Previsão do Índice Ibovespa Usando Análise Técnica e Modelagem Computacional

Flavio Barboza − flmbarboza@ufu.br
Universidade Federal de Uberlândia

Mariana Tognetti − marianatognetti@gmail.com
Universidade Federal de Uberlândia

Resumo

Este artigo tem como objetivo analisar a eficiência das previsões de tendência do índice de ações da B3 (IBOVESPA) em três diferentes horizontes de tempo, utilizando métodos computacionais (*Random Forest* e *Support Vector Machines, SVM*) como também a ferramenta estatística usual, a regressão logística (Logit). Com isso, é feito um comparativo de desempenho entre as previsões e o fato real. A implementação é feita em dados diários, semanais e mensais, os quais representam a ideia de curto, médio e longo prazos, respectivamente, com valores do IBOVESPA coletados num período de quase 20 anos (janeiro de 2000 e julho de 2018), tendo como objetivo prever a alta ou a baixa do IBOVESPA no período seguinte. A análise técnica serviu de fundamentação para a construção das variáveis explicativas, fornecendo indicadores com foco em determinantes de tendência. Os modelos de previsão foram então analisados em termos de acurácia média, erros do tipo I e II, e a curva ROC. Os resultados mostram os modelos baixo desempenho no curto prazo, com alguma melhora no médio prazo e, no longo praz, a capacidade preditiva chega a 68% com o modelo de SVM, mostrando melhor desempenho que logit (64%) e RF (61%). Tais resultados abrem discussão para a funcionalidade da análise técnica, uma vez que espera-se melhor performance no curto prazo, enquanto os achados mostram o contrário.
Palavras-chave: *Support Vector Machines*; *Random Forest*; Regressão Logística; Análise Técnica; Previsão do IBOVESPA.

Abstract

The focus of this research is to analyze the efficiency of brazilian market index (IBOVESPA) trend forecast in three different time horizons, by using computational methods (Random Forest and Support Vector Machines, SVM) as well as the usual statistical method, logistic regression (Logit). Then, a performance comparison is made among the forecasts and the actual facts. The implementation is done in daily, weekly and monthly data, which represent the idea of short, medium and long term, respectively, with IBOVESPA values collected in a period of almost 20 years (January 2000 and July 2018). IBOVESPA in the next period. The technical analysis is used as a basis for the explanatory variables design, providing indicators focused on trend determinants. The prediction models were then analyzed in terms of accuracy, type I and II errors, and the ROC curve. The results show the low performance models in the short term, with some improvement in the medium term and, in the long term, the predictive capacity reaches 68% with the SVM model, showing better performance than logit (64%) and RF (61%). These results open the discussion for the technical analysis funcionality, since better performance is expected in the short term, whereas the findings show the opposite.
Keywords: Support Vector Machines; Random Forest; Logistic Regression; Technical analysis; IBOVESPA forecast.

1. Introdução

A mudança na dinâmica das negociações de ações e seus derivativos ao longo dos últimos anos exigiu a aplicação de métodos e modelos para auxiliar a tomada de decisão sobre as precificações dos ativos nas bolsas de valores pelo mundo (Dash & Dash, 2016).

Considerada a maior bolsa da América Latina e uma das cinco maiores entre os mercados emergentes (Al Nasser & Hajilee, 2016), a Brasil, Bolsa, Balcão, ou ainda, B3, utiliza o Índice Bovespa (IBOVESPA) para representar o comportamento do mercado brasileiro, posto que é definido pelo conjunto das ações mais negociadas, a partir de metodologia específica. Assim, é natural o interesse dos investidores em saber o que vai acontecer com o IBOVESPA, para que possam prever melhor os preços dos ativos negociados no Brasil.

Entretanto, a movimentação dos preços é algo complexo (Basak, Kar, Khaidem, Saha, & Dey, 2018) e desafiador (Kim, 2003). Diante de um montante considerável de dados e informações, métodos para tentar prever tendências e preços, como também analisar riscos surgem a todo momento, como por exemplo, os métodos computacionais (Dash & Dash, 2016), que buscam formas variadas de combinar dados para encontrar padrões e, dessa forma, antecipar cenários futuros; além das técnicas estatísticas, que são frequentemente aplicadas em diversos problemas de previsão não somente na área de finanças, mas em muitos outros tópicos (Barboza, Kimura, & Altman, 2017).

Muitos dos métodos disponíveis na literatura utilizam variáveis baseadas nos preços históricos ou mesmo usando indicadores técnicos, que são frutos de operações e tratamentos em torno dos preços, o que é conhecido por análise técnica. Fang, Jacobsen e Qin (2014) estudaram a previsibilidade dos mercados usando a análise técnica num período de 25 anos e observaram que 26 estratégias diferentes baseadas em médias móveis foram inferiores a performance do mercado em si, assim como Širůček e Šíma (2016) ao analisar o índice S&P500 com indicadores otimizados identificaram algumas possibilidades, porém existem situações que os indicadores não provém retornos suficientes para superar o índice. Em contrapartida, estimativas promissoras da tendência do mercado foram encontradas por Kumar e Thenmozhi (2006) e Patel, Shah, Thakkar e Kotecha (2015), quando empregaram modelagem computacional com o uso de indicadores técnicos. Neste segundo caso, o percentual de acerto chega a 90%.

As principais fontes de dados da análise técnica são os preços, formando gráficos nos quais os analistas tentam visualizar os desejados padrões que, em tese, representariam tendências de alta e baixa dos preços (Fang et al., 2014). Por outro lado, existe a análise fundamentalista,

que, de acordo com Širůček e Šíma (2016), descreve o valor intrínseco da empresa, estando assim subjetivamente ligado ao preço da ação.

Neste trabalho, investiga-se o desempenho da modelos computacionais para prever a alta ou a baixa do IBOVESPA usando dez diferentes indicadores procedentes da análise técnica e também empregados por Patel et al. (2015): médias móveis simples e ponderada (SMA e WMA), momento, osciladores estocásticos %K e %D, *Moving Average Convergence Divergence* (MACD), *Relative Strength Index* (RSI), Oscilador de Larry Williams (%R), *Accumulation/Distribution* (A/D) e *Commodity Channel Index* (CCI), os quais foram também calculados e padronizados conforme Patel et al. (2015). A opção desta sobre a análise fundamentalista deve-se ao fato da análise técnica prever, além do momento de compra e venda, a direção que os preços dos ativos irão seguir no futuro (Fang et al., 2014).

Como a capacidade preditiva da Análise Técnica no curto prazo (um dia) é um assunto ainda em constante debate, como também o uso de técnicas computacionais (Basak et al., 2016; Kim, 2003). Assim, a fim de determinar o quão eficientes são os resultados da previsão, a base de dados completa foi estratificada considerando três horizontes de tempo, denominados de curto prazo, para previsões do dia seguinte; médio prazo, usando base de dados semanais; e, longo prazo, para uma base de dados mensais. Em suma, a pesquisa tenta fornecer informações a respeito da alta/baixa do IBOVESPA para o dia D+1 (curto prazo) a partir de dados conhecidos até o fechamento do dia D. A ideia é análoga ao considerar o horizonte de uma semana e um mês, referindo-se a médio e longo prazos, respectivamente.

Após o processo de tratamento dos dados, dividiu-se a base de dados em duas amostras, sendo a primeira para a criar o modelo e a segunda para testar a qualidade deles. Quatro modelos foram analisados: *Support Vector Machines* (SVM) com dois *kernels* diferentes, *Random Forest* e a regressão logística, por meio do software R.

Como resultados, constata-se uma melhor capacidade preditiva dos modelos no longo prazo, mas ainda distante dos resultados de Patel et al. (2015). As previsões para o índice semanal fica em segundo e a performance de curto prazo sendo o pior caso, com acurácia relativamente baixa (abaixo de 50% em todos os modelos). No que se refere a modelagem, SVM apresentou o melhor desempenho, alcançando 68% de acurácia na previsão do IBOVESPA mensal. *Random Forest*, neste caso, é superado pela regressão logística, o que contrasta com estudos anteriores, como o próprio trabalho de Patel et al. (2015). Como os indicadores foram construídos a partir da literatura internacional e o emprego de dados de outros países, é provável que isso explique esse desencontro com resultados anteriores.

O estudo está dividido em 5 seções: esta que é a introdução; a segunda que é a revisão da literatura; a terceira seção apresenta a metodologia; a análise dos resultados é feita na seção 4 e, por fim, são feitas as conclusões e considerações na seção 5.

2. Revisão da Literatura

Para o embase deste artigo, são apresentados trabalhos que envolvem análise técnica e também pesquisas que envolvam modelagem computacional.

Para o auxílio na tomada de decisão sobre a compra ou venda dos ativos, uma ferramenta utilizada é a análise técnica (Širůček & Šíma, 2016). De acordo com Fang et al. (2014), ela se baseia nos dados históricos do ativos como preços de abertura, fechamento, máximo e mínimo, além do volume dos ativos negociados.

Ao longo dos anos, essa técnica apresentou resultados positivos tanto para os profissionais da área como pesquisadores (Dash & Dash, 2016; Basak, 2016; Kumar & Thenmozhi, 2006). A ideia da análise técnica é entender o comportamento dos investidores que é refletido no comportamento dos preços. Para expor essa ideia a análise técnica utiliza diversos indicadores, tais como as médias móveis, momento, osciladores estocásticos, MACD (*Moving Average Convergence Divergence*), RSI (*Relative Strength Index*), A/D (*Accumulation/Distribution*) e CCI (*Commodity Channel Index*), PRC (*Price Rate of Change*), OBV (*On Balance Volume*), sendo todos calculados a partir dos dados históricos para indicar a tendência do mercado. Os indicadores são também visualizados em gráficos, evidenciando, quando analisado junto às linhas de suporte e resistência, mudanças de tendência quando as formações se inverterem, ou manutenção da tendência inicial, quando não houver alteração brusca no gráfico. Assim a análise técnica cumpre o seu objetivo de indicar tendências do mercado (Basak et al., 2016).

Porém, os analistas possuem uma outra opção para interpretar o mercado, a chamada análise fundamentalista. A grande diferença da análise técnica sobre a análise fundamentalista é que a primeira determina a direção dos preços, se haverá subida ou descida, e antecipar prováveis inversões de valores dos investidores, enquanto a análise fundamentalista se sustenta em indicadores: (i) financeiros relativos ao ativo, tais como lucratividade, liquidez, endividamento, entre outros (Wang, Li, Qin, & Ge, 2011), (ii) do mercado para fazer comparativos (KIM, 2003) e econômicos, no intuito de realizar projeções.

No tocante aos métodos utilizados para criar modelos de previsão do mercado, inúmeras alternativas são discutidas na literatura. As máquinas de suporte vetorial, ou SVM,

apresentaram resultados superiores a aqueles obtidos por modelos mais tradicionais, como as redes neurais, e também pelo seu desempenho para prever a direção dos preços quando utilizado dados temporais. Um trabalho em que a SVM destaca-se na área de finanças é o realizado por Kim (2003) no qual ele aplicou o modelo para prever a direção do Índice da Bolsa da Coreia (KOSPI) os resultados reforçam o uso desse modelo aplicado à base de dados temporais com performance superior ao modelo de redes neurais.

Outro método computacional frequente na literatura é o *Random Forest*, um algoritmo desenvolvido por Breiman (2001) que consiste na mineração dos dados de forma independente e aleatória em subgrupos, as árvores, indicando o quão favorável àquele dado é para determinado nó (as variáveis). Esse processo sofre centenas de repetições, sempre selecionando o nó e os dados de modo aleatório e independente, o que colabora para a minimização do chamado sobreajuste do modelo (Basak et al., 2016).

Métodos estatísticos são largamente empregado em finanças, mas para modelos preditivos o mais conhecido é a regressão logística (LOGIT). Sua forma permite melhor entendimento em comparação com outros métodos estatísticos, possui menos requisitos da distribuição dos dados, além de prover seus resultados em termos de probabilidade de ocorrência do evento.

Patel et al. (2015) comparou o poder preditivo de SVM e *Random Forest* com regressão logística usando indicadores técnicos em dados da Índia (da bolsa de Bombaim). Seus resultados mostraram que SVM e *Random Forest* são melhores, sendo o segundo o melhor deles. Kumar e Thenmozhi (2006) testaram também dados da Índia usando estes três métodos e dois outros (análise discriminante e redes neurais), e novamente constataram que SVM foi melhor na previsão de do índice S&P CNX NIFTY.

Fan e Palaniswani (2001) aplicaram SVM, *Randon Forest* e regressão logística no índice da bolsa de valores australiana. As conclusões reunidas por esses autores foram favoráveis ao uso dos modelos, em especial ao SVM que foi descrito como alternativa para a utilização de dados fundamentalistas.

3. Metodologia

Para a análise do IBOVESPA foram coletados dados de preços (máximos, mínimos, aberturas e fechamentos) em periodicidades diária, semanal e mensal, utilizando a base de dados da bolsa de valores brasileira, B3 (2018), que também foi usada por Wang et al. (2011). Em termos gerais, a amostra inicial abrange 4372 dias, 968 semanas e 223 meses de observações entre os dias 01 de Janeiro de 2000 a 15 de Julho de 2018, tendo assim um total de quase 18,5

anos de valores do índice. A Figura 1 mostra o histórico do índice no período analisado no qual é possível perceber os efeitos da crise financeira de 2008.

FIGURA 1 - Histórico do IBOVESPA entre os anos 2000 e 2018. Ao centro, destaca-se o movimento provocado pela crise financeira mundial. Fonte: elaborada pelos autores usando cotações históricas coletadas no site da B3 (2018).

Assim, toda metodologia baseia-se no estudo de previsão do altas e baixas do índice para o dia, a semana e o mês seguintes, dividindo-se a análise em três cenários, que foram denominados curto, médio e longo prazos, respectivamente.

Logo após a coleta, esses dados foram empregados nos cálculos de 10 indicadores da análise técnica que funcionarão como variáveis explicativas. São eles: SMA, WMA, momento, estocástico %K, estocástico %D, MACD, RSI, Larry Williams %R, Acumulação e Distribuição (A/D) e Índice de Canal de Commodities (CCI). As fórmulas de cada um estão presentes na Tabela 1, a seguir. Afim de elucidar o panorama geral destes indicadores, apresenta-se nos resultados a estatística descritiva dos mesmos.

Tanto a variável dependente (retorno do período seguinte) como os indicadores técnicos foram padronizados, conforme metodologia de Patel et al. (2015). Para os retornos positivos (em relação ao período anterior), adotou-se o valor 1, o contrário foi classificado como -1. Os indicadores MACD, %K, %D, %R, A/D, são classificados como 1 quando os valores do período são maiores do que aqueles do período anterior, caso contrário, atribuiu-se o valor -1. Para as médias móveis, é atribuiu-se 1 nos casos em que o preço de fechamento foi maior que o valor da média para o mesmo período, caso contrário, os valores são -1. Já o Momento é representado por 1 em períodos que apresentam valores positivos para esse indicador, caso contrário, é categorizado como -1. O CCI foi substituído por 1 quando seu valor estivesse abaixo de -200; caso o valor seja maior que -200, atribuía-se -1, e se ele se encontrasse entre -

200 e 200, era necessário observar o valor do período anterior. Desse modo, se o valor de CCI do período fosse maior que o anterior, recebia a classificação 1, caso contrário, era rotulado por -1. E no caso do RSI, se o valor fosse superior a 70, era classificado como -1, caso esteja abaixo de 30, era rotulado de 1. Se o valor de RSI estivesse entre 30 e 70, também era analisado o período anterior. Assim, caso seja fosse maior era classificado como 1, senão -1.

TABELA 1 - Descrição dos indicadores técnicos utilizados como variáveis explicativas na modelagem. C_t representa o preço de fechamento do dia t, LL_{t-9} é o menor mínimo dos preços observados os últimos 10 períodos, HH_{t-9} é o maior máximo dos últimos 10 períodos, UP e DW são dados pela médias móves das altas e das baixas dos últimos 10 períodos, respectivamente, DIF fornece a diferença entre as médias móveis exponenciais de 12 e 26 períodos, H_t é a máxima do dia t, L_t é a mínima do dia t, TP é a média aritmética entre H_t, C_t e L_t, e $D_t = (\sum_{i=1}^{10} |TP_{t-i+1} - SMA(TP, 10\ períodos)_t|)/10$.

Indicador	Fórmula
SMA (C, 10 períodos)	$C_{t-9} + C_{t-8} + \ldots + C_t/10$
WMA (C, 10 períodos)	$C_{t-9} + 2C_{t-8} + \ldots + 10C_t/(10 + 9 + \ldots + 1)$
Momento	$C_t - C_{t-9}$
%K	$100\ .(C_t - LL_{t-9})/(HH_{t-9} - LL_{t-9})$
%D	SMA (%K, 10 períodos)
Williams' %R	$1 - \%K$
RSI	$100 - 100/(1 + RS)$, tal que $RS = UP/DW$
MACD	$MACD_{t-1} + (2/11).(DIF - MACD_{t-1})$
A/D	$100.(H_t - C_{t-1})/(H_t - L_t)$
CCI	$\frac{TP_t - SMA(TP, 10\ períodos)}{0{,}015.D_t}$

Fonte: Patel et al. (2015) com adaptações.

No segundo momento, dividiu-se a base de dados em dois grupos, originando as amostras de treinamento (para a criação dos modelos) e teste (que verifica o poder preditivo dos modelos, isto é, se realmente são capazes de prever os retornos de períodos futuros), as quais possuem 67% e 33% da amostra final (elaborada após o computo dos indicadores técnicos), respectivamente. Ainda, a mostra teste é ocupada pelos dados mais recentes, assim como feito por Barboza et al. (2017), Fang et al. (2014) e Kumar e Thenmozhi (2006), e similar a Fan e Palaniswami (2001).

Com isso, foram criados modelos baseados nos métodos SVM (um com base linear e outro de base radial), *Random Forest* e Logit. Adotou-se a modelagem com parâmetros pré-definidos pelo programa R, por meio dos pacotes e1071, randomForest e gbm. A escolha destes métodos foi baseada nos bons resultados encontrados na literatura para prever preços quando estes métodos foram usados como base da criação dos modelos.

No caso do SVM, a proposta é criar um hiperplano que tem otimiza a distância entre as classificações de categorias diferentes, permitindo assim encontrar um divisor que separe as duas classes da melhor forma possível (Barboza et al., 2017). Para tanto utiliza-se uma função, denominada *kernel* que realiza um tratamento dos dados, por meio de uma mudança

de espaço n-dimensional. Para o *kernel* linear, a separação será sempre uma reta em qualquer dimensão, enquanto para *kernels* não lineares as formas são definidas pela função aplicada. A função de base radial (RBF) é uma delas e mais comumente aplicada. Em finanças, as aplicações são promissoras para esta técnica e tem alcançado um nível acentuado de acertos em modelos de previsão.

A técnica de *Random Forest* tem uma formatação razoavelmente simples. Usando do artefato de escolhas aletórias, cria várias árvores de decisão nas quais os nós são as variáveis preditoras escolhidas aleatoriamente e realizam a classificação para, no conjunto das árvores, definir a classificação de cada observação por maioria dos votos (cada árvore representa um voto). Conforme Basak et al. (2016), a técnica realiza internamente a seleção aleatória justamente para manter a estabilidade e aumentar o poder preditivo, e, paralelamente, contribui para reduzir a variância nas previsões e o sobreajuste (*overfitting*) do modelo (Wang et al., 2011).

A regressão logística é uma técnica estatística vastamente empregada em pesquisas na área de Finanças. Por ser adequada para modelagem com variáveis dependentes binárias (Kumar & Thenmozhi, 2006) e conhecida entre os pesquisadores da área, também foi aplicada neste estudo. De modo geral, o modelo é criado a partir de uma regressão baseada nas variáveis explicativas que fornece, a partir de uma função logística, uma probabilidade como resultado ou ainda, a previsão na forma percentual, ao invés de apenas classificar como 0 ou 1.

Para avaliar a qualidade destes modelos adotou-se medidas presentes em Basak et al. (2016), Patel et al. (2015) e Kumar e Thenmozhi (2006), tais como a matriz de confusão, o gráfico da curva ROC, e as medidas tradicionais de validação: erro do tipo I (percentual de erros de altas do IBOVESPA previstas como baixas em relação ao total de altas observadas), erro do tipo II (percentual de erros de baixas previstas como altas em relação ao total de baixas observadas) e a precisão (também conhecida por acurácia).

4. Análise Dos Resultados

Os resultados apresentados na sequência são oriundos dos processos descritos na seção anterior, nos quais foi utilizado dados diários para o curto prazo, dados semanais para o longo prazo e dados mensais para o longo prazo. As bases de dados foram construídas a partir de dados de preços entre 01 de Janeiro de 2000 e 15 de Julho de 2018. A Tabela 2 mostra a estatística descritiva dos indicadores técnicos adotados na pesquisa.

Observa-se uma série de curiosidades nestes dados. As médias móveis, tanto aritmética quanto ponderada, diminuem de valor a medida que o prazo aumenta, representando o comportamento esperado delas, suavidade intensificada em prazos mais longos, o que também é percebido pela variação total (entre os preços máximos e mínimos).

TABELA 2 - Estatística descritiva (máximos, mínimos, média e desvio-padrão) de cada um dos indicadores extraídos das amostras de curto, médio e longo prazos.

	MÁXIMO	MÍNIMO	MÉDIA	DESVIO PADRÃO
Curto Prazo (Dados diários)				
SMA	86293.91	8739.86	45469.99	19962.30
WMA	85809.88	8515.59	45004.84	19803.95
MOMENTO	7978.03	-13702.69	119.75	2181.40
%R	100	0	44.83	31.70
%K	100	0	55.17	31.70
%D	97.41	3.70	55.10	23.16
RSI	100	1.16	52.79	19.55
MACD	2102.46	-3804.44	95.03	796.27
A/D	250.04	-21.41	52.57	37.36
CCI	281.77	-308.64	10.29	105.05
Médio Prazo (Dados Semanais)				
SMA	85520,48	9492,12	44005,77	20417,03
WMA	85352,20	9314,89	44093,79	20442,24
MOMENTO	14290,90	-24368,58	535,69	4906,82
%R	100	0	42,80	31,35
%K	100	0	57,20	31,35
%D	96,20	9,95	57,34	23,13
RSI	100	2,78	53,66	20,86
MACD	4110,85	-6251,87	461,25	1654,33
A/D	105,45	-1,36	51,10	33,51
CCI	221,24	-295,56	14,13	106,51
Longo Prazo (Dados Mensais)				
SMA	67845.18	10434.68	41296.08	19000.15
WMA	67977.63	10455.65	41533.62	18955.14
MOMENTO	28861.13	-34409.19	1442.77	9541.25
%R	100	0	42.12	28.79
%K	100	0	57.88	28.79
%D	91.89	23.62	58.45	20.84
RSI	96.51	12.80	55.74	20.63
MACD	7670.82	-2331.99	1312.06	2799.98
A/D	100.62	-0.13	53.84	33.81
CCI	203.04	-217.14	17.07	102.27

Fonte: elaborada pelos autores.

Em movimento contrário, o momento aumenta em prazos maiores, o que é também esperado, já que representa uma medida de diferença entre preços de fechamento, ou seja, a diferença de preços de dez dias é, provavelmente menor que a diferença entre 10 semanas e, consequentemente entre dez meses, tanto que o desvio-padrão confirma esta constatação.

Com relação ao indicador RSI, é interessante observar que seu valor máximo permitido (100) não é atingido no longo prazo como também o valor mínimo (0), justamente pela questão de que, em um período tão longo é relativamente difícil os investidores permanecerem pressionando os preços numa mesma tendência, seja ela de compra ou de venda.

Já o MACD tem comportamento variante, não apresentando qualquer padrão. A/D também diminui a variação entre os extremos, mas com um desvio-padrão muito parecida em todos os cenários. Por fim, o valor médio de CCI aumenta em cenários de maior horizonte temporal, enquanto seu desvio padrão diminui, fazendo com que se torne um indicador menos disperso.

4.1 Desempenho dos Modelos

As Tabela 3, 4 e 5 apresentam, nesta mesma ordem, as previsões de cada um dos modelos no curto, médio e longo prazos. A coluna AC mostra a quantidade de Altas que o modelo identificou Corretamente; BC mostra a quantidade de Baixas previstas Corretamente. Já os erros de previsão são mostrados nas colunas AE (Altas Erradas) e BE (Baixas Erradas). O primeiro critério de análise do poder preditivo é a Acurácia (coluna ACC), que fornece o total de acertos (AC+BC) sobre o total de previsões (AC+BC+AE+BE). Os erros do Tipo I e II são dados por BE/(AC+BE) e AE/(AE+BC), respectivamente. Assim, as tabelas apresentadas a seguir tem como objetivo demonstrar se o modelo se comportou de modo adequado quando comparado com os dados reais, ou seja, se consegue antecipar o comportamento do índice IBOVESPA.

TABELA 3 - Matriz de confusão e medidas de erros (do Tipo I, ETI, e do Tipo II, ETII) para previsão de curto prazo (dados disponíveis até o dia D usados para prever alta ou baixa do período D+1). Os modelos SVMLin representa o método SVM com *Kernel* Linear, SVM-RBF é o SVM com *Kernel* de base radial (não linear), *Random Forest* e Logit, são os modelos baseados em *Random Forest* e regressão logística, respectivamente.

Modelo	AC	BC	AE	BE	ETI (%)	ETII(%)	ACC (%)
SVMLin	0	707	0	726	100	0	49,34
SVM-RBF	196	517	190	530	73	26,87	49,76
Random Forest	213	500	207	513	70,66	29,28	49,76
Logit	192	521	186	534	73,55	26,31	49,76

Fonte: elaborada pelos autores.

Cabe lembrar que tais resultados são previsões de dados que compõem a amostra de teste e que, dessa forma, não foram usados para a elaboração dos modelos.

Destaca-se na Tabela 3 que o modelo SVM com base Linear foi incapaz de distinguir o que aconteceria com o mercado no dia seguinte. De todo modo, o desempenho dos outros modelos ficou aquém do necessário, sequer chegando a 50%, o que se aproximaria de uma previsão a partir de jogar uma moeda. Outro fato a mencionar é a melhor capacidade destes três últimos em prever movimentos de baixa do IBOVESPA, sinalizando que tal modelagem é mais poderosa na antecipação de quedas do que subidas do mercado. Numa exemplificação simples

(com aproximações), é possível afirmar que a cada 100 quedas do índice, os modelos identificam aproximadamente 70 delas corretamente.

SVM com ambos os *kernels* decepcionam no médio prazo, conforme demonstrado na Tabela 4, mesmo tendo acurácia superior a 50%. Novamente o método não apresenta discernimento entre quedas e subidas e, seu melhor desempenho é justificado apenas pelo fato de que, ao longo do período pesquisado, ocorreram mais retornos negativos do IBOVESPA do que retornos positivos. Enquanto isso, o modelo baseado em *Random Forest* teve queda na sua qualidade quando comparado ao modelo de curto prazo. Uma possível explicação para este fato é a diminuição da amostra, que pode afetar a qualidade de modelos computacionais, ainda mais quando os indicadores não conseguem discriminar de maneira razoável as altas e as baixas do índice. Já Logit melhorou suas expectativas com relação ao curto prazo, errando ainda menos na previsão de quedas do mercado. Assim, para previsões de horizontes semanais e usando os indicadores técnicos, o modelo Logit teve melhor performance, principalmente em cenários de baixa.

TABELA 4 - Matriz de confusão e medidas de erros (do Tipo I, TEI, e do Tipo II, TEII) para previsão de médio prazo (dados disponíveis até a semana S usados para prever alta ou baixa do período S+1). Os modelos SVMLin representa o método SVM com *Kernel* Linear, SVM-RBF é o SVM com *Kernel* de base radial (não linear), *Random Forest* e Logit, são os modelos baseados em *Random Forest* e regressão logística, respectivamente.

Modelo	AC	BC	AE	BE	TEI (%)	TEII(%)	ACC (%)
SVMLin	0	164	0	151	100	0	52,06
SVM-RBF	0	164	0	151	100	0	52,06
Random Forest	51	103	61	100	66,23	37,2	48,89
Logit	25	136	28	126	83,44	17,07	51,11

Fonte: elaborada pelos autores.

Por último, os resultados para o longo prazo (dados disponíveis até o mês M para prever alta/baixa do mês M+1) mostram que os modelo de SVM com *kernel* não linear trouxe conseguiu uma performance superior aos demais, chegando a um patamar de 68% de precisão, conseguindo minimizar o erro do tipo II e, dessa forma, alcançar a casa de um dígito percentual (5,41%). Adicionalmente, o modelo Logit tem resultado de relativa qualidade, com quase 64% de acertos.

Numa visão geral destes resultados, o modelo de SVM para o longo prazo apresentou o maior poder preditivo, considerando apenas a acurácia e o erro do Tipo II (quando as quedas do mercado são previstas como altas). Nos outros cenários, é inadequado apontar um bom modelo, mas é interessante a capacidade do modelo Logit em identificar as quedas do mercado.

Tais resultados concordam com outras pesquisas, como Patel et al. (2015), Basak et al. (2016), somente no longo prazo, uma vez que SVM teve resultados insatisfatórios em curto e

médio prazos. Kim (2003) alerta para a necessidade de escolha dos parâmetros em SVM, assim como Wang et al. (2011) afirma que *Random Forest* também requer tais análises mais detalhadas acerca dos parâmetros e ainda o teste de outros indicadores que não somente advindos dos preços.

TABELA 5 - Matriz de confusão e medidas de erros (do Tipo I, TEI, e do Tipo II, TEII) para previsão de longo prazo (dados disponíveis até o mês M usados para prever alta ou baixa do período M+1). Os modelos SVMLin representa o método SVM com *Kernel* Linear, SVM-RBF é o SVM com *Kernel* de base radial (não linear), *Random Forest* e Logit, são os modelos baseados em *Random Forest* e regressão logística, respectivamente.

Modelo	**AC**	**BC**	**AE**	**BE**	**TEI (%)**	**TEII(%)**	**ACC (%)**
SVMLin	9	31	6	20	68,97	16,22	60,61
SVM-RBF	10	35	2	19	65,52	5,41	68,18
Random Forest	13	27	10	16	55,17	27,03	60,61
Logit	11	31	6	18	62,07	16,22	63,64

Fonte: elaborada pelos autores.

A evolução do poder preditivo ao longo de cenários maiores também está de acordo com o estudo de Basak et al. (2016). Ressalta-se que os autores chegaram a um nível de 90% de precisão em suas pesquisas.

4.3 Curvas ROC

Para auxiliar na tomada de decisão a partir dos modelos utilizados neste artigo (SVM, *Random Forest* e Logit), utilizou-se a curva ROC é uma das formas de avaliar a qualidade do modelo e pode ser usado como critério para definir o melhor modelo (Basak et al., 2016).

Assim, apresentam-se a seguir as figuras 2, 3 e 4, correspondentes às curvas ROC dos modelos em cada cenário (curto, médio e longo prazos). Analisando as curvas ROC da Figura 2 é possível dizer que para o curto prazo os modelos não apresentam resultados significativos, visto que a curva deles encontra-se em cima da diagonal, ou seja, os resultados apresentam em torno de 50% de acertos em suas previsões, o que não é ideal para a tomada de decisão na compra ou venda dos ativos.

Por fim, tem-se a curva ROC do longo prazo, a partir dela é possível chegar a um resultado satisfatório, visto que as curvas dos modelos de destoam da diagonal, demonstrando com clareza a melhor performance de todos os modelos. Assim, prever o comportamento do IBOVESPA no longo prazo, ou seja, no mês seguinte, exibe possibilidades de utilização por parte de investidores e especialistas que precisam de previsões mais acertadas a respeito da tendência do mercado.

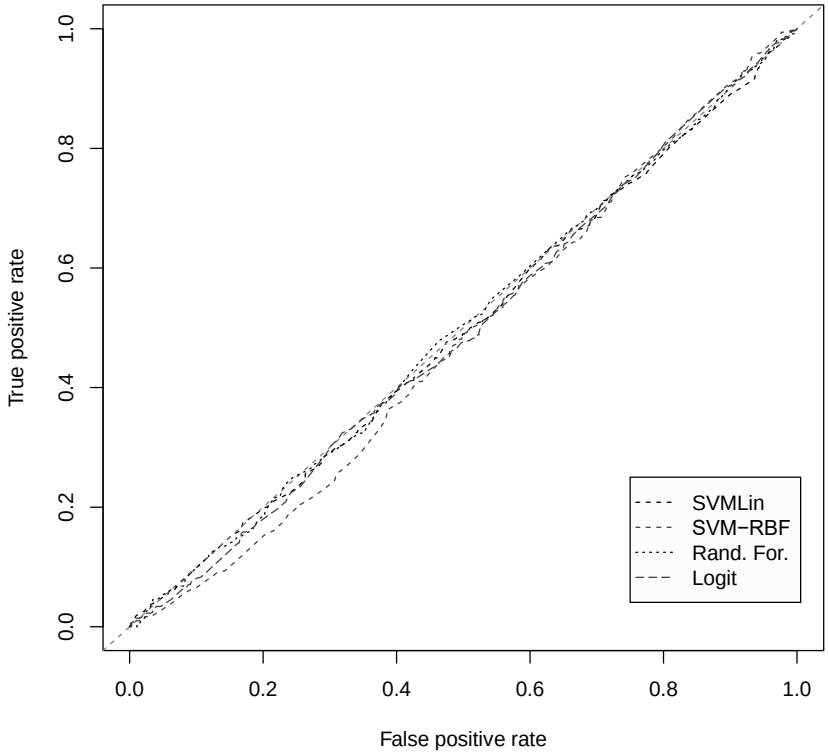

FIGURA 2: Curva ROC demonstrando o desempenho fraco de todos os modelos com o objetivo de prever os valores diários (curto prazo) do IBOVESPA. Tal resultado pode ser interpretado como tão preciso como jogar uma moeda.

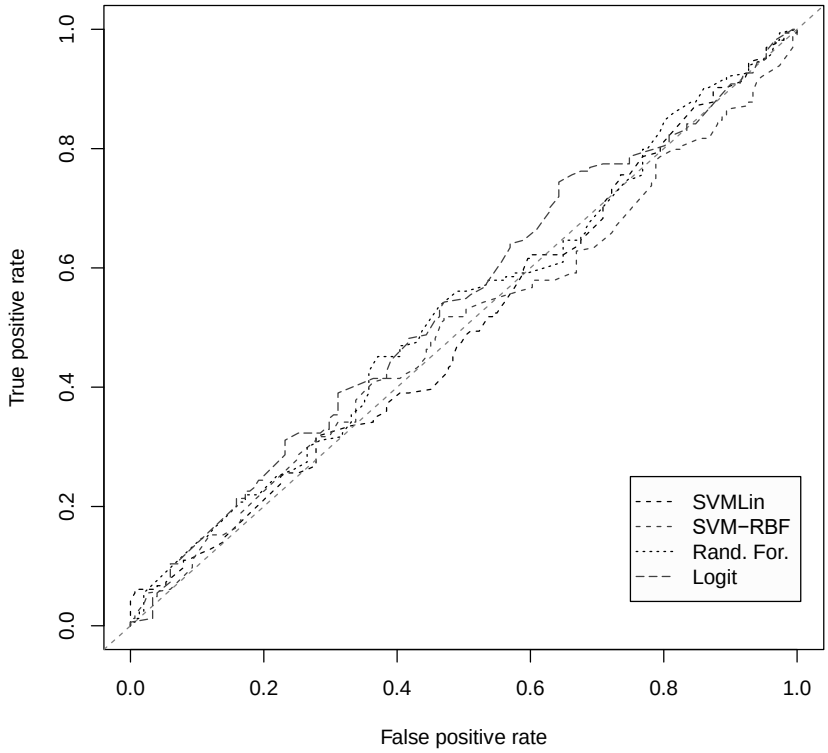

FIGURA 3: Curva ROC dos modelos de previsão para médio prazo (valores semanais do IBOVESPA). Comparado ao poder preditivo para curto prazo, a melhora é ainda insatisfatória, apesar de perceptível. Interessante notar que o modelo logit responde melhor a estes dados.

Para o médio prazo, tem-se que a curva dos modelos perto da diagonal, porém mais distantes do que as curvas no curto prazo. Apesar do distanciamento, ele ainda não o suficiente para a decisão da tomada de decisão da compra, já que os acertos da previsão continuam próximo de 50%.

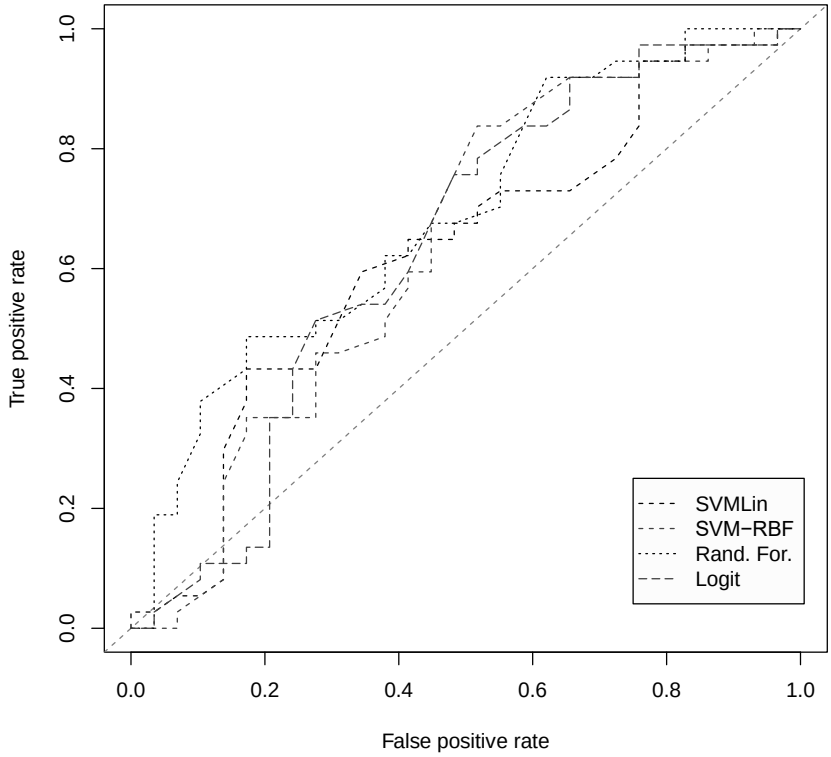

FIGURA 4: Curva ROC do longo prazo. Destaque para a melhoria de desempenho em relação aos modelos de curto e médio prazos.

Cabe ressaltar que os resultados ilustrados nas curvas ROC reforçam aqueles encontrados nas tabelas, nas quais constam as medidas de acurácia. Entretanto, as medidas de erro, que permitem analisar outras qualidades de cada modelo, estão implícitas nos gráficos e, sendo assim, calcular os erros do Tipo I e II, traz informação adicional relevante para a pesquisa.

5. Conclusão

Este artigo realizou o estudo empírico com o objetivo de demonstrar as eficiências dos seguintes modelos: máquina de suporte vetorial (SVM), *Random Forest* e regressão logística (Logit) para antever a direção dos preços dos ativos, no caso, do IBOVESPA.

Além dos resultados, foi reforçado os resultados de outros estudos citados anteriormente sobre a aplicação do SVM e *Random Forest* na área de finanças, comprando a eficiência para a previsão da direção da tendência, e consequentemente, para a tomada de decisão.

Os resultados, apresentados na seção acima, demonstram que tanto os modelos computacionais como a regressão logística obtêm melhores resultados no longo prazo, no qual

se prevê o preço para o mês t+1 a partir de dados do mês t e anteriores. Já os resultados no curto prazo, base de dados diária, são considerados insatisfatórios, visto que todos os dados beiram a 50% de acerto na indicação de alta ou baixa do índice, apesar de apresentarem relativa qualidade na previsão de quedas do índice.

Baseando-se na acurácia, calculada a partir da matriz de confusão de cada modelo, o método SVM apresenta melhor desempenho quando aplicado em horizontes mais longos (neste caso, em bases mensais), corroborando com estudos anteriores aplicados em outros contextos.

Apesar da superioridade apresentada, os modelos apresentam limitações. Uma delas é que os indicadores foram testados sob um parâmetro previamente fixado. Em estudos futuros, a seleção desse parâmetro poderá trazer melhorias nos resultados. A seleção da amostra também é defendida por alguns autores para que seja feita de forma aleatória (Dash & Dash, 2016; Patel et al., 2016; Wang et al., 2011), mas do ponto de vista prático, prever preços usando dados do futuro parece controverso, mesmo sendo defendido por pesquisadores. Assim, é mais uma oportunidade de investigação para a continuidade deste estudo.

No caso dos cenários de horizontes temporais, Basak et al. (2016) utilizaram 1, 2 e 3 meses para seu estudo. Como os resultados aqui apresentados são melhores no horizonte mensal, medir nestes intervalos seria mais uma possibilidade interessante de investigação, já que estes autores alcançaram um nível de 90% de acertos em suas previsões. Enfim, diversos caminhos podem dar continuidade e contribuir para o debate frequente nessa área de estudo.

Referências

Al Nasser, O. M., & Hajilee, M. (2016). Integration of emerging stock markets with global stock markets. *Research in International Business and Finance*, *36*, 1-12.

B3 - Brasil, Bolsa Balcão. Cotações Históricas | B3. Recuperado em 16 de Julho de 2018, de http://www.b3.com.br/pt_br/market-data-e-indices/servicos-de-dados/market-data/historicomercado-a-vista/cotacoes-historicas/

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, *83*, 405-417.

Basak, S., Kar, S., Saha, S., Khaidem, L., & Dey, S. R. (2018). Predicting the direction of stock market prices using tree-based classifiers. *The North American Journal of Economics and Finance*.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Dash, R., & Dash, P. K. (2016). A hybrid stock trading framework integrating technical analysis with machine learning techniques. *The Journal of Finance and Data Science*, *2*(1), 42-57.

Fan, A., & Palaniswami, M. (2001). Stock selection using support vector machines. In *Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on*(Vol. 3, pp. 1793-1798). IEEE.

Fang, J., Jacobsen, B., & Qin, Y. (2014). Predictability of the simple technical trading rules: An out-of-sample test. *Review of Financial Economics*, *23*(1), 30-45.

Kim, K. J. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, *55*(1-2), 307-319.

Kumar, M., & Thenmozhi, M. (2006). Forecasting stock index movement: A comparison of support vector machines and random forest. *Working Paper.*

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, *42*(1), 259-268.

Širůček, M., & Šíma, K. (2016). Optimized Indicators of Technical Analysis on the New York Stock Exchange. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, *64*(6), 2123-2131.

Wang, Q. G., Li, J., Qin, Q., & Ge, S. S. (2011, December). Linear, adaptive and nonlinear trading models for Singapore stock market with random forests. In *Control and Automation (ICCA), 2011 9th IEEE International Conference on* (pp. 726-731). IEEE.