eficientes e os investidores não possuírem acesso ilimitado à capital (MANGRAM, 2013).

3.3 Random Forest

Random Forest, ou Floresta Aleatória, é um estimador que utiliza um conjunto de árvores de decisão construídas a partir de um subconjunto ou amostragem com reposição (*bootstrapping*) da base de dados inicial.

As árvores de decisão são modelos simples e rápidos de construir, utilizar e interpretar, podendo inclusive trabalhar com variáveis categóricas e numéricas; pecam apenas, como sugere (HASTLE ROBERT TIBSHIRANI, 2009), em acurácia de previsão. A motivação para a utilização de conjuntos de árvores, como por exemplo o *Random Forest*, é solucionar este problema em troca de tornar seu resultado menos interpretável.

Outro aspecto do *Random Forest* é a utilização de uma amostra de variáveis explicativas para a realização de cada uma das quebras das árvores. A classificação é

então realizada a partir da agregação dos votos das árvores construídas (esta ação, quando combinada com *bootstrapping*, é denominada *bagging* (BREIMAN, 1996)).

O modelo inicial foi idealizado por Leo Breiman em 2001 (BREIMAN, 2001) e baseado nos trabalhos de (HO, 1998) e (AMIT; GEMAN, 1997). Em seu trabalho, Breiman sugere que o erro de generalização das florestas converge em um determinado limite quando o número de árvores é suficientemente extenso, e que ele depende da força individual das árvores e de suas correlações.

4 METODOLOGIA

Existem diversas maneiras de avaliar o benefício do uso de técnicas de *Machine Learning* junto à estratégia de momento na seleção de ativos financeiros. Este trabalho baseou-se na metodologia apresentada na Figura 1 para obter o ranqueamento de ativos financeiros e observar a influência desta ordenação nos resultados de portfólios de investimento.

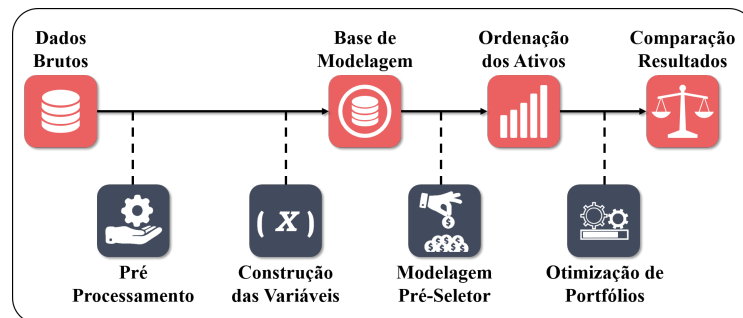


Figura 1 – Esquematização do processo de otimização de portfólios de investimentos com pré-seleção de ativos

Os seguintes passos que devem ser seguidos, cronologicamente, de modo a replicar este trabalho:

1. Obtenção de uma série temporal de preços de ativos financeiros com considerável liquidez em um período extenso de tempo (pelo menos 10 anos). Essas séries devem ser transformadas em log-retornos.
2. Separação da base de dados obtida em três conjuntos: treino, teste e validação. A razão é flexível, mas aconselha-se a utilização de pelo menos 20% do período para a validação (*backtest*).
3. Construção de uma variável resposta binária com base no log-retorno acumulado futuro das séries temporais dos ativos financeiros (1 para positivo e 0 para negativo).
4. Construção de variáveis históricas com potencial explicativo da variável resposta elaborada, com ênfase em variáveis de momento.
5. Treinamento de um modelo de *Machine Learning* para prever a variável resposta com base nas variáveis explicativas utilizando as bases de treinamento e teste.
6. Estruturação da base de validação em períodos segmentados contendo tempo necessário para a construção das variáveis explicativas e da resposta. O final de cada período será considerado como o momento de recalibração do portfólio, portanto

parte do período anterior será incluído para a construção de variáveis explicativas em cada período seguinte. Este racional é exemplificado na Figura 2

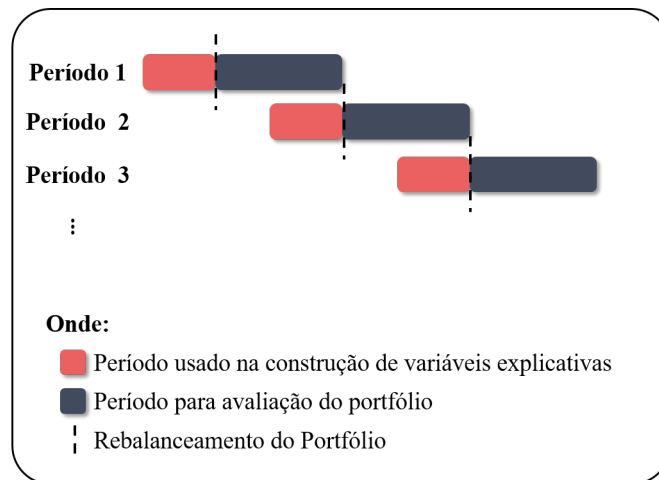


Figura 2 – Exemplificação da separação de períodos na base de validação

7. Cada período segmentado terá um registro para cada ativo financeiro contendo todas as variáveis necessárias para a aplicação do modelo de *Machine Learning* elaborado.
8. Aplicação do modelo construído nos períodos da base de validação e utilização das probabilidades extraídas do modelo para ordenar descendentemente os ativos financeiros (dos melhores ativos para os piores).
9. Construção de portfólios de investimentos incrementando sucessivamente o leque de ativos disponíveis para o otimizador de portfólios, considerando a ordenação obtida na etapa anterior.
10. Obtenção de métricas de performance dos portfólios com e sem a influência do pré-seletor para verificar a sua efetividade.

5 EXPERIMENTO

Esta seção irá expor os passos seguidos neste trabalho, baseando-se na metodologia previamente explanada.

Todos os tratamentos de dados e modelagens - tanto da etapa de pré-seleção dos ativos quanto da otimização de portfólios - foram realizados utilizando-se da linguagem de programação *python 3.0* junto ao aplicativo *web open-source Jupyter Notebook*. As principais bibliotecas utilizadas podem ser vistas na Tabela 1.

Tabela 1 – Bibliotecas utilizadas durante o desenvolvimento

| Biblioteca | Versão |
|--------------|--------|
| scikit-learn | 0.24.2 |
| pandas | 1.3.2 |
| numpy | 1.20.3 |
| matplotlib | 3.4.2 |
| seaborn | 0.11.2 |

5.1 Preparação dos Dados

A primeira etapa para a operacionalização do fluxo de pré-seleção de ativos foi a preparação dos dados. Nesta implementação, as informações mais importantes são as referentes ao preço e volume comercializado dos ativos em um determinado período (extenso suficiente elaborar o modelo de pré-seleção e realizar o *backtest* de seus resultados).

Foram incluídos no processo um *dataset* contendo os setores dos ativos, também utilizados na criação de variáveis explicativas para o pré-seletor de ativos.

O objetivo final desta etapa é a preparação de três bases de dados com propósitos distintos: treino, teste e validação. As bases de treino e teste serão utilizadas para a construção de um modelo capaz de distinguir quais ativos terão um retorno acumulado futuro positivo nos próximos 20 dias. Por outro lado a base de validação será utilizada para a realização de um *backtest* para comparar as carteiras de investimentos construídas com e sem a influência do modelo pré-seletor de ativos.

Um resumo de todas as variáveis utilizadas no modelo, presentes nas três bases de dados, podem ser vistas na Tabela 2.

Tabela 2 – Descrição de todas as variáveis utilizadas no modelo de *Machine Learning* para pré-seleção de ativos financeiros para portfólios de investimentos.

| Variável | Descrição |
|-----------------------------|--|
| lag0 | Valor do Log-Retorno referente ao dia referência (d0) do registro. |
| lag1 | Valor do Log-Retorno referente ao d-1. |
| lag2 | Valor do Log-Retorno referente ao d-2. |
| lag3 | Valor do Log-Retorno referente ao d-3. |
| lag4 | Valor do Log-Retorno referente ao d-4. |
| lag5 | Valor do Log-Retorno referente ao d-5. |
| lag6 | Valor do Log-Retorno referente ao d-6. |
| lag7 | Valor do Log-Retorno referente ao d-7. |
| lag8 | Valor do Log-Retorno referente ao d-8. |
| lag9 | Valor do Log-Retorno referente ao d-9. |
| lag10 | Valor do Log-Retorno referente ao d-10. |
| volume_total_comercializado | Total do Volume comercializado entre o d-10 e d0. |
| media_logretorno | Média dos Log-Retornos entre o d-10 e d0. |
| std_logretorno | Desvio Padrão dos Log-Retornos entre o d-10 e d0. |
| quantitative_momentum | Valor do ID (Momentum Absoluto Quantitativo) dos Log-Retornos entre d-10 e d0. |
| comunicacoes | 1 caso o ativo pertença ao setor de Comunicações, 0 caso contrário. |
| consumo_ciclico | 1 caso o ativo pertença ao setor de Consumo Cíclico, 0 caso contrário. |
| consumo_nao_ciclico | 1 caso o ativo pertença ao setor de Consumo não Cíclico, 0 caso contrário. |
| energia | 1 caso o ativo pertença ao setor de Energia, 0 caso contrário. |
| financeiro | 1 caso o ativo pertença ao setor Financeiro, 0 caso contrário. |
| materiais_basicos | 1 caso o ativo pertença ao setor de Materiais Básicos, 0 caso contrário. |
| saude | 1 caso o ativo pertença ao setor de Saude, 0 caso contrário. |
| ti | 1 caso o ativo pertença ao setor de TI, 0 caso contrário. |
| utilidade_publica | 1 caso o ativo pertença ao setor de Utilidade Pública, 0 caso contrário. |
| target | Variável Resposta. 1 Caso o Log-Retorno acumulado dos próximos 20 dias é positivo, 0 caso contrário. |

5.1.1 Preço e Volume Comercializado Diários

Os dados referentes aos ativos da bolsa brasileira foram extraídos utilizando um terminal da *Bloomberg* e contemplam o período entre 04/01/2010 e 17/11/2020, totalizando 2839 registros diários e aproximadamente 10 anos. Foram obtidas informações a respeito

de 77 ações que constituíram o IBOVESPA durante este período.

O primeiro passo foi aplicar um método de preenchimento para dados de preço faltantes, *forward fill*, onde valores vazios serão preenchidos com dados do último valor não vazio, com o propósito de obter-se mais dias de histórico aproveitável. Esse tipo de ferramenta, por outro lado, pode introduzir viés nos dados, principalmente considerando que o modelo de seleção irá utilizar variáveis de momento, e essas, por sua vez, buscam detectar a tendência dos valores históricos.

Buscando atenuar este efeito foi utilizado um limite de 20 dias de preenchimento. Desta maneira, exemplificando, caso uma coluna tenha apenas seu primeiro valor não nulo, apenas os próximos 20 dias serão preenchidos com o valor inicial.

Os dados referentes ao preço diário da ação foram transformados em log-retorno, uma vez que iremos trabalhar com o retorno acumulado, e sua utilização irá simplificar as análises e remover o impacto da tendência das séries de ativos. Com relação aos dados de volume foi apenas realizado o preenchimento dos valores vazios com 0, uma vez que os estes indicam que não ocorreu comercialização do ativo no dia.

É importante comentar que o log-retorno aqui mencionado, entre os dias t e $t + 1$, é definido pela equação 5.1.

$$R = \ln \left(\frac{P_{t+1}}{P_t} \right). \quad (5.1)$$

O log-retorno acumulado entre os dias t e $t + N$ é simplesmente

$$R_{acum} = \ln \left(\frac{P_{t+N}}{P_t} \right), \quad (5.2)$$

devido à propriedade de aditividade do log.

5.1.2 Setor

Foi obtida uma listagem da B3 contendo o setor econômico de cada uma das empresas já listadas em seu mercado, de modo que essa informação pudesse ser adicionada no modelo de pré-seleção de ativos como variáveis explicativas.

A próxima etapa foi o desenvolvimento de um dicionário contendo os valores disponíveis de setor na tabela original, junto de uma versão corrigida dessa informação sem acentos, caracteres maiúsculos e espaços, buscando tornar a informação mais fácil de se trabalhar posteriormente na fase de construção de variáveis.

Posteriormente buscou-se a remoção das linhas contendo os segmentos e o preenchimento da informação de setor econômico para todos os códigos dos ativos. Os dados foram então finalizados aplicando-se o dicionário previamente construído. Foi então preparada uma tabela contendo apenas o *ticker* dos ativos assim como seu setor corrigido.

5.2 Construção das Variáveis

Por se tratar de uma série temporal, a primeira preocupação na etapa de preparação das variáveis foi de separar o histórico disponível em três períodos, mantendo a ordenação original do *dataset* e garantindo que a criação de *lags* não acarretaria em vazamento de dados de um período para outro. A distribuição desses históricos, assim como o número total de registros de cada, pode ser observado na Tabela 3.

Tabela 3 – Separação dos períodos históricos e número total de registros

| Período | Início | Final | Número de Registros |
|-----------|------------|------------|---------------------|
| Treino | 05/01/2010 | 29/12/2016 | 1732 |
| Teste | 02/01/2017 | 28/12/2018 | 491 |
| Validação | 02/01/2019 | 17/11/2020 | 468 |

Como o objetivo inicial foi operacionalizar um fluxo de trabalho contendo a pré-seleção de ativos utilizando *Machine Learning*, a escolha dos parâmetros (como por exemplo: o número de *lags* passados, período futuro de previsão e limite mínimo da variável resposta) na criação das variáveis correspondem a uma dentre diversas possíveis implementações. Não foi verificado neste trabalho, porém, o efeito da variação destes parâmetros na comparação entre os portfólios construídos.

Uma vez que as bases destinadas às etapas de treino, teste e validação devem ter um formato semelhante com relação às variáveis construídas, elaborou-se uma função para os tratamentos dessas informações. Essa aplicação recebe a base de dados contendo os log-retornos diários do período (novamente: treino, teste ou validação), assim como outras informações necessárias, e retorna a base pronta para a modelagem.

Outros parâmetros dessa função incluem: base de dados contendo os volumes diários comercializados, construída na etapa 5.1.1; tabela contendo o setor corrigido dos ativos da bolsa brasileira, preparada na etapa 5.1.2; número de *lags* passados desejado; número de dias do período futuro de previsão e limite mínimo do retorno acumulado.

A construção da variável resposta e das explicativas foi realizada individualmente para cada ativo (coluna do *dataset* de log-retornos) e seu resultado foi então acumulado em um *dataset* final. A Figura 3 exemplifica esse processo.

5.2.1 Variável Resposta

Seguindo a proposta de se classificar os ativos disponíveis - buscando pré-selecionar os melhores - a variável resposta do modelo foi idealizada com base no log-retorno desses ativos ao longo de um período pré-determinado. Esse período é variável dentro da construção do fluxo de trabalho, porém para efeito de teste a implementação realizada considerou 20 dias, aproximando à quantidade de dias úteis em um mês.

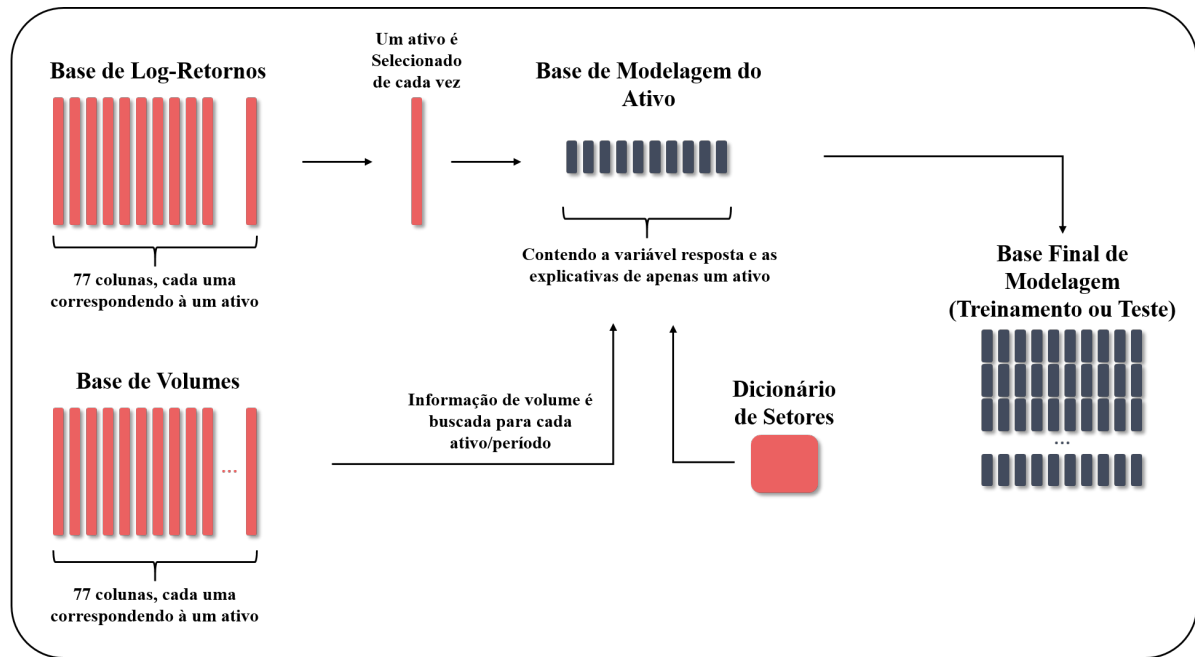


Figura 3 – Esquematização do processo de construção de variáveis para as bases de modelagem

Uma alternativa proposta nesse trabalho foi a de utilizar o log-retorno acumulado dos ativos durante o período selecionado como variável resposta. Dessa maneira não é necessária a previsão diária das séries temporais financeiras, o que costuma acarretar uma grande quantidade de erro nos resultados.

Buscando facilitar ainda mais o problema e garantir melhores resultados no pré-seletor, mais uma etapa de tratamento foi aplicada à variável resposta. A variável numérica R_{acum} foi transformada em categórica, com base em um determinado limite mínimo. Caso R_{acum} seja maior do que 0, sua classificação será 1 (ou *True*); caso contrário 0 (ou *False*).

Dessa maneira, o modelo de *Machine Learning* irá calcular a probabilidade de determinado ativo financeiro ter um log-retorno acumulado nos próximos 20 dias acima de 0. Essa porcentagem será então utilizada para ordenar os ativos, considerando os cuja probabilidade é maior como melhores e vice-versa.

O valor de limite mínimo de 0 pode ser variado em futuros trabalhos para se identificar se é o melhor possível; podendo também variar com relação ao número de dias previstos.

5.2.2 Variáveis Explicativas

As variáveis explicativas utilizadas buscam agregar informações referentes à qualidade dos ativos financeiros, posto que o modelo pretende ordená-los de acordo com a sua propensão a ter um log-retorno acumulado positivo nos próximos 20 dias.

Seguindo o padrão da literatura econométrica, umas das primeiras variáveis explicativas criadas foram os *lags* de cada série temporal de log-retornos dos ativos financeiros. Para a implementação deste trabalho foram considerados 10 *lags* para cada registro da base de modelagem.

No processo de criação dos lags foram identificados algumas particularidades na base de dados. Quando a base foi separada nos três períodos (treinamento, teste e validação), alguns ativos não tinham um número mínimo de log-retornos preenchidos para gerar ao menos um registro para as bases de modelagem (sendo necessários 31 log-retornos consecutivos para esta implementação, considerando o dia atual, 10 lags e 20 dias futuros para a variável resposta). Isso acontece quando as ações de uma determinada empresa não foram comercializadas durante o respectivo período.

Buscando contornar essa característica, os ativos sem o número mínimo foram excluídos das bases. Essa remoção pode prejudicar o resultado do pré-seletor de ativos, uma vez que o modelo não possui informação referente ao comportamento do ativo em seu treinamento; porém ela também simula o advento da inclusão de novas empresas na bolsa brasileira. Os ativos que foram excluídos das bases de treinamento e teste estão indicados na Tabela 4.

Tabela 4 – Ativos excluídos das bases de Treinamento e Teste na construção da base de modelagem

| Período | Número de Registros | Ativos Considerados | Ativos Excluídos |
|-------------|---------------------|---------------------|--|
| Treinamento | 110208 | 68/77 | IRBR3, PCAR3, AZUL4, SUZB3, BRDT3, HAPV3, GNDI3, BPAC11, CRFB3 |
| Teste | 33560 | 76/77 | PCAR3 |

Uma outra informação agregada como variável explicativa foi o volume total comercializado do ativo. Foi considerado o período entre o dia atual do registro e os 10 dias considerados no lags. Essa informação foi incluída como uma *proxy* da liquidez dos ativos.

Complementando a informação fornecida pelos *lags* de cada ativo, outras métricas foram criadas com base nos log-retornos passados de cada série temporal financeira. As duas primeiras, mais elementares, foram a média e o desvio padrão dos valores.

A terceira métrica foi construída com base na equação 3.1. Esta foi aplicada nos log-retornos dos 10 dias de *lags* considerados e busca qualificar o momento do ativo, verificando o quão contínuo ele foi.

O dicionário de setores dos ativos financeiros foi utilizado na construção de uma variável categórica indicando as respectivas Áreas. Entre os possíveis níveis dessa variável estão: "Bens Industriais"; "Comunicações"; "Consumo Cíclico"; "Consumo não Cíclico";

"Energia", "Financeiro"; "Materiais Básicos"; "Utilidade Pública"; "Saúde"; "TI" e "Outros". Essa variável categórica foi transformada em um conjunto de variáveis *dummies* de modo a informação pudesse ser agregada ao modelo de pré-seleção de ativos.

A implementação realizada considera apenas uma pequena gama de variáveis explicativas, de modo a construir uma versão inicial do modelo de pré-seleção de ativos. O processo, porém, permite a aprimoração dos resultados com a inclusão de outras informações relevantes a respeito dos ativos financeiros e sua qualidade.

5.3 Base de Validação

A base de validação sofreu um tratamento distinto das demais. Como a proposta da validação é verificar os resultados dos portfólios de investimentos com e sem a influência do pré-seletor de ativos, o período separado para validação foi dividido em períodos de 31 dias (tempo dos 10 *lags*, data de referência e 20 dias para a variável resposta). Nesse caso, considera-se que na data de referência o portfólio de investimentos será rebalanceado de acordo com os resultados das previsões e será portanto mantido pelo período de 20 dias. A figura 2 exemplifica a divisão destes períodos.

A mesma função utilizada para as bases de modelagem foi então aplicada individualmente em cada um dos 21 períodos de validação. A janela de tempo considerada em cada período pode ser visto na Tabela 5.

5.4 Classificação das Ações

Após a criação das bases de modelagem, o próximo passo foi elaborar um fluxo de trabalho para a construção do pré-seletor de ativos. A primeira etapa neste processo foi a separação das bases de treino em teste entre variáveis explicativas (X) e resposta (y). As dimensões dessas bases podem ser vistas na Tabela 6.

A técnica utilizado neste trabalho foi o *Random Forest*, porém a implementação é agnóstica ao modelo de *Machine Learning*. Desfrutou-se da implementação de *Random Forest* contida na biblioteca de *Machine Learning* para *python*: *scikit-learn* versão 0.24.2. A avaliação de outras técnicas pode ser feita, de modo a otimizar os resultados do pré-seletor, porém não fez parte do escopo deste trabalho.

Foram construídos três modelos com base no *Random Forest*, buscando otimizar os resultados de sua previsão, alterando seus hiper-parâmetros. Os hiper-parâmetros são configurações externas ao modelo cujos valores não podem ser estimados a partir dos dados, e portanto devem ser especificados manualmente. Heurísticas podem ser consideradas, mas é complexo saber *a priori* seus valores ótimos. Uma alternativa para este problema é a busca aleatória por tentativa e erro, o que foi utilizado neste trabalho.

Tabela 5 – Períodos de Validação

| Período | Início | Fim | Ativos Excluídos |
|---------|------------|------------|------------------|
| 1 | 02/01/2019 | 14/02/2019 | PCAR3 |
| 2 | 01/02/2019 | 19/03/2019 | PCAR3 |
| 3 | 06/03/2019 | 17/04/2019 | PCAR3 |
| 4 | 04/04/2019 | 20/05/2019 | PCAR3 |
| 5 | 07/05/2019 | 18/06/2019 | PCAR3 |
| 6 | 05/06/2019 | 19/07/2019 | PCAR3 |
| 7 | 05/07/2019 | 19/08/2019 | PCAR3 |
| 8 | 06/08/2019 | 17/09/2019 | PCAR3 |
| 9 | 04/09/2019 | 16/10/2019 | PCAR3 |
| 10 | 03/10/2019 | 14/11/2019 | |
| 11 | 01/11/2019 | 17/12/2019 | |
| 12 | 04/12/2019 | 21/01/2020 | |
| 13 | 08/01/2020 | 19/02/2020 | |
| 14 | 06/02/2020 | 23/03/2020 | |
| 15 | 10/03/2020 | 23/04/2020 | |
| 16 | 08/04/2020 | 25/05/2020 | |
| 17 | 12/05/2020 | 24/06/2020 | |
| 18 | 10/06/2020 | 23/07/2020 | |
| 19 | 10/07/2020 | 21/08/2020 | |
| 20 | 10/08/2020 | 22/09/2020 | |
| 21 | 09/09/2020 | 22/10/2020 | |

Tabela 6 – Bases de Modelagem Separadas

| Dataset | Linhas | Colunas |
|----------|--------|---------|
| X_treino | 110208 | 24 |
| y_treino | 100208 | 1 |
| X_teste | 33560 | 24 |
| y_teste | 33560 | 1 |

A três versões foram feitas da seguinte maneira, com relação aos seus hiper-parâmetros: primeira, utilizando os valores padrão para todos eles; segundo, aplicando a técnica de busca randomizada (*Randomized Search Cross Validation*); terceiro, realizando o ajuste fino da versão anterior com base no *Grid Search Cross Validation*. Os hiper-parâmetros que foram otimizados estão descritos na Tabela 7.

A utilização de tentativa e erro na otimização dos hiper-parâmetros é de alto custo computacional. Uma maneira de contornar esta questão é a utilização de amostras e do *Randomized Search* em seu processo. Uma amostra de 10% da base de treino foi utilizada neste caso. O *Randomized Search* realiza a amostragem de determinados valores para os parâmetros do modelo dentre uma lista pré-definida buscando obter o melhor modelo possível, até atingir um determinado e pré-especificado número de iterações, com base em

Tabela 7 – Hiper-Parâmetros considerados na otimização do *Random Forest* para pré-seleção de ativos.

| Hiper-Parâmetro | Definição |
|-------------------|---|
| bootstrap | Se as árvores serão, ou não, construídas com base em amostragem com reposição |
| max_depth | A profundidade máxima de cada árvore. |
| max_features | Número de variáveis consideradas na busca da melhor divisão de nó. |
| min_samples_leaf | Número mínimo necessário de amostras em cada nó após divisões. |
| min_samples_split | Número mínimo de amostras necessárias para dividir um nó da árvore. |
| n_estimators | Número de árvores de decisão de cada floresta aleatória. |

validação cruzada. Foram utilizadas 100 iterações e três divisões na validação cruzada. A lista utilizada pode ser vista na Tabela 8.

Tabela 8 – Lista pré-definida de valores para os hiper-parâmetros considerada no *Randomized Search*.

| Hiper-Parâmetro | Valores |
|-------------------|--|
| bootstrap | True, False |
| max_depth | 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, None |
| max_features | sqrt |
| min_samples_leaf | 1, 2, 4 |
| min_samples_split | 2, 5, 10 |
| n_estimators | 200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000 |

Os resultados do *Randomized Search* foram utilizados na preparação da lista de valores para o *Grid Search*, onde cada combinação possível de hiper-parâmetro será testada, de modo a realizar o ajuste fino do modelo. A lista está exposta na Tabela 9.

Tabela 9 – Lista pré-definida de valores para os hiper-parâmetros considerada no *Grid Search*.

| Hiper-Parâmetro | Valores |
|-------------------|------------------|
| bootstrap | False |
| max_depth | 10, 15, 20, None |
| max_features | sqrt |
| min_samples_leaf | 1, 2 |
| min_samples_split | 4, 5, 6 |
| n_estimators | 150, 200, 250 |

Os hiper-parâmetros finais utilizados em cada uma das versões do modelo estão indicados na Tabela 10.

Os modelos foram inicialmente validados com base na matriz de confusão de seus resultados (composta pela proporção de verdadeiros positivos e negativos; e falsos positivos

Tabela 10 – Lista dos valores de hiper-parâmetro de cada modelo criado.

| Hiper-Parâmetro | Padrão | Randomized | Grid |
|-------------------|--------|------------|-------|
| bootstrap | True | False | False |
| max_depth | None | 10 | 20 |
| max_features | sqrt | sqrt | sqrt |
| min_samples_leaf | 1 | 1 | 1 |
| min_samples_split | 2 | 5 | 6 |
| n_estimators | 100 | 200 | 250 |

e negativos). Porém foi identificado que essa não seria a melhor maneira, uma vez que expõe apenas uma faixa do resultado: a que considera um limiar de 50% (onde as ações cuja probabilidade resultante, obtidas pelo modelo, excede esse valor são classificadas como boas, pertencentes à classe 1, e as demais como ruins, classe 0). Essa validação foi expandida ao observar as curvas de *Precisão* e *Recall* dos modelos, que consideram todos os possíveis limiares, assim com as áreas abaixo delas, que representam seu potencial preditivo.

Uma última métrica utilizada no estudo, e considerada como mais relevante devido ao seu propósito, foi a ordenação dos decis das probabilidades dos modelos. Essa métrica observa a quantidade de bons ativos por faixa de probabilidade do modelo. Como o objetivo principal do modelo é ser capaz de identificar os melhores ativos - e consideramos que os melhores são assim definidos por terem uma maior chance de ter seu log-retorno acumulado R_{acum} em 20 dias superior à 0 - desejamos que o modelo seja capaz de acumular esses ativos nos decis de maior probabilidade. A Figura 4 exemplifica a diferença entre um resultado que apresenta ordenação de outro que não o faz.

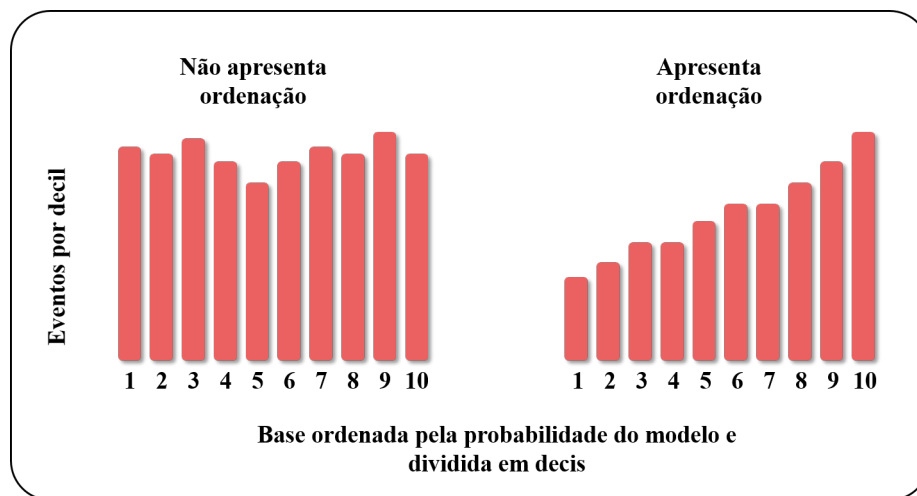


Figura 4 – Exemplo de resultados com e sem ordenação dos decis de probabilidade.

5.5 Otimização dos Portfólios

Para validar o pré-seletor de ativos criados no item 5.4 foi necessário utilizar um modelo de otimização de portfólios, uma vez que desejamos validar o impacto do pré-seletor nos resultados de uma carteira de investimentos. O modelo escolhido para esse fim foi um dos principais paradigmas desse campo: Média-Variância - concebido por Harry Markowitz na década de 50. Apesar de não ser frequentemente empregado no mercado, uma vez que possui limitações com base em suas suposições, é um marco normalmente utilizado para comparar os resultados de modelos mais complexos e portanto seu uso, para a finalidade deste trabalho, é satisfatório.

A implementação do modelo de Média-Variância foi realizada a partir de um processo de Monte Carlo. Números randômicos foram gerados para cada um dos ativos e depois normalizados para representarem o peso do ativo na carteira de investimentos, podendo variar de 0 a 1, sendo sua soma, necessariamente, igual à 1 (ou 100%).

O retorno e a volatilidade esperados são calculados com base nos 10 dias anteriores ao rebalanceamento de cada período de acordo com as equações 5.3 e 5.4 a seguir.

$$E[R] = \sum \bar{M} \bullet \bar{W}, \quad (5.3)$$

onde \bar{M} representa a matriz das médias dos log-retornos anualizados dos ativos nos 10 dias anteriores e \bar{W} a matriz contendo os pesos de cada ativo.

$$E[V] = \sqrt{\bar{W}^T \bullet \bar{Cov} \bullet \bar{W}}, \quad (5.4)$$

sendo \bar{Cov} a matriz das covariâncias anualizadas dos ativos nos 10 dias anteriores.

Com base nestes valores o índice Sharpe do portfólio é obtido e reservado. Esse processo é repetido 10000 vezes e os pesos referentes ao portfólio de maior índice Sharpe são selecionados.

5.6 Avaliação e comparação de Portfólios

O primeiro passo da validação dos resultados foi a aplicação do pré-seletor nos diversos períodos previamente separados para *backtest*, enumeradas na Tabela 5. Entre os três modelos desenvolvidos o terceiro, e último, foi o selecionado devido aos resultados expostos anteriormente na seção 6.1. A aplicação dessa forma simula uma carteira de investimentos que é rebalanceada periodicamente a cada 20 dias úteis, como indicado na Figura 2.

O processo de otimização não foi efetuado para o período 15 (10/03/2020 a 23/04/2020), uma vez que corresponde à um período crítico da bolsa, devido, entre

outros possíveis motivos, à pandemia do COVID-19. Durante esse período o mercado não possuía liquidez, logo não seria possível realizar o rebalanceamento de um portfólio de investimentos. Nesse caso foram utilizados para o período 15 os mesmos pesos do período anterior, tornando os resultados obtidos no *backtest* mais concordantes com o a realidade do período.

Para cada período de validação, os ativos presentes foram ordenados em ordem decrescente com base no resultado do pré-seleção, ou seja, na probabilidade estimada do ativo, nos próximos 20 dias, obter um log-retorno positivo.

Optou-se por determinar o impacto do pré-seleção da seguinte maneira: seguindo a ordenação de ativos encontrada a partir do modelo, foram sendo incluídos ativos no leque de possibilidades do otimizador de portfólio e calculadas as métricas a partir das carteiras resultantes. Dessa maneira, o primeiro portfólio contempla apenas o melhor ativo, de acordo com o pré-seleção; já o segundo considera o primeiro e o segundo melhores. Eventualmente a última carteira poderá utilizar todos os ativos e representará, portanto, o portfólio construído sem a influência do pré-seleção.

Seguindo esse processo, foram avaliadas as seguintes métricas para os portfólios resultantes: índice Sharpe e log-retorno total. Também observou-se o log-retorno acumulado ao decorrer dos 21 períodos de validação de portfólios construídos a partir da disponibilização de 1, 5, 10, 20 e 76 ativos ao otimizador de portfólios.

6 RESULTADOS

Os resultados finais do estudo serão focados em dois pontos principais: primeiramente nas métricas dos modelos de *Machine Learning* utilizados para a pré-seleção de ativos, e, em um segundo momento, nos resultados financeiros de sua aplicação nas bases de validação durante um *backtest*, eventualmente observando seu efeito em uma simulação de sua aplicação.

6.1 Comparação de Modelos de Pré-Seleção

O modelo *baseline*, ou seja, construído sem processo algum de otimização de seus hiper-parâmetros, já demonstrou, em um primeiro momento, resultados relevantes para a pré-seleção de ativos. Na Figura 5.a podemos observar uma alta porcentagem de verdadeiros positivos: 52,43%, portanto o modelo é capaz, considerando um limiar de 50%, de identificar os bons ativos com uma boa precisão (80,64%), já que a quantidade de falsos positivos é baixa. Da mesma forma seu *recall* (67,18%), ou sensibilidade, também é consideravelmente alto devido à sua baixa quantidade de falsos negativos.

Podemos observar que mesmo variando o limiar de classificação, como exposto na Figura 5.b na curva de precisão-*recall* do modelo, a sua precisão permanece acima de 75%. Outro bom indicativo do potencial de predição do modelo *baseline* é sua *AUCPR* - *area under precision-recall curve* - que possui o valor de 0,832.

A *AUCPR* é uma média da precisão ponderada pela probabilidade de um determinado limiar e é utilizada para sumarizar em um valor escalar a curva de precisão-*recall* (BOYD KEVIN H. ENG, 2013), buscando facilitar a comparação entre classificadores (semelhante à *AUC* da curva *ROC*, porém voltada para problemas com classes desbalanceadas).

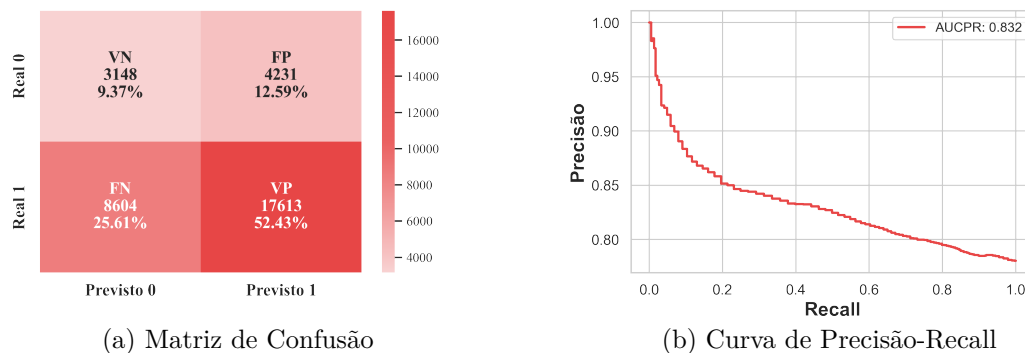


Figura 5 – Resultados do Modelo Baseline

Importante ressaltar que a *AUCPR* considera toda a área do gráfico de precisão e *recall* (ambos variando entre 0 e 1) e os gráficos apresentados somente expõem a precisão variando de 1 até o percentual de positivos, de modo a facilitar sua visualização.

Observando, porém, a quantidade de eventos por decil de probabilidade na Figura 6 (como explanado no Item 5.4) podemos ver que o modelo não é capaz de ordenar os bons ativos efetivamente, ou seja: não existe uma distinção clara entre a quantidade de bons ativos entre as altas e baixas probabilidades.

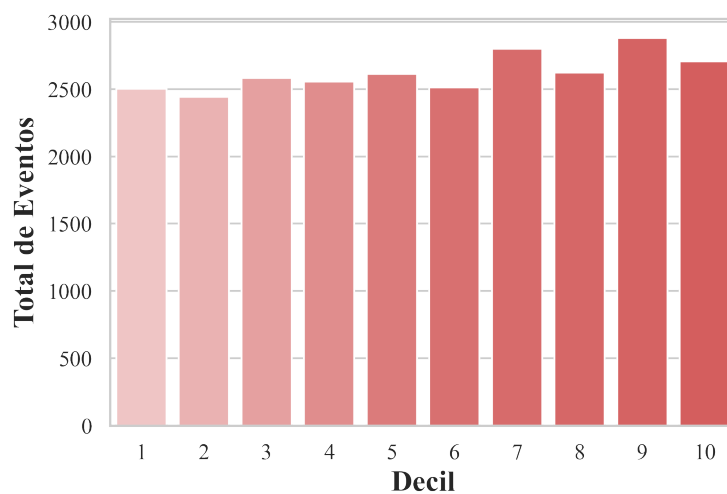


Figura 6 – Ordenação por Decis do modelo Baseline

O modelo cujos hiper-parâmetros foram otimizados a partir da técnica de busca randomizada apresentou, inicialmente, resultados semelhantes aos do *Baseline*, ainda que inferiores, como exposto na Figura 7. Este modelo demonstrou, com limiar de 50%, precisão de 77,88% e *recall* de 60,91%, ambos abaixo do modelo inicial; porém, observando os resultados ao longo dos diversos limiares, depara-se com uma *AUCPR*, e portanto um potencial preditivo, semelhante: 0,831.

Apesar do suposto revés, a otimização dos hiper-parâmetros ocasionou um resultado significativo: a ordenação dos decis de probabilidade. Como é exposto na Figura 8, pode-se notar uma considerável melhora nesta métrica, tendo apenas uma falha na região entre o terceiro e quarto decil, mas apresentando uma definida distinção entre a quantidade de eventos entre os decis superiores, como espera-se, e os inferiores.

Examinando os resultados do modelo cujos hiper-parâmetros foram ainda mais otimizados a partir da técnica de *Grid Search CV*, pode-se observar que, por outro lado, esse processo não garantiu melhoras significativas na ordenação das probabilidades. A ordenação apresentada na Figura 10 é muito semelhante à da versão anterior.

Entretanto, podemos observar na Figura 9 valores de precisão (80,82%), *recall* (66,17%) e *AUCPR* (0,832) mais próximos aos do modelo *baseline*. Neste caso o modelo

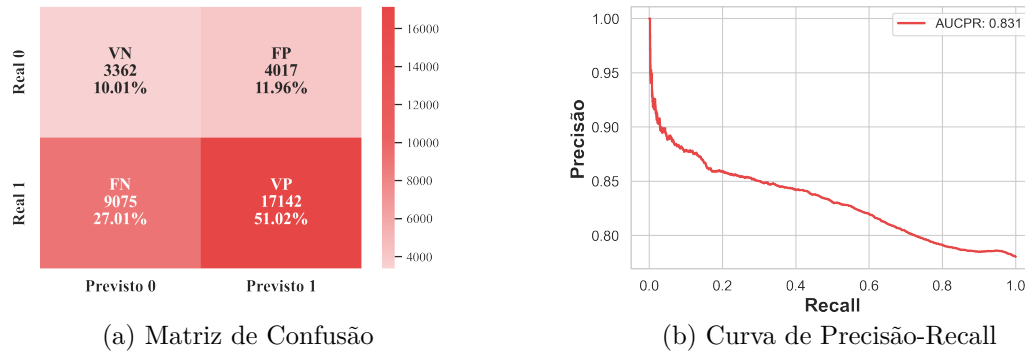


Figura 7 – Resultados do Modelo Randomized Search CV

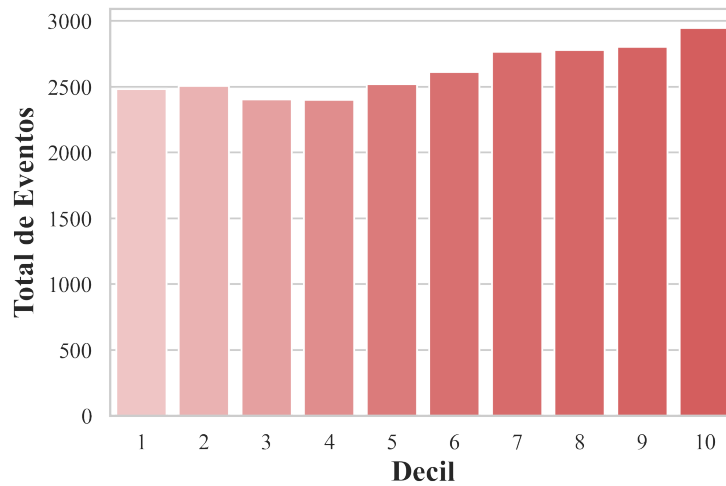


Figura 8 – Ordenação por Decis do modelo Randomized Search CV

de *Grid Search* possui valores superiores nas métricas básicas, quando comparado exclusivamente com o de busca aleatória, mas mantém a mesma ordenação superior, sendo, portanto, a melhor opção entre as analisadas.

Devido a estes resultados, na etapa posterior de construção e comparação dos portfólios de investimentos serão utilizadas apenas as probabilidades do modelo de *Grid Search*.

6.2 Comparação dos Portfólios

Uma vez que o melhor portfólio selecionado, quando otimiza-se pelo método de Markowitz, corresponde àquele cujo Sharpe é máximo, a primeira métrica utilizada na comparação dos portfólios foi deliberadamente esta.

Pode-se observar na Figura 12, exposta no Anexo A, um padrão com relação aos valores obtidos no decorrer dos 21 períodos: portfólios construídos disponibilizando ao

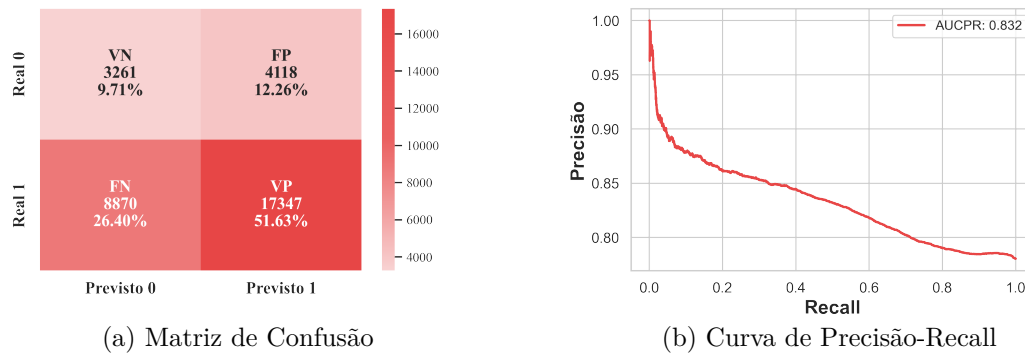


Figura 9 – Resultados do Modelo Grid Search CV

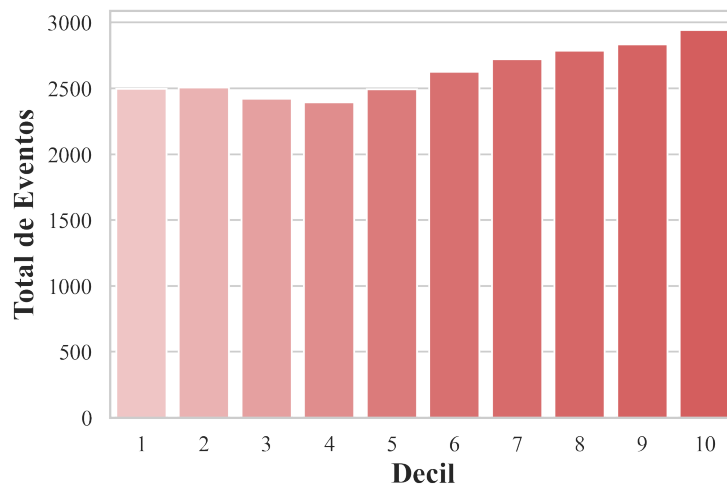


Figura 10 – Ordenação por Decis do modelo Grid Search CV

otimizador apenas os melhores ativos, segundo o pré-seletor, conseguem índices sharpes superiores ao portfólio sem limitações em alguns períodos, porém o resultado é inconclusivo. Este resultado era esperado uma vez que portfólios de momento apresentam mais risco.

Apesar de um conjunto pequeno de ativos - à primeira vista 20 aparenta ser o valor ótimo - ser capaz de obter valores de Sharpe superiores, observa-se contudo que esse resultado não é de todo estável. Em casos como o apresentado no período 3 verifica-se que a inclusão de um único ativo pode, subitamente, proporcionar um valor de Sharpe inferior ao obtido sem a influência do pré-seletor. Este fato é sintoma de uma necessidade de melhorias na capacidade de ordenação do modelo de pré-seleção.

A Figura 13 apresenta a mesma análise, porém normalizando os valores mínimos e máximos do gráfico, de modo a facilitar a comparação dos resultados entre os períodos.

Outra métrica observada durante o *backtest* foi o log-retorno total obtido, pelos diversos portfólios, ao longo dos 20 dias de cada período. Estes resultados podem ser vistos na Figura 14, também no Anexo A, onde novamente observa-se a vantagem dos

portfólios construídos com a intervenção do pré-seletor de ativos de investimento, porém não é possível determinar um número ótimo de ativos neste caso. A normalização, como mostra a Figura 15, dificulta a comparação dos resultados, uma vez que a faixa de valores do período 14 destoa significativamente das demais.

Buscando simular a utilização do método como estratégia de investimento, analisou-se o log-retorno acumulado ao decorrer dos 21 períodos de portfólios construídos a partir de um subconjunto de ativos - indicados como os melhores pelo modelo. Dessa forma o portfólio foi calibrado e seu resultado observado ao longo de 20 dias, quando então teria seus pesos recalculados de acordo com as novas informações obtidas. O resultado desta análise está exposto na Figura 11.

A análise mostra evidentemente que com apenas o melhor ativo obtém-se o melhor resultado total durante o período de *backtest*, mantendo-se com retorno acumulado consistentemente acima do portfólio sem restrições. O portfólio com os 5 melhores ativos possui um período de baixo resultado porém eventualmente também o supera, o que pode ser explicado pelo fato de portfólios de momento serem considerados estratégias de longo prazo.

Pode-se observar também que, apesar das quedas significativas de log-retorno em todos os portfólios durante os períodos 14 e 15 - o qual a bolsa brasileira mais foi impactada pela pandemia do COVID-19 - os com calibrados com influência do pré-seletor apresentaram *drawdowns* inferiores. Isto sugere que a estratégia pode ser utilizada para, além de seus propósitos já evidenciados, proteger o capital investido; sendo necessário, porém, maiores estudos neste sentido.

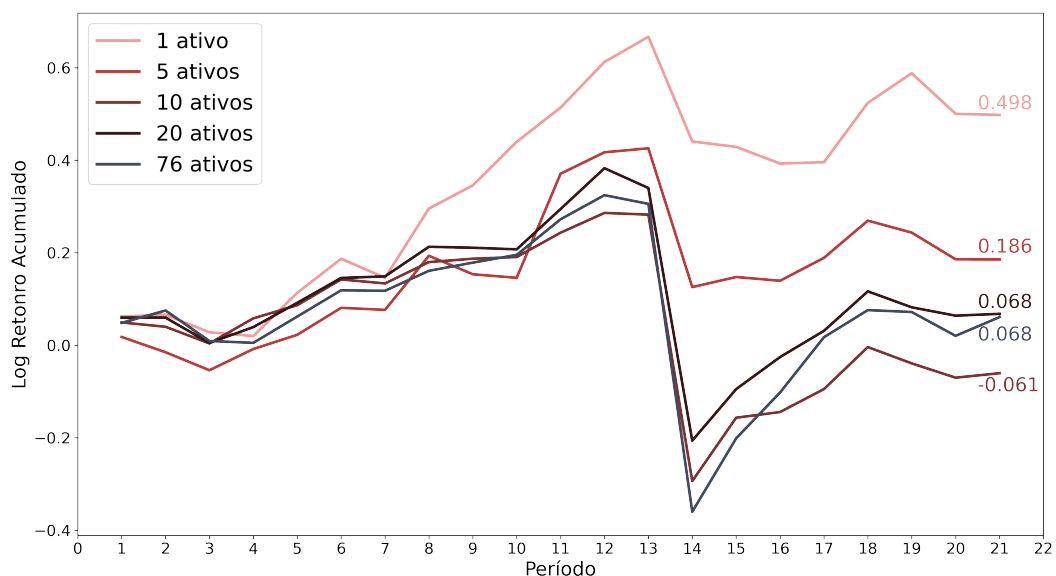


Figura 11 – Log-Retorno acumulado obtido ao longo de todos os períodos

7 CONCLUSÃO

O maior intuito deste trabalho foi demonstrar uma aplicação prática alternativa para modelos de *Machine Learning* no domínio das finanças, aproveitando o potencial de retorno da utilização do fator momento em portfólios de investimento.

É possível observar que de fato existe um espaço para a utilização deste tipo de modelo na pré-seleção de ativos. Este trabalho buscou, ao realizar uma implementação no mercado financeiro brasileiro, reforçar esta tese.

Notou-se durante o *backtest* empreendido a significativa vantagem de sua utilização nos retornos acumulados a longo prazo, uma vez que um portfólio construído com base em apenas 20 ativos financeiros foi capaz de obter um log-retorno acumulado igual ao de um feito com 76 ativos (0.068), como indicado na figura 11

Além disso a maioria dos portfólios construídos com a influência do pré-seletor obtiveram resultados superiores ao portfólio sem esta influência. Dentre as opções a que mais chama atenção é o contendo apenas 1 ativo, que, mesmo a curto prazo, possui log-retorno acumulado maior em quase todo o período de *backtest* (0.498).

Entretanto recomenda-se a utilização do portfólio que contempla os 5 melhores ativos, pois apresenta maior diversificação quando comparado ao de 1 ativo, mesmo que possua menor log-retorno acumulado a longo prazo (0.186), ainda que superior ao *benchmark* sem a influência do pré-seletor.

Apesar dos resultados positivos, porém, foi possível constatar a existência de diversas possíveis melhorias em sua utilização, uma vez que os modelos construídos apresentaram significativa instabilidade em alguns períodos analisados, como exemplificado pelo resultado do portfólio construído com base nos 10 melhores ativos, que obteve o pior log-retorno acumulado entre os avaliados (-0.061).

Uma dessas possíveis melhorias seria a construção e utilização de novas variáveis explicativas no modelo de pré-seleção, principalmente a inclusão do momento de prazo intermediário. Outras informações que podem ser investigadas em futuros trabalhos, são: análise de texto natural de informações do mercado obtidas de fontes como *google analytics* e *twitter*; correlação da série dos ativos no período com os resultados de grandes fundos de investimentos; indicadores de análise técnica, como, por exemplo, o MACD.

Outro possível foco para aprofundar a análise seria a experimentação de outros modelos de *Machine Learning*, como *XGBoost*, *Support Vector Machine*, entre outros; podendo eventualmente avaliar técnicas de *Deep Learning* como as redes neurais.

Importante ressaltar que os resultados foram obtidos considerando apenas a otimi-

zação de portfólios pelo método de Markowitz. É necessário, além dos aprofundamentos previamente indicados, avaliar a influência deste tipo de pré-seletor em outras técnicas de otimização de portfólios mais complexas e com menos limitações, como por exemplo o *Risk Budget* e *Black-Litterman*.

REFERÊNCIAS

- AMIT, Y.; GEMAN, D. Shape quantization and recognition with randomized trees. 1997.
- ASNESS TOBIAS J. MOSKOWITZ, L. H. P. C. S. Value and momentum everywhere. 2013.
- BARBERIS ANDREI SHLEIFER, R. V. N. A model of investor sentiment. 1998.
- BOYD KEVIN H. ENG, C. D. P. K. Area under the precision-recall curve: Point estimates and confidence intervals. 2013.
- BREIMAN, L. Bagging predictions. 1996.
- BREIMAN, L. Random forests. 1 2001.
- CARHART, M. M. On persistence in mutual fund performance. 1997.
- DA UMIT G. GURUN, M. W. Z. Frog in the pan: Continuous information and momentum. 2014.
- EMERSON, S. et al. Trends and applications of machine learning in quantitative finance. 5 2019.
- FABOZZI FRANCIS GUPTA, H. M. M. F. J. The legacy of modern portfolio theory. 2002.
- FAMA, K. R. F. E. F. Common risk factors in the returns on stocks and bonds. 1993.
- FAMA, K. R. F. E. F. A five-factor asset pricing model. 2015.
- FRAZZINI RONEN ISRAEL, T. J. M. A. Trading costs of asset pricing anomalies. 2014.
- FU, X. et al. A machine learning framework for stock selection. 06 2018.
- GRAY, J. R. V. W. R. *Quantitative Momentum: A practioner's guide to building a momentum-based stock selection system*. [S.l.]: Wiley, 2016.
- HASTLE ROBERT TIBSHIRANI, J. F. T. *The Elements of Statistical Learning: Data mining, inference, and prediction*. [S.l.]: Springer, 2009.
- HO, T. K. Random decision forests. 1998.
- HUANG, C.-F. A hybrid stock selection model using genetic algorithms and support vector regression. *Applied Soft Computing*, 2012.
- JEGADEESH, S. T. N. Returns to buying winners and selling losers: Implications for stock market efficiency. 1993.
- LEE, M.-C. Using support vector machine with a hybrid feature selection method to the stock trend prediction. 2009.
- LEVY, R. A. Relative strength as a criterion for investment selection. 1967.

MANGRAM, M. E. A simplified perspective of the markowitz portfolio theory. 2013.

MARKOWITZ, H. M. Portfolio selection. 1952.

MARKOWITZ, H. M. *Portfolio Selection*: Efficient diversification of investments. [S.l.]: Yale University Press, 1959.

PORDEUS, G. G. N. Seleção de carteiras por meio de machine learning e da influência da informação assimétrica. 10 2018.

RASEKHSCHAFTE, K.; JONES, R. Machine learning for stock selection. 2 2019.

SHARPE, W. F. Mutual fund performance. 1966.

SHARPE, W. F. The sharpe ratio. 1994.

ANEXOS

ANEXO A – RESULTADOS POR PERÍODO DE VALIDAÇÃO

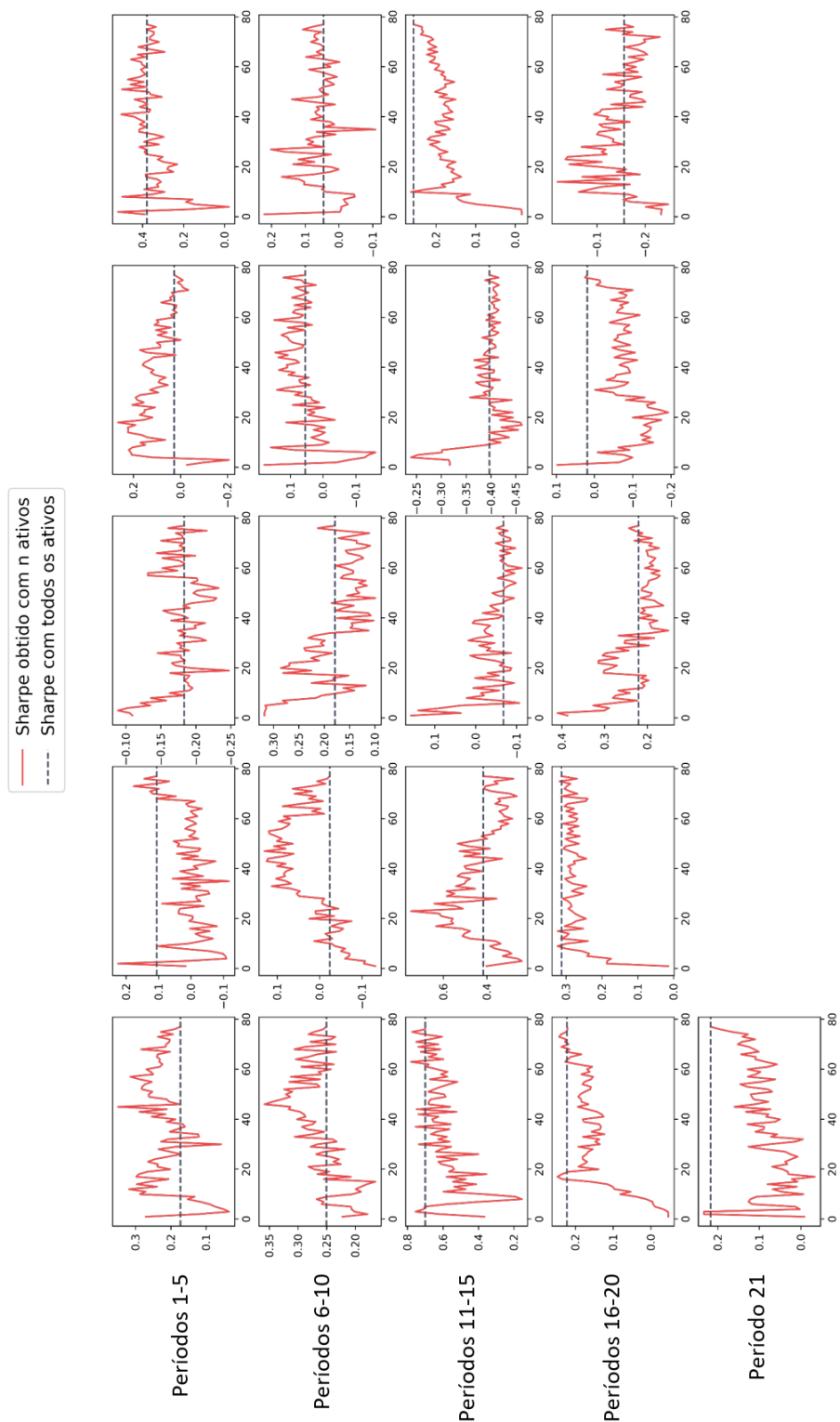


Figura 12 – Sharpe obtido de acordo com o número de ativos disponibilizados ao otimizador de portfólios

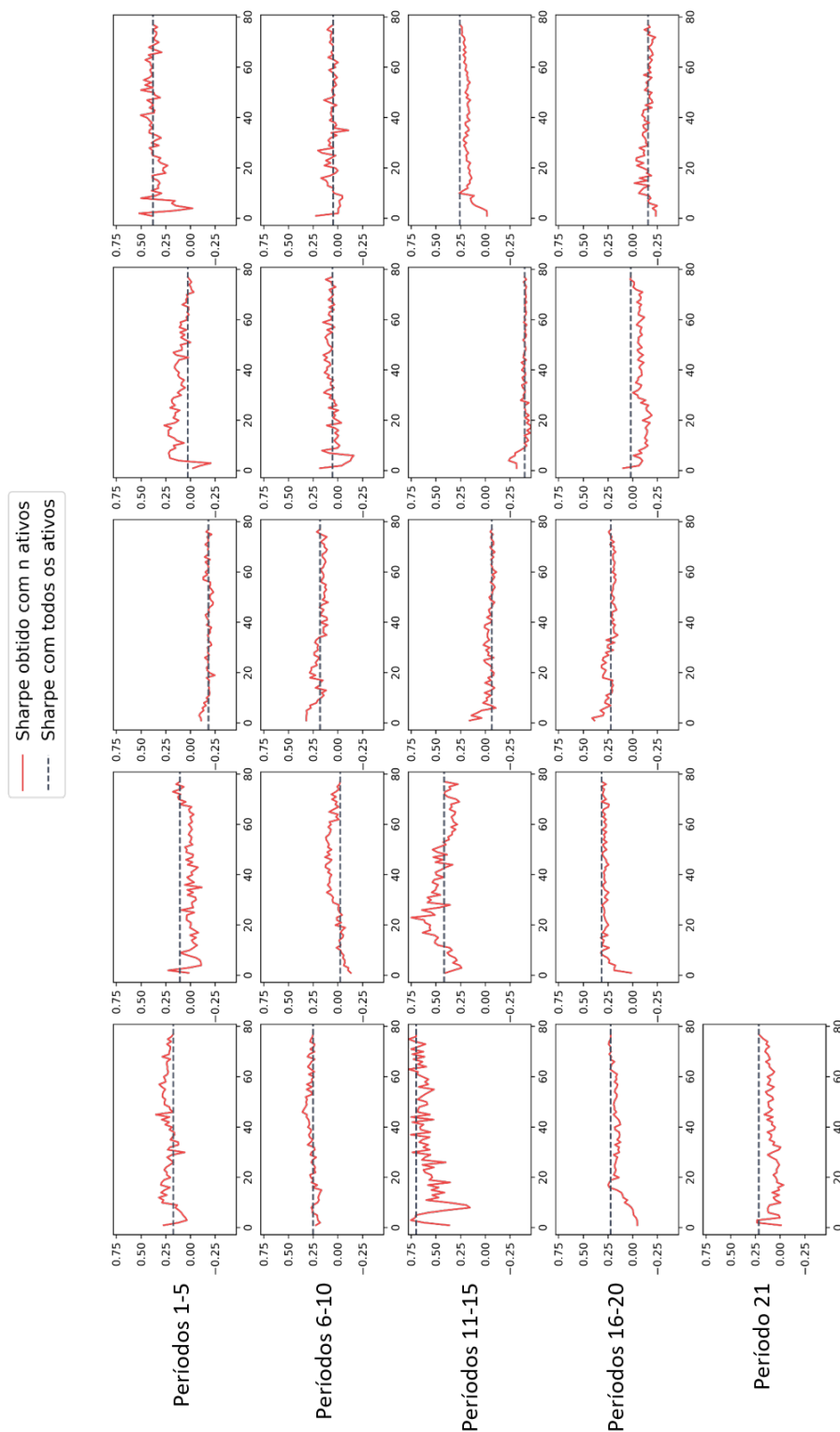


Figura 13 – Sharpe obtido de acordo com o número de ativos disponibilizados ao otimizador de portfólios com eixo normalizado

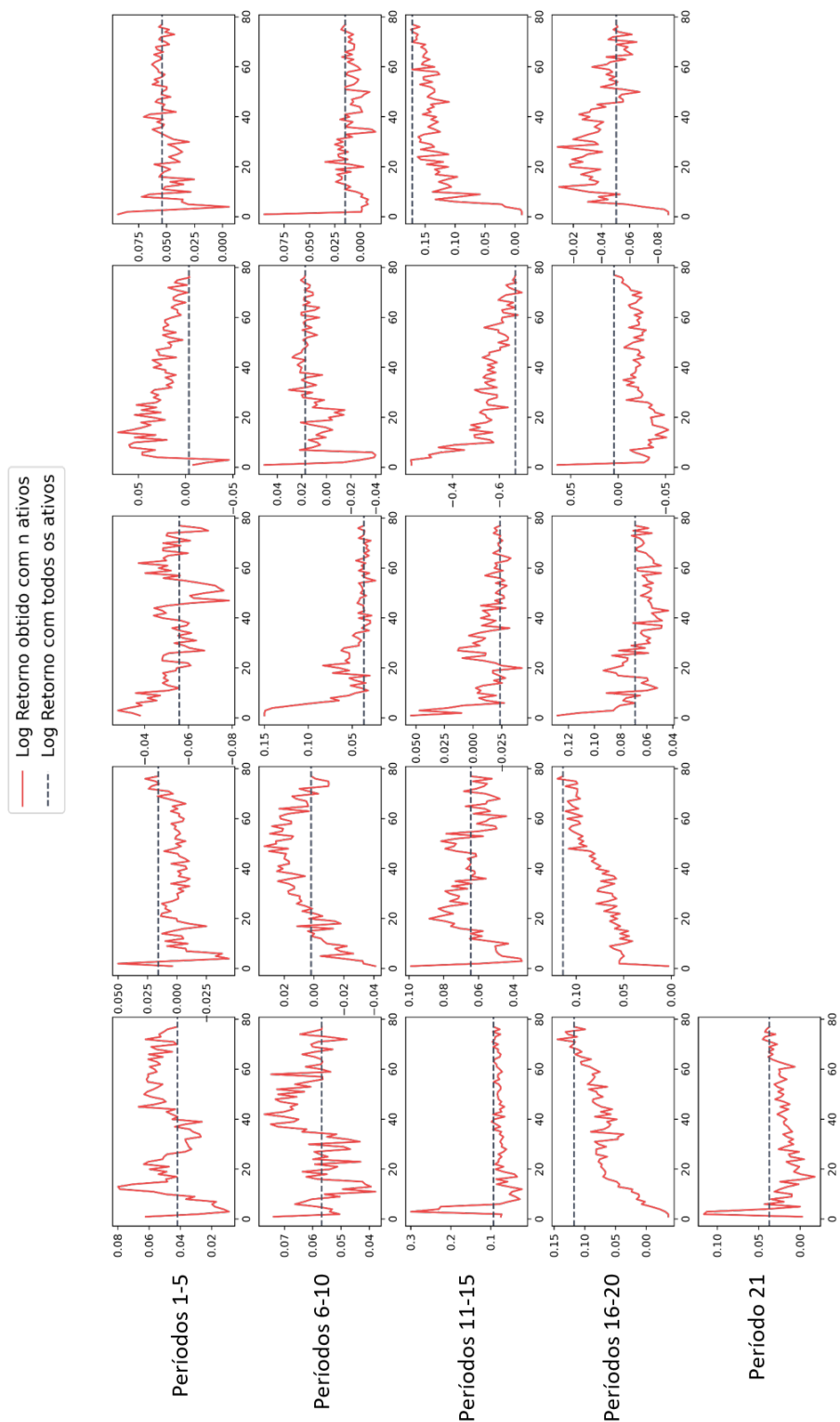


Figura 14 – Log-Retorno obtido de acordo com o número de ativos disponibilizados ao otimizador de portfólios

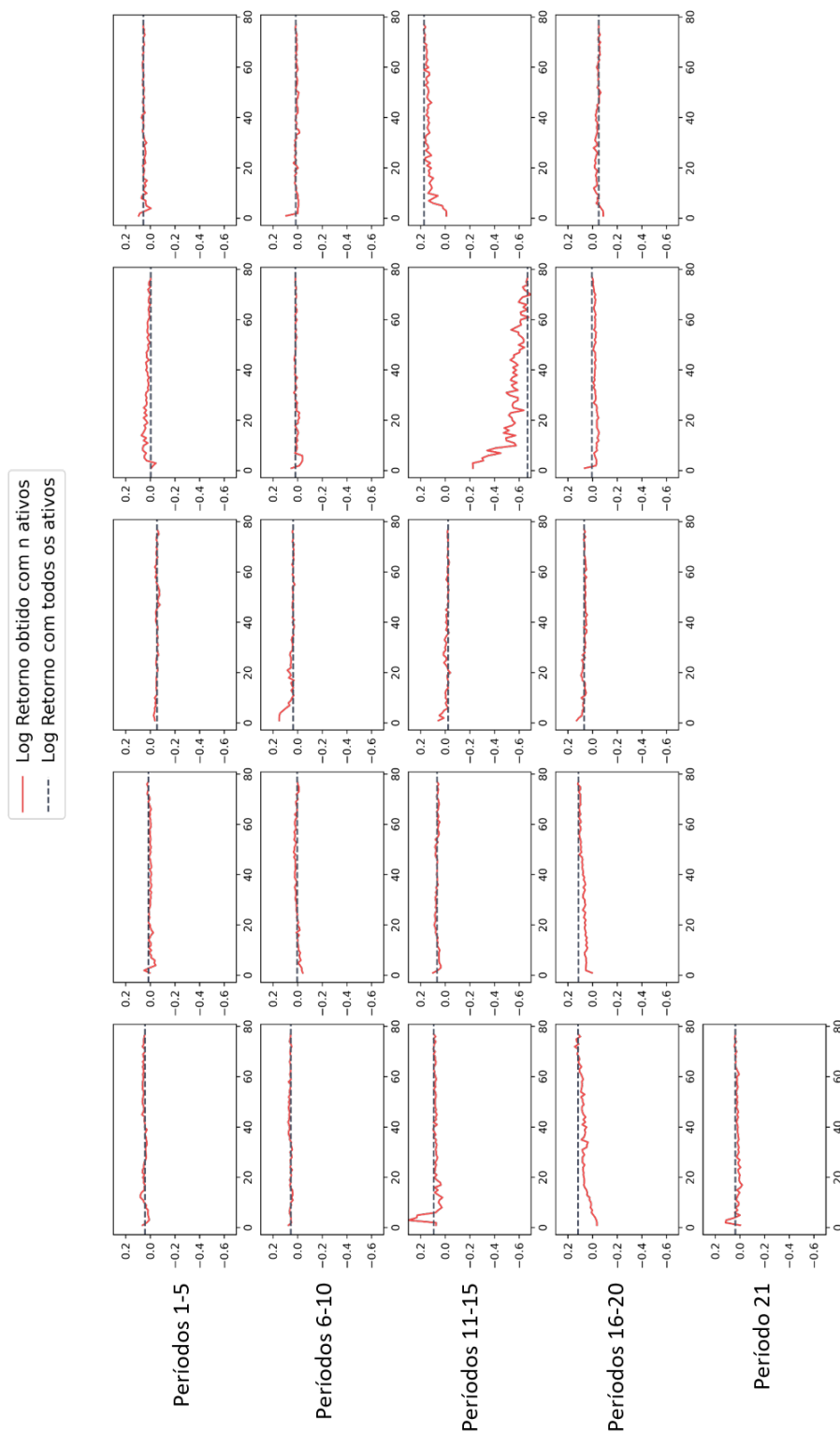


Figura 15 – Log-Retorno obtido de acordo com o número de ativos disponibilizados ao otimizador de portfólios com eixo normalizado