Previsão do Prêmio de Risco no Mercado de Ações Brasileiro Utilizando Notícias Financeiras, Fatores de Risco, Indicadores Técnicos e Aprendizado de Máquina Supervisionado

Elvira Helena Oliveira de Medeiros [*]
Lucas Lúcio Godeiro[†]
Kléber Formiga Miranda[‡]

Resumo

O objetivo do artigo é utilizar técnicas de mineração de dados textuais com aprendizado de máquina supervisionado recursivamente via Ridge, LASSO e *Elastic Net*, na construção do Índice de Sentimento para o mercado de ações brasileiro. A robustez do Índice foi verificada considerando seu poder explicativo sobre o prêmio acionário em relação ao *benchmark* (média histórica) do mercado no período dentro da amostra de 2008:01 a 2014:03 e fora da amostra de 2014:04 a 2021:01. Posteriormente foi feita a sua comparação por meio de Fatores de Risco e Indicadores Técnicos . Os resultados encontrados revelaram que o Índice de Sentimento proposto não consegue prever estatisticamente / economicamente o retorno acionário brasileiro. Por outro lado, o fator de risco *Winners Minus Losers* (WML) e cinco dos seis modelos da Média Móvel (MA) conseguem ganhos econômicos e financeiros, superando o *benchmark*.

Palavras-Chave: Aprendizado de Máquina Supervisionado; Sentimento Textual; Fatores de Risco; Indicadores Técnicos.

Abstract

The objective paper utilizes text mining techniques with supervised machine learning through Ridge, LASSO and Elastic Net in the development of a sentiment indexes to the Brazilian Stock Market. In order to test the sentiment indexes predictability over the Brazilian stock market equity premium, We performed out-of-sample forecasts in the period of 2014:04 to 2021:01. We found that the benchmark, the historical average, outperforms the forecasts generate by the sentiment indexes. In the other hand, the risk factor predictor winner minus loser(WML) and five out of six models generated by the moving average technical indicator predictors outperformed the historical average.

Keywords: Supervised Machine Learning; Textual Sentiment; Risk Factors; Technical Indicators.

Área 4 - Macroeconomia, Economia Monetária e Finanças
JEL classification: C53, E44, G17.

[*]Doutoranda em Economia Aplicada - UFJF e Mestre em Economia Aplicada - UERN. e-mail: ravilelenna@yahoo.com.br
[†]Doutor em Economia Aplicada - UFPB e Professor da Universidade Federal Rural do Semi Árido. e-mail: lucasgodeiro@ufersa.edu.br
[‡]Doutor em Ciências Contábeis - UFPB e Professor da Universidade Federal Rural do Semi Árido. e-mail: mirandakf@ufersa.edu.br

1 Introdução

O Prêmio de Risco acionário é o retorno incremental em relação a uma taxa livre risco que os agentes esperam receber quando investem seus recursos em ações. Em meados dos anos 60 e 70 a hipótese do mercado eficiente (HME) interpretava que o verdadeiro prêmio de risco era constante no tempo. Esse ponto de vista era associado ao uso da média histórica, sendo refutado por Fama and French (1988) quando sugeriram que o verdadeiro prêmio de risco ou retorno acionário não é um resultado fixo no tempo, podendo ser estimado por meio de dados observados.

Assim, dada a dificuldade de previsibilidade dos retornos, por não ser uma variável observável, o uso de dados econômicos ou Financeiros para inferí-los tem sido um assunto bastante discutido na econometria financeira. Rapach and Zhou (2013a) detectaram inúmeros preditores macroeconômicos com o intuito de analisar o seu resultado .Rapach and Strauss (2010) propuseram melhorar o prêmio de risco *out-of-sample* (OOS) combinando informações de previsões distintas com o intuito de observar quem possui o melhor poder de decisão ou se melhoram seu desempenho prevendo conjuntamente. Li et al. (2013) analisaram se o custo de capital seria uma boa *proxy* para medir o retorno no decorrer do tempo, descobrindo que esta variável prevê fortemente e o seu poder preditivo persiste mesmo durante ciclos de negócios tanto para os dados dentro e fora da amostra . Neely et al. (2014) verificaram se os indicadores técnicos conseguem prever de forma acurada quando comparado a dados macroeconômicos, verificando que um é complemento do outro.

Mais recentemente, economistas financeiros acrescentaram o índice de sentimento textual como um novo preditor Godeiro et al. (2018). Esses índices caracterizam-se como uma análise linguística das palavras pertencentes a uma fonte de conteúdo informacional que, por meio de técnicas específicas, conseguem um parecer subjetivo dos textos. Na literatura, essa fonte de conteúdo é conhecida por dados não estruturados. A sua utilização é caracterizada por meio de uma variedade de informações, originando em um grande armazenamento de elementos, chamado de *Big Data* Silva (2018). Dessa forma, tomadores de decisões como economistas financeiros e investidores individuais passaram a utilizar esse mecanismo como uma *proxy* para medir o sentimento do mercado. O sentimento positivo / negativo encontrado em jornais/artigos que possuem coluna sobre notícias financeiras tem sido um dos responsáveis pelas informações em massa, na hora dos agentes decidirem qual melhor decisão a tomar.

Normalmente, para se fazer tal análise utiliza-se de métodos aprimorados pelo aprendizado baseado em máquinas (*Machine Learning*) ou por meio do uso de dicionários. Um dos métodos utilizado em aprendizado de máquinas é o supervisionado recursivamente por meio de técnicas como: LASSO, Ridge e *Elastic Net*. Essas técnicas nos permitem criar um dicionário que mude conforme o teor das notícias no decorrer do tempo, permitindo captar o sentimento conforme as notícias naquele determinado momento. Já o procedimento baseado em dicionário determina o sentimento por meio de algoritmo que lê e classifica as palavras conforme categorias pré-estabelecidas, o que caracteriza que o sentimento seja constante no tempo. O desenvolvimento desse dicionário advêm com o Harvad General Inquirer (*GI/Harvard*) criado para o uso na psicologia, mas adaptado para o mercado financeiro. Porém, devido aos seus erros de especificações, foi posteriormente aprimorado apenas para conteúdo de finanças numa grande amostra *10-Ks* de empresas listadas na NYSE, Amex e NASDAQ entre 1994 a 2008. Este dicionário ficou conhecido na literatura como Loughran e McDonald (L& M), devido as iniciais dos seus autores

Alguns autores têm utilizado notícias financeiras para criar o índice de sentimento com o intuito de investigar até que ponto ele pode ser utilizado como uma *proxy* para verificar o sentimento do mercado. Dentre os primeiros trabalhos podemos destacar Cutler et al. (1988) analisando a ligação entre as notícias financeiras e as oscilações dos preços do mercado.Tetlock (2007) por sua vez, verifica se as notícias diárias

de uma coluna especializada em notícias financeiras do *Wall Street Journal* (WSJ) eram suficientes para prever pressões sobre os preços no mercado, por meio do uso do dicionário General Inquire (*GI/Harvard*) para os períodos de 1984 a 1999. Garcia (2013) construiu uma medida de sentimento diário entre 1905 a 2005, com base em duas colunas financeiras, *New York Times* e *Wall Street Journal*, para prever a relação entre o retorno das ações com as notícias do mercado financeiro durante períodos de recessões fazendo uso do dicionário fixo *Loughran e McDonald* (L& M).

Além deles encontramos Godeiro et al. (2018), que buscaram encontrar por meio das notícias financeiras e dados macroeconômicos o melhor preditor para o prêmio de risco que superasse o modelo de referência (*benchmark*) do mercado Americano. Para tal análise foram utilizados dados de notícias financeiras publicadas no "*The Wall Street Journal*"e "*The New York Times*" da base de dados *Dow Jones Factiva* e 15 variáveis macroeconômicas no período de janeiro de 1980 a dezembro de 2017. Os autores por meio de um processo recursivo dentro e fora da amostra e com o uso do aprendizado de máquina supervisionado via *Elastic Net* para selecionar as palavras mais preditivas conseguiram observar que o modelo baseado em notícias proporcionam ganhos tanto estatístico quanto econômico.

Isto posto, o presente artigo tem por objetivo utilizar técnicas de mineração de dados textuais com aprendizado de máquina supervisionado recursivamente com as técnicas Ridge, LASSO e *Elastic Net* para constuir um índice de sentimento a partir de palavras que mudam ao longo do tempo visando observar se o mesmo prevê de forma acurada o prêmio de risco acionário brasileiro quando comparado ao *benchmark* (média histórica) do mercado. Os dados foram calculados mensalmente, sendo de janeiro de 2008 a janeiro de 2021, permitindo a sua comparação com outros preditores. Dessa forma, foi feita a sua comparação com cinco Fatores de Risco Fama and French (2015) e 14 Indicadores Técnicos proposto por Neely et al. (2014). Além disso, dividimos a amostra dentro da amostra de 2008:01 a 2014:03 e fora da amostra de 2014:04 a 2021:01.

Além desta parte introdutória, o artigo está dividido em mais três seções: a seguir será apresentada de forma detalhada a metodologia. Posteriormente, serão apresentados resultados e discussões e por fim, na última seção são apresentadas as conclusões

2 Metodologia e Dados

2.1 Filtragem de Palavras mais Preditivas

Para a determinação do índice de sentimento textual, seja ela aplicada ao setor financeiro ou qualquer área da Economia, deve-se levar em conta a escolha de quais notícias serão utilizadas. A escolha dessas notícias ajudam na eliminação de erros de precificações. Dessa forma, para prever o prêmio de risco acionário brasileiro foram utilizadas notícias financeiras como *proxy* para análise de sentimento. Nesse sentido, foi feito junto ao "Banco de Dados FACTIVA" a busca por esses textos. Os arquivos foram baixados com informações mensais de janeiro de 2008 a janeiro de 2021 e os noticiários são jornais de grandes circulação no Brasil, como podemos destacar: o "Globo", "Estadão"e "Valor Econômico".

Posteriormente, após baixar os dados salvaremos em um "corpus", segundo (Godeiro et al., 2018, p. 07) " um corpus é uma coleção de textos escritos, ou seja, um conjunto de notícias financeiras ". Porém é sabido que grandes jornais de circulações além desses noticiários contêm outros tipos de notícias. Dessa forma, com o uso do *Software R-Stúdio*, foi aplicado uma filtragem de todos os textos buscando os termos com referência apenas financeiras.

- **Filtro de Fontes:** "Apenas Notícias Financeiras". Este filtro especifica que queremos apenas as

notícias do "O GLOBO", "Estadão"e "Valor Econômico".

- **Filtro de Assunto:** "Mercado Financeiro ". Este filtro específica que queremos apenas as notícias "Mercado Financeiro", a fim de evitar reportagens não relacionadas ao mercado financeiro.
- **Filtro de Idiomas:** "Portuguese". Este filtro especifica que queremos apenas os textos escritos em português.

Uma forma de diminuir ou até mesmo eliminar palavras distintas e deixando somente as que estejam dentro do contexto específico é fazendo um pré-processamento bruto dos dados textuais por meio de algumas etapas. A primeira etapa que também foi adotada por Godeiro et al. (2018) tem o objetivo de classificar palavras que contenham sequências de termos agrupadas para um único termo. Posteriormente, utilizando da programação *R-Stúdio* podemos encontrar pacotes que nos ajudem a selecionar as palavras mais preditiva. Por meio do pacote *"tm"* removeremos os *"stopwords"* como : o, aquele,qual, o que, entre outras; para remover números *(removenumber)*; para capturar palavras que tenham o mesmo sentido *(steming)* como por exemplo, "Economia" e "Econômico" que pode ser contabilizado para uma única palavra *"Econom"*. Adotando o método em Godeiro et al. (2018) *apud* Hansen et al. (2017) as palavras restantes foram classificadas usando o termo frequência-inverso (*tf-idf*) que tem como característica punir palavras raras, também foi descartada palavras selecionadas como 100 ou menor que 100.

Por conseguinte, após a realização dessa etapa foi observado uma redução de palavras que são agrupadas em colocações, (j) mensalmente (s) numa série de tempo $X_{j,t} = \{X_{j,s}\}_{s=1}^{t}$, sendo *s=1,....,p* e *j=1,....,t* . Ainda, segundo Godeiro et al. (2018) foi feito uma coleta de todas as séries temporais por meio de uma matriz de tamanho $p \times t$, $X_t = (X_{1,t,...,}X_{p,t})'$ em que as linhas p de X_t correspondem aos termos de notícias financeiras p (palavras ou colocações). Na prática, o número de palavras e colocações (p) são bastante grandes e podem até ser maior que o tamanho da amostra, ou seja, $p >> T$. Isso cria um problema de alta dimensão que pode ser resolvido usando métodos de regularização, como o Hidge, LASSO e *Elastic Net*.

2.2 Uma lista de Palavras Variantes no Tempo

Nesta seção, explicamos como foi construído o índice de sentimento usando apenas as palavras mais preditivas do prêmio de risco acionário brasileiro. Usamos a abordagem de sacola de palavras ("*bag of words*") para quantificar o tom textual nos documentos, mas não dependemos de uma lista predeterminada de palavras. Em vez disso, nossa abordagem usa aprendizado de máquina supervisionado recursivamente para selecionar as palavras mais preditivas ao longo do tempo. Para fazer isso, regredimos o prêmio de risco r_{t+1} no saco de palavras X_t encontrado na seção anterior.

$$r_{t+1} = X_t' \boldsymbol{\beta_t} + \epsilon_{t+1} \tag{1}$$

onde β_t é estimado minimizando a seguinte função de objetivo:

$$\min_{\boldsymbol{\beta}_t} \sum_t (r_{t+1} - X_t' \boldsymbol{\beta_t})^2 + \lambda_1 \parallel \boldsymbol{\beta_t} \parallel_{\ell_1} + \lambda_2 \parallel \boldsymbol{\beta_t} \parallel_{\ell_2} \tag{2}$$

Com $\|\cdot\|_{\ell_1}$ e $\|\cdot\|_{\ell_2}$ e $\|\cdot\|_{\ell_2}$ indicando as normas ℓ_1 e ℓ_1 , respectivamente. Quando $\lambda_1 = \lambda_2 = 0$ então a função objetivo (2) torna-se igual à soma usual dos resíduos quadrados. Quando $\lambda_2 = 0$, então (2) se torna o chamado estimador LASSO. Embora *LASSO* seja bem sucedido na seleção de variáveis, no

caso particular em que o número de palavras é maior que o tamanho da amostra, ie $p >> T$, *LASSO* seleciona no máximo T palavras antes de saturar, excluindo, portanto, grandes porções do conjunto de informações de condicionais e reduzindo potencialmente a precisão das previsões. O caso com $\lambda_1 = 0$ corresponde à Ridge , que não faz seleções de modelo porque não reduz os coeficientes para zero. Por estas razões, Zou and Hastie (2005) sugeriu usar uma combinação de restrição-ℓ_1 e restrição-ℓ_2 que corresponde ao chamado estimador de "*Elastic Net*". Nesse caso, os coeficientes $\hat{\beta}_t$ são reduzidos para zero de duas maneiras diferentes, promovendo tanto a esparsidade como a estabilidade. Isso evita o conhecido "*overfitting*"[1] nos dados ao definir os coeficientes sem importância como zero e identifica apenas as palavras preditivas mais relevantes. Os valores ótimos de λ_1 e λ_2 são obtidos a partir do procedimento sugerido no pacote GLMNET R (seção de regressão linear) desenvolvido por Trevor Hastie e Junyang Qian. [2]

Na literatura de previsão, uma pequena lista de trabalhos que empregaram com sucesso *Elastic Net* incluem Bai and Ng (2008), Li et al. (2015) e Lima et al. (2019). Com o objetivo de utilizar supervisionamento de máquina, o artigo utilizou as técnicas de regularização, Ridge, LASSO e *Elastic Net* para classificar as palavras mais preditivas no tempo. Em nosso exercício de previsão fora da amostra, a equação (2) é estimada recursivamente até final da amostra, ou seja, para cada origem da previsão $t = R, ..., T-1$ [3], regredimos as observações r_{s+1} em X_s para $s = 1, ..., t-1$, implicando que o vetor de coeficiente β_t tem permissão para mudar com o tempo. Esta estimativa recursiva da equação (2) produz a seguinte regra de classificação dinâmica:

(a) excluímos séries temporais neutras(palavras), ou seja, excluímos a série temporal na qual o coeficiente estimado na equação (2) é igual a zero. Este é um procedimento padrão no aprendizado de máquina. As séries temporais restantes(palavras) serão salvas em uma nova matriz (conjunto de palavras) X_t^*. Ao contrário de X_t, as linhas de X_t^* podem ser interpretadas como a lista das palavras mais preditivas disponíveis até o momento t. Além disso, como a estimativa ocorre recursivamente para cada $t = R, ..., T-1$, o valor de β_t e, portanto, o conteúdo da lista, mudará com o tempo, gerando um tempo lista variável de palavras.

(b) classificamos as séries temporais(palavras) como notícias financeiras "positivas" se seu coeficiente correspondente na Equação (2) for positivo. A ideia aqui é que um coeficiente positivo sugere uma correlação positiva parcial entre o prêmio de risco acionário no momento $t+1$ e a respectiva palavra no tempo t;

(c) classificamos as séries temporais(palavras) como notícias financeiras "negativas" se seu coeficiente correspondente na Equação (2) for negativo. A ideia aqui é que um coeficiente negativo sugere uma correlação negativa parcial entre o prêmio de risco acionário no momento $t+1$ e a respectiva palavra no tempo t;

As palavras selecionadas de acordo com a regra de classificação dinâmica acima farão a lista das palavras mais preditivas. A frequência com que essas palavras selecionadas aparecem ao longo do tempo é salva em X_t^*, que nada mais é do que uma matriz contendo a série temporal mais preditiva com tom textual positivo e negativo disponível até t. Mais importante, dada a estimativa recursiva de (2), o valor de β_t e, consequentemente, o conteúdo de X_t^* muda ao longo do tempo, permitindo que novas palavras sejam

[1] No aprendizado de máquina, o *overfitting* ocorre quando dividimos a amostra em treino e teste e o modelo decorou tão bem a primeira parte que não generaliza para a etapa seguinte.

[2] Nós empregaremos validações cruzadas para dados dependentes como em Elliott and Timmermann (2013). Este procedimento de validação cruzada executa uma seleção *ex-ante* dos parâmetros de ajuste λ_1 e λ_2, o que é essencial para evitar *overfitting*.

[3] Terminamos em $T-1$ porque precisamos usar a observação T para avaliar as previsões feitas em $T-1$

selecionadas e não selecionadas à medida que nos aproximamos do final da amostra. Finalmente, observe que X_t^* é carregado com notícias positivas ou negativas, o que implica que, para instante de tempo t, o número de palavras positivas é necessariamente igual a um menos o número de negativas.

Com a regra de classificação acima em mente, a etapa final é calcular um índice de sentimento. Definimos um índice de sentimento positivo SI_{TV}^{pos} como o número de palavras positivas em X_t^* dividido por um mais a contagem total de palavras, em que TV é usado para denotar que o índice de sentimento é calculado a partir de uma lista de palavras que varia no tempo. Da mesma forma, definimos um índice de sentimento negativo, SI_{TV}^{neg}, como o número de palavras negativas em X_t^* dividido por um mais a contagem total de palavras. Como o número de palavras positivas em X_t^* é idêntico a um menos o número de palavras negativas, temos que SI_{TV}^{pos} e SI_{TV}^{neg} terá um índice de correlação próximo ao negativo.[4] Por esse motivo, utilizamos apenas SI_{TV}^{neg} em nosso exercício de previsão. Finalmente, uma média móvel de quatro meses do índice é aplicada para remover sazonalidade e saltos idiossincráticos e o índice resultante é então padronizado para ter média zero e variância unitária para facilitar a interpretação e comparação entre outros preditores.

2.3 Outros Dados

2.3.1 Fatores de Risco

Com o intuito de verificar a veracidade do Índice de Sentimento apresentado, SI_{TV}^{neg} foi feito a sua comparação com outros dados. O primeiro a ser utilizado foram os Fatores de Risco proposto por Fama and French (2015). Este modelo fatorial tem como propósito captar anomalias do mercado que não foram observadas pelo CAPM. Ele foi utilizado a fim de simular os fatores de risco brasileiro. Inicialmente o mesmo foi um aperfeiçoamento do Modelo de Três Fatores de Risco Fama and French (1993) acrescentando a ele mais dois fatores de mercado: Lucratividade e Investimento. Inicialmente seu modelo foi proposto da seguinte forma:

$$R_{it} - Rf_t = \alpha_i + \beta_i(R_m - Rf_t) + s_i SMB_t + h_i HML_t + r_i RMW_t + c_i CMA_t + \epsilon_{it} \qquad (3)$$

Em que: $R_m - Rf_t$ é o fator mercado, calculado pela diferença entre o retorno de todo o mercado e taxa livre de risco; SMB_t é fator tamanho, representado pela diferença entre os retornos de carteiras de ações com baixo valor de capitalização (**S**mall) e carteiras com alto valor de capitalização (**B**ig); HML_t é fator de valor, mensurado pela diferença entre os retornos de carteiras de ações com alto valor de índice *book-to-market* (**H**igh) e carteiras com baixo valor de índice *book-to-market* (**L**ow); RMW_t é o fator de desempenho, relativo à diferenaça entre os retornos de carteiras de ações com lucratividade operacional robusta (**R**obust) e ações com lucratividade operacional fraca (**W**eak); CMA_t é fator de intensidade na expansão dos ativos, calculado pela diferença entre os retornos de carteiras diversificadas de ações de empresas conservadoras (**C**onservative) com menor intesidade de expensão dos ativos e ações de empresas agressivas (**A**greesive) com maior expansão; s_i, h_i, r_i e c_i representam os coeficientes de cada fator para estimação do retorno esperado.

A inclusão dos fatores de risco na previsão do prêmio de risco do Ibovespa ($Retorno_{ibov,t} - Rf_t$ parte da noção de que cada fator de risco pode influenciar de forma diferente no retorno futuro das

[4] Esta correlação será exatamente igual à negativa se dividirmos o número de palavras positivas (negativas) pela contagem total de palavras, em vez de 1 mais a contagem total de palavras

ações componentes do índice ibovespa. Assim, os fatores foram avaliados individualmente em relação ao retorno do Ibovespa, permitindo verificar qual dos fatores contribuem mais para a geração do prêmio para se investir na carteira do Ibovespa. Os fatores de risco foram coletados no sitio *web* do Centro Brasileiro de Pesquisa em Economia Financeira da Universidade de São Paulo (NEFIN) [5].

Dada a exigência de liquidez imposta pelos critérios para seleção da amostra do NEFIN, o fator mercado apresentou alta correlação em relação ao prêmio de risco do Ibovespa. Ainda assim, as análises foram realizadas, dada a utilização de defasagem nas relações, considerando a pouca memória entre os retornos.

2.3.2 Indicadores Técnicos

Na literatura de previsão do prêmio de risco acionário com o uso da construção de um índice de sentimento é comumente observado a sua comparação com variáveis econômicas vinculadas aos fundamentos macroeconômicas. Contudo, variáveis técnicas que são amplamente utilizadas por profissionais do mercado financeiro estão sendo vistos como bons indicadores, dado o seu valor de prever estatisticamente e economicamente.

Os estudiosos dessa técnica denotam que o seu desenvolvimento advêm inicialmente por Charles Down (1851-1902), pioneiro da análise técnica, tinha como principal pressuposto analisar comportamentos e padrões dos níveis de preços. Tais procedimentos por sua vez, eram observados por meio de índices como indicativo de mercado, conforme ainda visto nos dias atuais. Segundo Neely et al. (2014) esse indicador toma como base padrões de preços e volumes anteriores visando identificar persistência no futuro.

É sabido que na análise técnica tais comportamentos ocorrem por meio de demonstrações gráficas que são comumente observadas por tendências. Significativamente, essas tendências possuem características próprias, dentre as quais podemos destacar: a Média Móvel e Volume Financeiro. A primeira, caracteriza-se por meio de linhas gráficas que são sucessivamente as médias dos níveis de preços de um determinado período, sendo estas calculadas a partir de intervalos específicos de tempos. Pequenos intervalos são conhecidos por média móveis curtas e intervalos maiores são denotados por médias móveis longas.

Já o segundo por sua vez, foi inicialmente introduzido por Joseph Granville (1960) demonstrando se o Volume Financeiro de um determinado ativo cresce ou diminui no decorrer do tempo, Gonçalves (2018). Quando era observado um aumento havia divulgação de que o volume era alto, caracterizando possivelmente em aumento dos níveis de preços. Já quando o volume era menor, os níveis de preços diminuíam, originando em uma menor transação.

Sendo assim, o artigo comparou o Índice de Sentimento proposto, SI_{TV}^{neg} com a técnica adotada no trabalho de Neely et al. (2014), onde os mesmos constroem quatorze Indicadores Técnicos a partir de três categorias: Média Móvel (MA), Momento (MOM) e Volume (VOL). Para isso, foram utilizados dados do Índice IBOVESPA considerado o principal indicador de desempenho do mercado Brasileiro. Nesse sentido, foi feito junto ao site do *Yahoo Finance* uma busca por tais indicadores.[6]

A Média Móvel[7] (MA) tem como objetivo dá o valor de compra e venda de uma determinada ação S_i no tempo $t(0,1)$ conforme apresentado na equação abaixo.

[5] http://nefin.com.br/risk_factors.html

[6] A estimação dos ocorreu com o *Software R- Stúdio* através dos pacotes *"quantmod"*, *"xts"*, *"BETS"* e *"CARET"*.

[7] As combinações para cada classificação dos Indicadores Técnicos (Média Móvel, Momento e Volume) são apresentados no Apêndice 2.

$$S_{i,t} = \begin{cases} 1 se MA_{s,t} \geq MA_{l,t} \\ 0 se MA_{s,t} \leq MA_{l,t} \end{cases} \tag{4}$$

Quando

$$MA_{i,j} = \frac{1}{j} \sum_{i=0}^{j-1} P_{t-i} para j = s, l \tag{5}$$

Onde: P_t é o nível de preço de uma determinada ação S_i; a Média Móvel *(MA)* é representada como longa *(l)* e curta *(s)*, sendo *(s < l)* denotando que uma MA_s será mais sensível ao movimento de preço do que MA_l. Sendo assim, quando um determinado preço começar a subir a MA_s será mais sensível que a MA_l caracterizando sinal de compra que é representada pelo valor 1 na equação acima e um valor de venda quando o resultado for 0. Os valores atribuídos as médias móveis (MA) são: MA_s(1,2,3) e MA_l (9,12).

Já os Indicadores de Momentos (MOM) são os valores de compra e venda de uma ação, S_i. Automaticamente, quando o preço de uma ação S_i no período atual *(t)* é superior aos preços *(t-m)* indica que o momento *(t)* é positivo, gerando um sinal de compra da ação, S_i. Assim, o modelo de Momento (MOM) é representado pela seguinte equação:

$$S_{i,t} = \begin{cases} 1 se P_t \geq P_{t-m} \\ 0 se P_t \leq P_{t-m} \end{cases} \tag{6}$$

Já com relação ao Volume (VOL), segundo Neely et al. (2014) especialistas da análise técnica constantemente utilizam o volume financeiro em grupo com os preços dos períodos anteriores visando identificar tendência no mercado. Assim, nossa estratégia termina acrescentando esse Volume (OBV) da seguinte forma:

$$OBV_t = \sum_{k=1}^{t} VOL_k D_k, \tag{7}$$

Onde: OBV_t é o retorno do Índice IBOVESPA no momento *t*; VOL_k é o volume no período *k* e D_k é um binário com valores P_k e $P_k \leq$ *k* e -1 caso contrário. Assim, conforme o que foi descrito será apresentado um sinal de negociação para o o Índice IBOVESPA, OBV_t da seguinte forma:

$$S_{i,t} = \begin{cases} 1 se MA_{s,t}^{OBV} \geq MA_{l,t}^{OBV} \\ 0 se MA_{s,t}^{OBV} \leq MA_{l,t}^{OBV} \end{cases}, \tag{8}$$

Quando

$$MA_{i,j}^{OBV} = \frac{1}{j} \sum_{i=0}^{j-1} OBV_{t-i}, para j = s, l. \tag{9}$$

Se a média móvel MA_s do IBOVESPA for maior que a média móvel MA_l indicará momento favorável de compra, caso contrário representará sinal de venda.

2.4 Prevendo Fora da Amostra

Especificamente, dividimos a amostra total de observações de $T = R + P$ em porções dentro da amostra e fora da amostra. As observações dentro da amostra abrangem 1 a R, enquanto as observações fora da amostra abrangem $R+1$ até T para um total de P previsões fora da amostra. Para cada origem de previsão $t = R, ..., T-1$, estimamos recursivamente as regressões preditivas fora da amostra regredindo r_{s+1} sobre a constante e um único preditor observado w_s^i para $s = 1, ..., t-1$ e, portanto, calculamos a previsão como $\hat{r}_{t+1} = \hat{\alpha}_t + \hat{\phi}_t w_t^i$, onde $\hat{\alpha}_t$ e $\hat{\phi}_t$ são o estimativas OLS e w_t^i denotam um dos preditores (índice de sentimento, fator de risco ou preditor técnico) introduzidos nas seções anteriores. Observe que, a estimação recursiva da equação de previsão implica que os coeficientes α_t e ϕ_t têm permissão para mudar ao longo do tempo.

Os dados na janela de estimativa inicial começa em 2008.01 e termina em 2014.03 (75 observações). A previsão fora da amostra varia de 2014.04.1 a 2021.01, totalizando $P = 79$ observações fora da amostra. Dividindo a amostra dentro e fora da amostra nos permitiu estimar com precisão os dados iniciais e ter um período maior fora da amostra.

Na análise fora da amostra, as observações do índice de sentimento proposto, SI_{TV}^{neg}, também são computadas recursivamente. Em outras palavras, usamos a amostra de estimação inicial 2008.1-2014.03 para estimar a equação (2) a partir da qual classificamos as palavras para o período 2008.8-2014.03 e construímos o índice SI_{TV}^{neg}. Em seguida, adicionamos a observação 2014.04 à amostra de estimação, re-estimamos a equação (2) e classificamos as palavras. Esta nova classificação foi usada para gerar as observações de 2014.04 de SI_{TV}^{neg}. Isso continua até que recursivamente geramos a observação 2021.01 de SI_{TV}^{neg}. Assim, na análise fora da amostra, as palavras são recursivamente classificadas como positivas, negativas ou neutras à medida que nos aproximamos do final da amostra. Isto implica que os valores de SI_{TV}^{neg} no período fora da amostra são calculados levando-se em conta a lista mais preditiva de palavras disponíveis até a origem prevista $t = R, ..., T-1$.

Nosso procedimento de avaliação de previsão é baseado no R^2, R_{OS}^2 fora da amostra, que compara a previsão fora da amostra do modelo de previsão $\hat{r}_{t+1} = \hat{\alpha}_t + \hat{\phi}_t w_t^i$, para a previsão de *benchmark*[8] representada pela média histórica,[9] $\overline{r}_{t+1} = \frac{1}{t}\sum_{s=1}^{t} r_s$. Segundo Welch and Goyal (2008), a média histórica é uma boa referência de ser passada.

Dessa forma, o R^2 fora da amostra é calculado da seguinte forma:

$$R_{OS}^2 = 1 - \frac{\sum_{t=R}^{T-1}(r_{t+1} - \hat{r}_{t+1})^2}{\sum_{t=R}^{T-1}(r_{t+1} - \overline{r}_{t+1})^2}.$$

Se a previsão $\hat{r}_{t+1}$ supera a previsão do *benchmark* então $R_{OS}^2 > 0$. Por esse motivo, a estatística R_{OS}^2 está medindo a redução no erro de previsão quadrática média (MSPE) dos modelos de previsão em relação ao modelo de referência (média histórica). Em prática, relatamos o valor de R_{OS}^2 em termos percentuais que implica em multiplicar R_{OS}^2 por 100, ou seja, $R_{OS}^2\,(\%) = 100 \times R_{OS}^2$.

Ainda, além de observar a magnitude desta estatística verificamos se um determinado modelo preditivo possui uma estatística significante. Para isso aplica-se o método adotado por DIEBOLD and MARIANO (1995). No entanto, devido ao fato de estarmos trabalhando com um modelo aninhado tal método

[8] Nosso modelo *benchmark* é representada pela média histórica.

[9] Este (*benchmark*) nos permite uma visão razoável do comportamento futuro do mercado acionário brasileiro

se torna inviável. Assim, para modelos aninhados Clark and West (2007)[10] aprimoraram o modelo de DIEBOLD and MARIANO (1995) denominado de Erro Quadrado Médio de Previsão Ajustado (MSPE) que significa testar a hipótese nula contra a hipótese alternativa. Logo, tal teste pode ser estimado da seguinte forma:

$$g_{t+1} = (r_{t+1} - \overline{r}_{t+1})^2 - \left[(r_{t+1} - \hat{r}_{t+1})^2 - (\overline{r}_{t+1} - \hat{r}_{t+1})^2\right], \tag{10}$$

Desta forma, encontramos o MSPE regredindo a série $\{g_{t+1}\}_{t=R}^{T-1}$ sobre o intercepto e calculamos sua estatística t. Logo, podemos verificar se as previsões baseadas em modelos $\hat{r}_{t+1}$ têm um MSPE significativamente menor que o *benchmark* que corresponde a testar a hipótese nula de que $R^2_{OS} \leq 0$ contra a hipótese alternativa de que $R^2_{OS} > 0$.

Na literatura de previsão de retorno, os valores percentuais de $R^2_{OS}\,(\%)$ são tipicamente pequenos, mas isso não significa que seu valor econômico é insignificante. De fato, como argumentado por Campbell and Thompson (2008), mesmo um $R^2_{OS}\,(\%)$ muito pequeno como 1.0 % para dados mensais podem ainda sinalizar um grau de previsibilidade do retorno economicamente significativo em termos de aumento do retorno anual da carteira para um investidor de média variância. Calculamos esse valor econômico de uma previsão pelo conhecido *equivalente de certeza do retorno (ou ganho de utilidade)*, que pode ser interpretado como a taxa de administração que um investidor está disposto a pagar para ter acesso às informações adicionais fornecidas pelo modelo de previsão baseada no modelo em relação à informação disponível no modelo da média histórica (*benchmark*). Neste artigo, calculamos o ganho de utilidade dos modelos de previsão usando o método introduzido por Campbell and Thompson (2008) e Rapach and Strauss (2010).

O método pressupõe um investidor avesso ao risco que tem uma função de utilidade de média variância e considera como alocar otimamente a riqueza total entre um um ativo arriscado e um ativo livre de risco no momento t com base na taxa livre de risco atual e na previsão do prêmio de risco, $\hat{r}_{t+1}$. Dessa forma, o peso atribuído ao ativo arriscado é calculado da seguinte forma: $\omega_t = \frac{1}{\gamma}\frac{\hat{r}_{t+1}}{\hat{\sigma}^2_{t+1}}$.

Onde γ é o parâmetro de aversão ao risco e $\hat{\sigma}^2_{t+1}$ é a variância estimada do prêmio de risco. Assim, o retorno realizado do portfólio no momento $t+1$ é $R_{t+1} = \omega_t r_{t+1} + (1-\omega_t) r^f_{t+1}$. Impomos a restrição $\omega_t \in (0, 1.5)$ para garantir que não haja venda a descoberta ou alavancagem.

Assim, durante o período fora da amostra, o investidor percebe um nível médio de utilidade de:

$$U = \hat{\mu} - \frac{1}{2}\gamma\hat{\sigma}^2 \tag{11}$$

onde $\hat{\mu} = \frac{1}{P}\sum_t R_t$, $\sigma^2_p = Var(R_t) = \frac{1}{P}\sum_t (R_t - \hat{\mu}_p)^2$, e P é a quantidade total de observações fora da amostra. O ganho de utilidade é a diferença entre a utilidade obtida usando um modelo baseado na previsão, $\hat{r}_{t+1}$, e o modelo de previsão baseado na média histórica(benchmark), $\overline{r}_{t+1}$. Para facilitar a interpretação, multiplicamos os ganhos de utilidade por 1200, o que nos dá a taxa de administração anual que um investidor estaria disposto a pagar para obter acesso à informação adicional da previsão baseada no modelo $\hat{r}_{t+1}$. Além disso, reportamos os resultados utilizando $\gamma = 3$ [11] .

[10] Para modelos aninhados Clark and West (2007) mostram que as estatística de DIEBOLD and MARIANO (1995) tem distribuição fora do padrão, uma vez que, seu teste poderia ser subdimensionado sob a hipótese nula e com baixo poder de previsão sob a hipótese alternativa.

[11] Conforme adotado em Lima and Meng (2017) e Godeiro et al. (2018)

3 Resultados Empíricos

Para se fazer a avaliação do prêmio de risco no mercado de ações brasileiro foi feita uma estimação dentro da amostra de 2008:01 a 2014:03 (75 observações) e fora da amostra de 2014:04 a 2021:01 com um total de $P = 79$ observações. Os dados foram estimados recursivamente ao longo do tempo e a divisão dos mesmos teve como objetivo ter uma amostragem fora da amostra que seja superior. Em um segundo momento se procedeu a análise fora da amostra, traçando os resultados das previsões condicionais de acordo com a estatística $R^2_{OS}(\%)$ e sua significância através dos p valores do procedimentos do *Erro Quadrado Médio de Previsão Ajustado* (MSPE) de Clark and West (2007), teste (CW) demonstrando que para valores positivos um dos modelos (Dados Textuais, Indicador Técnico ou Fatores de Risco) têm performance superior ao modelo de referência (*benchmark*), além do ganho de utilidade do investidor com preferência de média variância, $\Delta U\%$ (anual) associado a cada modelo de previsão. Por fim, os resultados das previsões dos preditores individuais foram expostos nas tabelas de 1 a 5. [12]

Na tabela 1 tem-se os resultados referente ao modelo baseado em Dados Textuais (não estruturado) onde de (02) observamos a frequência de palavras negativas das notícias financeiras ao longo do tempo dos jornais "O GLOBO ", "Estadão"e "Valor Econômico". O objetivo foi observar se as técnicas de regularização Ridge, LASSO e *Elastic Net* geram um índice de sentimento SI^{neg}_{TV} que seja superior ao modelo de *benchmark* (Média Histórica) na estimativa do prêmio de risco para o mercado acionário brasileiro. Assim, foi observado que os modelos baseados em sentimentos são estatisticamente inferiores ao de referência quando observado a estatística $R^2_{OS}(\%)$ sendo o LASSO de -3.47 %, Ridge -0.01 % e o *Elastic Net* -2.92 %. Embora o *Elastic Net* seja aninhado aos dois anteriores, o seu resultado não obteve uma performance significativa. Essa análise corrobora o trabalho de Godeiro et al. (2018) ao demonstrar que o *Elastic Net* tem uma performance superior ao LASSO e Ridge uma vez que, os dois últimos são aninhados ao primeiro. Ainda da tabela 1 notamos que o MSPE do teste Clark and West (2007) demonstra que de fato ambos os dicionários não apresentam superioridade em relação ao *benchmark*, sendo seus resultados acima de 80% para o LASSO e *Elastic Net*.

Dando continuidade as classes dos demais modelos, as tabelas de 2 a 4 tem como escopo os dados referentes ao modelo de Indicadores Técnicos conforme suas classificações (Média Móvel, Momento e Volume). Inicialmente na tabela 2 temos que as melhores combinações que obteve um $R^2_{OS}(\%)$ positivo e estatisticamente significativo maior que zero, foi a MA_03_09 com um desempenho superior ao da média histórica em 1.19 %. Vale ainda destacar que a combinação MA_03_12 foi exatamente igual a média histórica, sendo o seu valor igual a 0. Assim, para esta combinação temos que a média dos retornos utilizados contém informações para inferir o retorno esperado, que neste caso seria o *benchmark*. Desta forma, o mercado estaria observando as médias do prêmio de risco acionário anterior para então ter expectativa do seu futuro desempenho. Posteriormemte, na tabela 3 observamos que o indicador Momento (MOM) não obteve nenhuma combinação estatísticamente positiva, sendo a média histórica superior aos mesmos. Por fim,verificamos na tabela 4 este mesmo resultado para o Indicador de Volume(OBV) .

Na tabela 5 atribuimos os resultados das previsões baseado no modelo de Cinco Fatores de Risco Fama and French (2015). Observamos que apenas *Winners Minus Losers Factor* (WML) obteve uma melhor previsão, sendo um $R^2_{OS}(\%)$ positivo em 0.39 % .

Finalmente, são apresentados os resultados dos gráficos que têm como objetivo mostrar a evolução ao longo do tempo do retorno de portfólio [13] de um investidor de média variância, cuja alocação foi

[12] Os resultados encontrados para cada preditor são apresentados em figuras e tabelas.

[13] O Portfólio constitui um conjunto de ativos de um investidor (pessoa física ou jurídica), cujo objetivo é reduzir o risco por

estabelecida a partir das previsões geradas pelos modelos de Dados Textuais, Indicadores Técnicos e os Cincos Fatores de Fama and French (2015) comparados ao modelo da Média Histórica. Utilizamos o mesmo peso adotado em Lima and Meng (2017) para calcular o ganho de utilidade, conforme o parâmetro de aversão ao risco $\gamma = 3$. Por exemplo, com base nas previsões fora da amostra $(t + 1)$ o investidor aloca um peso em ativos com risco r_{t+1} e em ativo sem risco, r^f_{t+1} . Posteriormente calcula o retorno do portfólio, R_{t+1} conforme definido em (2.4).

Assim, quando a tendência dos gráficos mostrar-se negativo, o retorno de portfólio baseado na média histórica naquele determinado período proporciona um resultado maior que o modelo individual (dados textuais, técnicos ou cinco fatores). Caso seja positivo, a resposta de cada modelo é superior ao modelo de referência (*benchmark*). A sua análise complementa os resultados do ganho de utilidade apresentadas nas tabelas de 1 a 5. Por exemplo, um investidor que se baseou na combinação MA_02_12 teve em média um retorno de 1.27% maior que o retorno do portfólio baseado na média histórica, proporcionando um significativo ganho econômico / financeiro para os investidores.

Por fim, destacamos que as combinações MA_03_09 foram as que apresentaram maior ganho de utilidade, superando a média histórica em 5.29% ao ano. Graficamente observamos que o seu comportamento foram positivos para quase todo o período fora da amostra. Além dele encontramos o Ridge que apresentou apenas um pico negativo, mas no decorrer dos períodos foram positivos, obtendo um retorno no ganho de utilidade em 0.32% . Porém, quando observado os ganhos de utilidades referente ao modelo de Momento (MOM), o consumidor não maximiza sua utilidade, sendo seus valores de -1.54% para MOM_01_09 e -2.30% para MOM_01_12. Graficamente, o retorno de portfólio destes preditores têm maiores picos para valores negativos não havendo, portanto, diverisficação entre ativo sem risco e ativo arriscado.

4 Considerações Finais

O presente artigo teve como objetivo investigar se os modelos baseadas em notícias financeiras superam as amostras de previsões fundamentadas na Média Histórica (*benchmark*), Indicadores Técnicos e Fatores de Risco de Mercado na previsão do prêmio de risco acionário bracileiro.

Para alcançar o objetivo proposto foram utilizadas notícias dos jornais "O Globo", "Estadão" e "Valor Econômico" e o aprendizado de máquina supervisionado via Ridge, LASSO e *Elastic Net* para selecionar as palavras mais preditivas no decorrer do tempo conforme o teor das notícias (positivas / negativas) para os dados fora da amostra. Em seguida, estimou-se o índice de sentimento verificando a sua influência na análise do prêmio de risco acionário brasileiro quando comparado ao *benchmark* do mercado. Nesse sentido, foi observado que esses modelos não superam o modelo da média histórica, sendo estatíticamente insignificantes para todos os períodos *out-of-sample*. Este resultado é contrário a Godeiro et al. (2018) que ao acrescentarem o modelo de dados textuais observaram que o seu resultado potencializa alguns modelos de previsões encontrado na literatura e superando o valor de referência (*benchmark*).

Posteriormente, fez-se a análise dos Indicadores Técnicos conforme sua classificação, identificando que alguns apresentaram ganhos econômicos e financeiros superando a média histórica, como: (MA_01_12), (MA_02_09), (MA_02_12), (MA_03_09) e (MA_03_12). Por outro lado, quando observado o modelo de Cinco Fatores de Risco Fama and French (2015) apenas um tem um valor estatísticamente positivo e acima de zero, *Winners Minus Losers* (WML) com R^2_{OOS} igual a 0.39% superando o modelo de referência

meio da diversificação entre ativos arriscado e não arriscado.

(*benchmark*).

Já quando analisado os ganhos de utilidade do agente tomador de decisão destacamos que os mesmos obtiveram ganhos significativos na (MA_03_09) em que foi observado um retorno de 5.29 % maior que o *benchmark*. Assim, esse preditor consegue repassar mensagem ao investidor que um dado mês ou meses, a melhor decisão é ficar fora do ativo de risco, maximizando sua utilidade. Além dele, temos que o *Market Factor* e o *Small Minus Big* (SMB Factor) conseguem passar mensagem ao investidor, com 3.15% para o primeiro e 4.20% para o segundo. Também foi observado que, na classificação de palavras ao longo do tempo, apenas Ridge apresentou um bom resultado, com 0.32 % ganho de utilidade.

Desse modo, o presente artigo revela que os modelos baseados em notícias financeiras a partir de palavras variante no tempo não apresentaram ganhos estatíticos e econômicos, bem como, algumas classificações dos Indicadores Técnicos: Momento (MOM) e Volume (OBV). No entanto, observa-se que o preditor (WML Factor) e cinco das seis combinações da Média Móvel (MA) foram os melhores modelos de previsão para o prêmio de risco acionário brasileiro.

Referências

Bai, J. and Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2):304–317.

Campbell, J. Y. and Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, 21(4):1509–1531.

Clark, T. E. and West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of econometrics*, 138(1):291–311.

Cutler, D. M., Poterba, J. M., and Summers, L. H. (1988). What moves stock prices? Technical report, National Bureau of Economic Research.

DIEBOLD, F. and MARIANO, R. (1995). Comparing predictive accuracy. journal of business and economics statistics, v. 13.

Elliott, G. and Timmermann, A. (2013). *Handbook of economic forecasting*. Elsevier.

Engelberg, J. (2008). Costly information processing: Evidence from earnings announcements.

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

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22.

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance*, 68(3):1267–1300.

Godeiro, L. L. et al. (2018). Ensaios sobre modelos de previsão econômica.

Gonçalves, L. I. (2018). O uso de indicadores técnicos como suporte à tomada de decisões no mercado financeiro.

Hansen, S., McMahon, M., and Prat, A. (2017). Transparency and deliberation within the fomc: a computational linguistics approach. *The Quarterly Journal of Economics*, 133(2):801–870.

Li, J., Tsiakas, I., and Wang, W. (2015). Predicting exchange rates out of sample: Can economic fundamentals beat the random walk? *Journal of Financial Econometrics*, 13(2):293–341.

Li, Y., Ng, D. T., and Swaminathan, B. (2013). Predicting market returns using aggregate implied cost of capital. *Journal of Financial Economics*, 110(2):419–436.

Lima, L. R. and Meng, F. (2017). Out-of-sample return predictability: A quantile combination approach. *Journal of Applied Econometrics*, 32(4):877–895.

Lima, L. R., Meng, F., and Godeiro, L. (2019). Quantile forecasting with mixed-frequency data. *International Journal of Forecasting*.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

Neely, C. J., Rapach, D. E., Tu, J., and Zhou, G. (2014). Forecasting the equity risk premium: the role of technical indicators. *Management science*, 60(7):1772–1791.

Oliveira, A. D. C. M. d. (2016). Identificando emoções em manchetes de notícias escritas em português do brasil utilizando naïve bayes.

Rapach, D. and Zhou, G. (2013a). Forecasting stock returns. In *Handbook of Economic Forecasting*, volume 2, chapter 6, pages 328–383. Elsevier.

Rapach, D. and Zhou, G. (2013b). Forecasting stock returns. In *Handbook of economic forecasting*, volume 2, pages 328–383. Elsevier.

Rapach, D. E. and Strauss, J. K. (2010). Bagging or combining (or both)? an analysis based on forecasting us employment growth. *Econometric Reviews*, 29(5-6):511–533.

Silva, M. D. d. O. P. d. (2018). O efeito do sentimento das notícias sobre o comportamento dos preços no mercado acionário brasileiro.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, 62(3):1139–1168.

Vasconcelos, B. F. B. d. (2017). Poder preditivo de métodos de machine learning com processos de seleção de variáveis: uma aplicação às projeções de produto de países.

Welch, I. and Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4):1455–1508.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Tabelas e Figuras

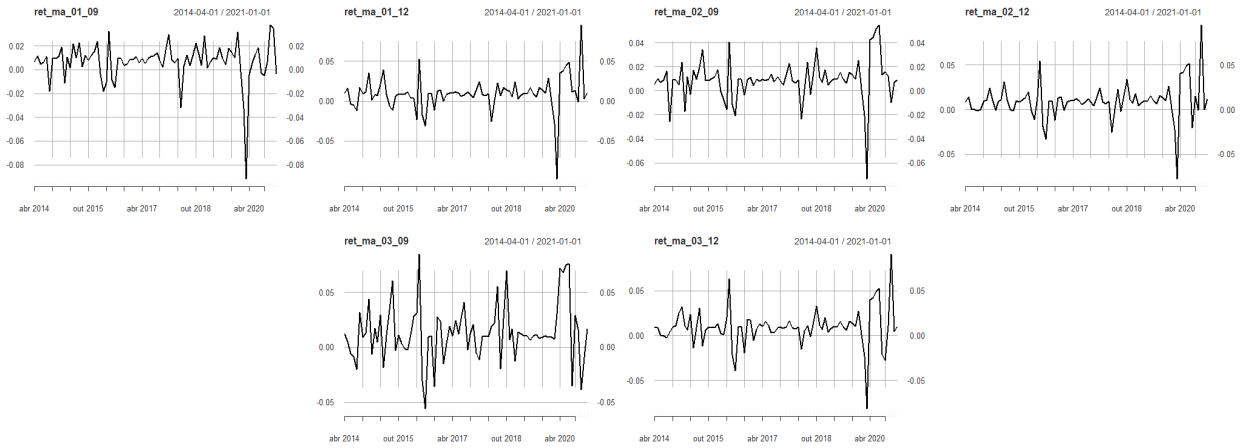

Figura 1: Gráficos dos Retornos de Portfólio do Indicador Técnico -Média Móvel(MA)

O gráfico mostra o comportamento dos retornos de Portfólio com base na série temporal fora da amostra, 2014:04 a 2021:01 do Indicador Técnico baseado na Média Móvel (MA), comparado ao modelo de referência do mercado (*benchmark*).

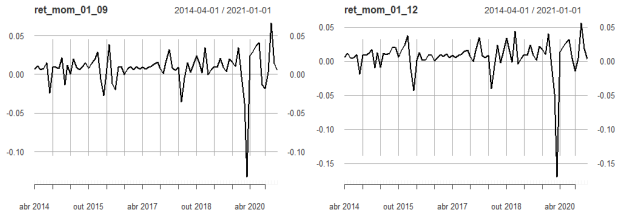

Figura 2: Gráficos dos Retornos de Portfólio do Indicador Técnico - Momento (MOM)

O gráfico mostra o comportamento dos retornos de Portfólio com base na série temporal fora da amostra, 2014:04 a 2021:01 do Indicador Técnico baseado no Momento (MOM), comparado ao modelo de referência do mercado (*benchmark*).

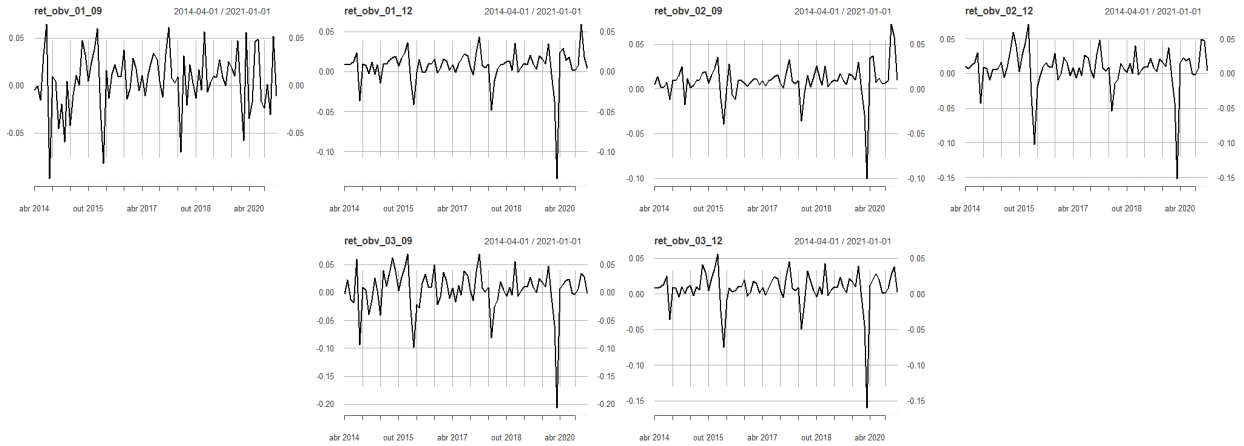

Figura 3: Gráficos dos Retornos de Portfólio do Indicador Técnico - Volume (OBV)

O gráfico mostra o comportamento dos retornos de Portfólio com base na série temporal fora da amostra, 2014:04 a 2021:01 do Indicador Técnico baseado no Volume (OBV), comparado ao modelo de referência do mercado (*benchmark*).

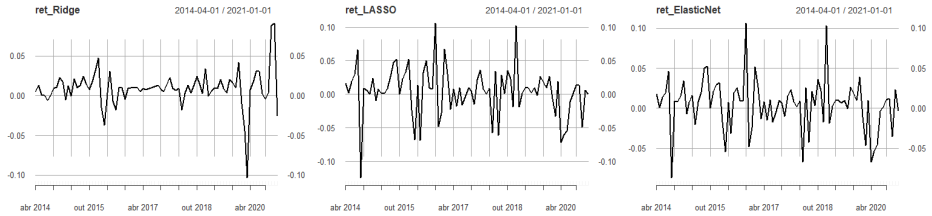

Figura 4: Gráficos dos Retornos de Portfólio dos Modelos Basedados em Dados Textuais

O gráfico mostra o comportamento dos retornos de Portfólio com base na série temporal fora da amostra, 2014:04 a 2021:01 do Modelo baseados em Dados Textuais, comparado ao modelo de referência do mercado (*benchmark*).

Figura 5: Gráficos dos Retornos de Portfólio dos Cincos Fatores de Fama e French

O gráfico mostra o comportamento dos retornos de Portfólio com base na série temporal fora da amostra, 2014:04 a 2021:01 do Modelo de Cinco Fatores de Fama and French (2015), comparado ao modelo de referência do mercado (*benchmark*).

Tabela 1: Resultado das Previsões Fora da Amostra para o Prêmio de Risco Acionário

Painel D: Modelo Baseado em Dados Textuais			
Modelo	$R^2_{OOS}\%$	CW	$\Delta U\%$
Ridge	-0.01	0.45	0.32
LASSO	-3.47	0.80	-6.44
Elastic Net	-2.92	0.83	-5.52

A tabela 1 retrata dos valores $R^2_{OS}\,(\%)$ fora da amostra em termos percentuais (%) e sua significância através dos *p* valores do teste de Clark and West (2007) (CW) além do ganho de utilidade do investidor com preferência média variância em termos de $\Delta\%$ .Se $\mathrm{R}^2_{OS} > 0$ a previsão condicional do Modelo Baseado em Dados Textuais supera a média histórica do mercado (*benchmark*). O ganho de utilidade anual é a taxa de administração que o investidor estaria disposto a pagar para ter acesso às informações adicionais.

Tabela 2: Resultados das previsões fora da amostra do Prêmio de Risco Acionário

Painel A: Indicadores Técnicos- Média Móvel (MA)			
Modelo	$R^2_{OOS}\%$	CW	$\Delta U\%$
MA_01_09	-1.08	0.99	-1.82
MA_01_12	0.05	0.40	0.64
MA_02_09	0	0.45	0.86
MA_02_12	0.25	0.32	1.27
MA_03_09	1.19	0.15	5.29
MA_03_12	0.02	0.41	0.73

A tabela 2 retrata dos valores $R^2_{OS}\,(\%)$ fora da amostra em termos percentuais (%) e sua significância através dos *p* valores do teste de Clark and West (2007)(CW) além do ganho de utilidade do investidor com preferência média variância em termos de $\Delta\%$ .Se $\mathrm{R}^2_{OS} > 0$ a previsão condicional do Indicador Técnico basedo na Média Móvel (MA) supera a média histórica do mercado (*benchmark*). O ganho de utilidade anual é a taxa de administração que o investidor estaria disposto a pagar para ter acesso às informações adicionais.

Tabela 3: Resultado das Previsões Fora da Amostra para o Prêmio de Risco Acionário

Painel B: Indicadores Técnicos- Momento (MOM)			
Modelo	$R^2_{OOS}\%$	CW	$\Delta U\%$
MOM_01_09	-0.59	0.88	-1.54
MOM_01_12	-0.78	0.95	-2.30

A tabela 3 retrata dos valores R^2_{OS} (%) fora da amostra em termos percentuais (%) e sua significância através dos *p* valores do teste de Clark and West (2007) (CW) além do ganho de utilidade do investidor com preferência média variância em termos de $\Delta\%$ .Se $R^2_{OS} > 0$ a previsão condicional do Indicador Técnico Baseado no Momento (MOM) supera a média histórica do mercado (*benchmark*). O ganho de utilidade anual é a taxa de administração que o investidor estaria disposto a pagar para ter acesso às informações adicionais.

Tabela 4: Resultado das Previsões Fora da Amostra para o Prêmio de Risco Acionário

Painel C: Indicadores Técnicos- Volume (OBV)			
Modelo	$R^2_{OOS}\%$	CW	$\Delta U\%$
OBV_01_09	-2.24	0.82	-5.14
OBV_01_12	-0.75	0.93	-2.13
OBV_02_09	-0.16	0.59	0.01
OBV_02_12	-0.81	0.73	-2.58
OBV_03_09	-2.66	0.92	-7.41
OBV_03_12	-0.87	0.86	-2.63

A tabela 4 retrata dos valores R^2_{OS} (%) fora da amostra em termos percentuais (%) e sua significância através dos *p* valores do teste de Clark and West (2007)(CW) além do ganho de utilidade do investidor com preferência média variância em termos de $\Delta\%$ .Se $R^2_{OS} > 0$ a previsão condicional do Indicador Técnico baseado no Volume (OBV)supera a média histórica do mercado (*benchmark*). O ganho de utilidade anual é a taxa de administração que o investidor estaria disposto a pagar para ter acesso às informações adicionais.

Tabela 5: Resultado das Previsões Fora da Amostra para o Prêmio de Risco Acionário

Painel E: Modelo Baseado nos Cincos Fatores de Fama e French			
Modelo	$R^2_{OOS}\%$	CW	$\Delta U\%$
Market Factor	-1.35	0.38	3.15
SMB Factor	-2.48	0.38	4.20
HML Factor	-1.28	0.86	-3.07
WML Factor	0.39	0.16	2.81
IML Factor	-1.15	0.72	-1.37

A tabela 5 retrata dos valores R^2_{OS} (%) fora da amostra em termos percentuais (%) e sua significância através dos *p* valores do teste de Clark and West (2007) (CW) além do ganho de utilidade do investidor com preferência média variância em termos de $\Delta\%$ .Se $R^2_{OS} > 0$ a previsão condicional do Modelo de Fama and French (2015) supera a média histórica do mercado (*benchmark*). O ganho de utilidade anual é a taxa de administração que o investidor estaria disposto a pagar para ter acesso às informações adicionais.

Apêndice 1: Dicionário das Palavras mais Preditivas

Dicionário Negativo

aberto; abril; abrir, abriram; acentuada; acesita; acho; ação; acrescentou; administração; adr; afetada; afetado; endividado; agosto; alteração; estagnado; estragar; excluída; evitar; escandalosa; erro; esquecer; escassez; alan; alberto; alcoa; alimento; alteração; escandaloso; escândalos; operassem; parada; reclama; reclamam; parada; recuavam; violência; zeraram; reduzissem; prejudicassem; baixa; baixo; renunciaria;

Apêndice 2: Combinações dos Indicadores Técnicos

Tabela 6: Combinações de Pares dos Indicadores Técnicos

Preditores	**Descrição dos Pares conforme o Indicador Técnico**
(MA_01_09)	Média Móvel Curta do Primeiro Período com a Média Móvel Longa do Nono Período
(MA_01_12)	Média Móvel Curta do Primeiro Período com a Média Móvel Longa do Décimo Segundo Período
(MA_02_09)	Média Móvel Curta do Segundo Período com a Média Móvel Longa do Nono Período
(MA_02_12)	Média Móvel Curta do Segundo Período com a Média Móvel Longa do Décimo Segundo Período
(MA_03_09)	Média Móvel Curta do Terceiro Período com e Média Móvel Longa do Nono Período
(MA_03_12)	Média Móvel Curta do Terceiro período com a Média Móvel Longa do Décimo Segundo Período
(MOM_01_09)	Retorno dos Níveis de Preços no Primeiro Período com Retorno dos Níveis de Preços do Nono Período
(MOM_01_12)	Retornos dos Níveis de Preços no Primeiro Período com os Retornos dos Níveis do Décimo Segundo período
(OBV_01_09)	Saldo de Volume do Primeiro Período com o Saldo de volume do Nono Período
(OBV_01_12)	Saldo de Volume do Primeiro Período com o Saldo de Volume do Décimo Segundo Período
(OBV_02_09)	Saldo de Volume do Segundo Período com o Saldo de Volume do Nono Período
(OBV_02_12)	Saldo de Volume do Segundo Período com o Saldo de Volume do Décimo Segundo Período
(OBV_03_09)	Saldo de Volume do Terceiro Período com o Saldo de Volume do Nono Período
(OBV_03_12)	Saldo de Volume do Terceiro Período com o Saldo de Volume do Décimo Segundo Período

A tabela 6 retrata das combinações de pares para cada Indicador Técnico conforme seus períodos, formando seis pares para a Média Móvel (MA); dois pares o Momento (MOM) e seis pares para o Volume (OBV).