Congresso Internacional de Administração
ADM 2021
Administração Ágil
Inovação e Trabalho Remoto
25 a 27 de outubro
Ponta Grossa - Paraná - Brasil

Machine Learning aplicada a finanças: Previsão por meio de indicadores econômicos-financeiros

Machine Learning applied to finance: Profit forecast through economic and financial indicators

ÁREA TEMÁTICA: FINANÇAS

Olavo Soares da Conceição, Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP), Brasil,

E-mail: crmsca@gmail.com

Ricardo Maroni Neto, Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP), Brasil,

E-mail: maroni@ifsp.edu.br

Resumo

Este trabalho aplica um modelo de *machine learning* que emprega a regressão linear para fazer previsões. Utilizando como fonte de dados indicadores econômico-financeiros de liquidez e rentabilidade, o presente estudo verifica qual a eficácia do modelo em relação aos valores reais do lucro líquido projetado. Constatou-se a previsão do lucro líquido com precisão de 90,55%, ou seja, uma margem eficaz de previsibilidade por meio de dados históricos.

Palavras-chave: *Machine learning*, regressão linear, liquidez, rentabilidade, lucro líquido.

Abstract

This work applies a machine learning model that employs linear regression to make predictions. Using economic-financial indicators of liquidity and profitability as a data source, this study verifies the effectiveness of the model in relation to the actual values of the projected net income. The forecast of net income was found to be accurate to 90.55%, that is, an effective predictability margin through historical data.

Keyword: *Machine learning; linear regression; liquidity; profitability; net profit.*

1 INTRODUÇÃO

O aprendizado de máquina consiste no reconhecimento de padrões ocultos em dados históricos a partir de técnicas estatísticas, matemáticas e ferramentas computacionais. A partir desses dados estrutura-se um modelo capaz de fazer uma previsão aproximada quando apresentado a dados novos nunca antes vistos.

Existem diversos trabalhos que aplicam o aprendizado de máquina relacionado a finanças. Muitos realizam a análise de variáveis categóricas e discorrem sobre o risco de

crédito. Estes efetuam a classificação das variáveis através do uso de modelos de regressão logística, dentre outros métodos, para a verificação da probabilidade de inadimplência dos clientes do setor bancário. (SINHORIGNO, 2007; LUKOSIUNAS, 2018; MARRA, 2019)

O problema de pesquisa proposto neste artigo procura responder a seguinte questão: qual a eficácia e precisão da previsibilidade do modelo em relação aos valores reais do lucro líquido da companhia?

O objetivo geral deste projeto é aplicar um modelo de *machine learning* que utiliza a regressão linear para fazer uma previsão do lucro líquido futuro das empresas. Os objetivos específicos do artigo são:

- Apresentar os principais conceitos de *machine learning*;

-Apresentar o método de regressão linear;

-Construir um modelo de predição de lucro com base em indicadores econômico-financeiros.

2 REFERENCIAL TEÓRICO

2.1.1 Aprendizagem de máquina

A Aprendizagem de Máquina pode ser definida genericamente como a área multidisciplinar que se ocupa com o desenvolvimento, análise e aplicação de métodos para a detecção automática de padrões em conjuntos de dados. Seja qual for o tipo de dado, o objetivo final da aprendizagem automática de padrões é a mineração de informações úteis que possam orientar a tomada de decisões sobre o problema em estudo (FINKLER, 2017).

Géron (2017) afirma que o aprendizado de máquina é a ciência e a arte de programar computadores para que eles possam aprender com os dados. Nascimento (2020) que utiliza o termo *machine learning* para descrever uma coleção diversificada de modelos de alta dimensão para a previsão estatística combinada com o chamados métodos de regularização para a seleção de modelos e mitigação do super ajuste e algoritmos eficientes para pesquisar entre um grande número de possíveis especificações de modelo.

Segundo Raschka (2017), a divisão da amostra em conjunto de dados de treinamento e teste é realizada com o intuito de verificar se o modelo apresenta boa predição não apenas nos dados que foram utilizados no ajuste (treinamento), mas também na capacidade de generalização para uma nova amostra (teste).

Silva, Almeida & Ramalho (2020) destacam que os modelos preditivos são compostos por dois principais objetivos: seleção e avaliação. No que se refere a selecionar, a performance de diferentes modelos é avaliada por meio de critérios de medidas de desempenho para que, a partir de um equilíbrio entre viés-variância, seja selecionado o modelo que resulta em uma melhor acurácia e desempenho no conjunto de treinamento. Já no que se refere à finalidade de avaliar, após a definição da melhor performance, busca-se estimar o modelo em novas observações, na base de teste.

A maioria dos problemas de aprendizagem de máquina podem ser categorizados com aprendizado supervisionado ou não supervisionado. De acordo com Athey (2018) o aprendizado não supervisionado envolve encontrar grupos de observações que são semelhantes em termos de suas variáveis, e, a saída de um típico modelo não supervisionado descrevem uma mistura de tópicos ou grupos aos quais uma observação possa pertencer.

Neste artigo é aplicado o aprendizado supervisionado em que para cada observação da base de dados do preditor $Xi, i = 1, .. , n$ há uma resposta associada Yi. O principal objetivo é ajustar um modelo que relaciona a resposta aos preditores, com o objetivo de prever com

precisão a resposta para eventos futuros. Alguns métodos clássicos do aprendizado supervisionado são a regressão linear para variáveis quantitativas e regressão logística para variáveis categóricas. (JAMES et al., 2017)

2.1.2 Regressão linear

A análise de regressão avalia a amplitude da alteração em uma variável explicada, decorrente de alterações em outra variável explicativa, com base em um modelo. O modelo de regressão serve paraprever comportamentos cuja base é a associação entre duas variáveis que geralmente possuem uma boa correlação.

A análise de regressão linear estuda a relação entre a variável dependente e uma ou várias variáveis independentes. Esta relação representa-se por meio de um modelo matemático, ou seja, por uma equação que associa a variável dependente com as variáveis independentes (FINKLER, 2017).

Quanto aos tipos de regressão linear existem duas categorias: a regressão linear simples e a regressão linear múltipla. O modelo de regressão linear simples define-se como a relação linear entre a variável dependente e uma variável independente. Enquanto que na regressão linear múltipla assume-se que exista uma relação linear entre uma variável dependente e várias variáveis independentes (RODRIGUES; NUNES, 2012).

Utiliza-se neste trabalho a abordagem da regressão linear múltipla que tem como variáveis independentes (explicativas) os indicadores econômicos financeiros de rentabilidade e liquidez e como variável dependente a ser prevista tem-se o lucro líquido futuro.

2.1.3 Especificação do modelo

O modelo escolhido para realização deste artigo baseia-se no estudo de Costa, Macedo, Câmara e Batista (2013) que buscaram compreender como a gestão do capital de giro influencia a rentabilidade das empresas, considerando o setor em que estão inseridas. O estudo desses autores contribui para o avanço do tema no Brasil, podendo servir como base para pesquisas futuras que considerarem outras variáveis moderadoras.

Diferente do sugerido por Costa et al (2013) neste trabalho são utilizados alguns indicadores distintos e com a finalidade pertinentes ao objetivo deste artigo que é a maximização da previsibilidade do lucro líquido futuro (LL'). Descreve-se matematicamente o modelo de regressão linear proposto:

$$LL' = \propto + \beta1(LC) + \beta2(LO) + \beta3(LI) + \beta4(LG) + \beta5(ROE) + \beta6(ROA) + \varepsilon$$

Onde: LL' = Representa o Lucro líquido que o modelo tentará prever; $\propto$: É uma constante, que representa a interceptação da reta com o eixo vertical; β: Representa a inclinação (coeficiente angular) em relação à variável explicativa; LC: Índice de Liquidez Corrente; LO: Índice de Liquidez Operacional; LS: Índice de Liquidez Seca ; LI: Índice de Liquidez Imediata; ROE: Retorno sobre o capital próprio; ROA: Retorno sobre ativos; ε: Representa todos os fatores residuais mais os possíveis erros de medição.

2.3.1 Correlação entre as variáveis

Uma correlação é uma relação entre duas variáveis. Os dados podem ser representados por pares ordenados (x, y), onde x é a variável independente e y é a variável dependente. O coeficiente de correlação linear (r) entre duas variáveis é calculado como:

Equação 1 - Correlação entre duas variáveis

$$r = \frac{n.\sum X.Y - \sum X.\sum Y}{\sqrt{[n.\sum X^2 - (\sum X^2)].[n.\sum Y^2 - (\sum Y)^2]}}$$

Através da análise do valor do coeficiente de correlação linear (r) entre duas variáveis é possível classificá-las como: ausente de correlação quando r = 0; fraca correlação quando | r | < 0.5 ; forte correlação | r | > 0.5 e perfeita correlação quando | r | = 1 (SOUZA, 2020). Em um primeiro momento é realizada uma análise exploratória através da visualização de gráficos para uma melhor compreensão dos dados obtidos. Os resultados obtidos são úteis na análise e seleção das variáveis a serem incluídas no modelo de regressão.

2.4 Métricas de avaliação

A análise exploratória de dados emprega grande variedade de técnicas gráficas e quantitativas, visando maximizar a obtenção de informações ocultas na sua estrutura, descobrir variáveis importantes em suas tendências, detectar comportamentos anômalos do fenômeno, testar se são válidas as hipóteses assumidas, escolher modelos e determinar o número ótimo de variáveis. (JAMES, 2017).

Nos modelos de regressão linear uma métrica de avaliação é a diferença entre os valores reais observados e os valores de resposta que o modelo previu. Essa diferença é chamada de resíduo e considera-se como pressuposto do modelo que os resíduos sejam independentes e que tenham distribuição normal.

Uma forma padrão de medir o desempenho de um modelo ou estimador na previsão de dados quantitativos é através da raiz do erro quadrático médio[1] (RMSE). Conforme demonstrado por Géron (2017) na equação 2 o RMSE é a raiz quadrada da média das diferenças quadradas entre a previsão e a observação real.

Equação 2 - Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Yi - \hat{Y})^2}{n}}$$

O RMSE mede o desvio padrão dos erros que o sistema comete em suas previsões. O desvio padrão, geralmente denotado por σ (a letra grega sigma), é a raiz quadrada da variância, que é a média do desvio quadrado da média. O propósito primário é minimizar o erro quadrático médio das previsões (RMSE) e atingir um coeficiente de determinação satisfatório (NASCIMENTO, 2020).

O erro médio absoluto é uma medida de erros entre observações pareadas que expressam o mesmo fenômeno. Essa métrica consiste em calcular o residual de cada ponto, cujos valores residuais negativos e positivos não se anulam. Após esse agrupamento, calcula-se então a média desses ‘residuais’ (SSE). Essa é a variação inexplicada dos valores observados em relação aos valores estimados (GÉRON, 2017).

Em contrapartida, a variação explicada dos valores observados em relação aos valores estimados é a soma de quadrados devido aos resíduos (SSR). Vieira (2004) afirma que ambas métricas são necessárias para o cálculo da soma dos quadrados totais (SST) conforme observado na equação 3.

[1] Sigla do inglês Root Mean Square Error (RMSE)

Equação 3 - Soma dos Quadrados dos Erros

$$SST = SSE + SSR$$

$$\sum_{i=1}^{n}(Yi - \bar{Y})^2 = \sum_{i=1}^{n}(Yi - \widehat{Yi})^2 + \sum_{i=1}^{n}(\widehat{Yi} - \bar{Y})^2$$

A soma dos quadrados totais (SST) e a soma do quadrado dos resíduos são determinantes para o cálculo do coeficiente de determinação (R^2).

O quociente entre SSR e SST revela uma medida de proporção da variação total que é explicada pelo modelo de regressão. A esta medida dá-se o nome de coeficiente de determinação (R^2), onde $0 \leq R^2 \leq 1$ e quanto mais próximo de 1 significa que grande parte da variação de Y é explicada linearmente pelas variáveis independentes.

Equação 4 - Coeficiente de determinação

$$R^2 = 1 - \frac{SSR}{SST}$$

Este coeficiente pode ser utilizado como uma medida da qualidade do ajustamento, ou como medida da confiança depositada na equação de regressão como instrumento de previsão.

Um valor grande de R^2 não indica necessariamente que seja um bom modelo. A adição de uma variável ao modelo sempre aumenta o R^2, sem importar se a variável é ou não estatisticamente significativa. Portanto, são necessárias análises conjuntas de outras informações para determinar a competência do modelo. Outro critério similar ao R^2 é o coeficiente de determinação ajustado ($\bar{R}^2$) que leva em conta o número de variáveis do modelo:

Equação 5 - Coeficiente de determinação ajustado

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n - 1}{n - p - 1}$$

Onde: n = número de observações;

p = número de variáveis independentes (VIEIRA, 2004)

Rodrigues e Nunes (2012) confirmam que é preferível utilizar o coeficiente de determinação ajustado para saber se o modelo de regressão providencia um bom ajustamento aos dados, pois, é uma medida ajustada do coeficiente de determinação que é "penalizada" quando são adicionadas variáveis pouco explicativas.

2.2.1 Indicadores

Após o processo de registro e mensuração dos eventos econômicos que alteram o patrimônio de uma empresa é por meio da análise das demonstrações financeiras que se faz a adequada avaliação da sua situação econômico-financeira.

Para tanto é necessário primeiro realizar ajustes do balanço patrimonial, ou seja, é preciso realizar uma reclassificação, distinguindo as contas do Ativo e do Passivo (AIKAWA; MARONI NETO, 2020). Esta atividade é chamada de padronização e tem como motivo: a simplificação, comparabilidade, adequação aos objetivos da análise, precisão nas classificações de contas e a descoberta de erros (MATARAZZO, 2010). Para este trabalho utiliza-se como fonte de dados os indicadores econômico-financeiros de liquidez e rentabilidade.

Os indicadores de liquidez visam medir a capacidade de pagamento de uma empresa, ou seja, sua habilidade em cumprir corretamente as obrigações assumidas (SOARES; OLIVEIRA, 2019). Os principais índices para avaliar a situação econômica da empresa estão divididos em índice de liquidez geral, índice de liquidez corrente, índice de liquidez seca e índice de liquidez imediata (FERRAZ; SOUSA; NOVAES, 2017).

A rentabilidade é medida pela relação entre o lucro e o investimento realizado na empresa (SOUSA et al., 2017). O principal objetivo de qualquer empresa é maximizar a riqueza dos proprietários por meio da alta rentabilidade. Mas manter a liquidez da empresa também é um objetivo importante. O problema é que aumentar os lucros à custa de liquidez pode trazer problemas para a empresa (COSTA et al., 2013). Os índicadores utilizados neste artigo, a sua aplicabilidade e a fundamentação matemática são baseados no estudo realizados por Ferraz et al (2017).

3 METODOLOGIA

3.1 Delineamento da Pesquisa

Quanto à natureza, a pesquisa pode constituir-se em um trabalho científico original ou em uma pesquisa aplicada (ANDRADE, 2010). Para Prodanov e Freitas (2013), a pesquisa aplicada tem como finalidade gerar conhecimento, abrangendo verdades e interesses locais para solucionar problemas específicos, para realizar uma aplicação prática. A natureza desta pesquisa configura-se como aplicada pois busca gerar conhecimento para a aplicação prática e dirigida à solução de problemas que contenham objetivos anteriormente definidos.

A observação envolve o registro de padrões de comportamentos, bem como dados sobre objetos e eventos de forma sistemática, para obter informações sobre o fenômeno de interesse (MALHOTRA, 2019). Quanto ao procedimento utilizado para obtenção dos dados deste trabalho foi realizado o método de observação dos demonstrativos econômicos financeiros da empresa WEG.

A pesquisa descritiva é um tipo de pesquisa conclusiva que tem como objetivo a descrição de características ou funções de mercado (MALHOTRA, 2019). Os fatos são observados, registrados, analisados, classificados e interpretados, sendo que uma de suas características é a técnica padronizada da coleta de dados (ANDRADE, 2010). Este estudo caracteriza-se com pesquisa descritiva, pois, foram realizadas observações e padronização financeira dos relatórios contábeis trimestrais da empresa WEG desde o ano de 2009 até o ano de 2020.

Quanto à abordagem, a pesquisa se classifica como exploratória, pois conforme Prodanov e Freitas (2013), este tipo de pesquisa tem por objetivo proporcionar maiores informações sobre o assunto que será analisado. Embora o objeto de análise deste trabalho seja a empresa WEG os procedimentos metodológicos aplicados são extensíveis a qualquer empresa de setor e porte diverso, desde que haja dados contábeis disponíveis.

3.2 Procedimentos metodológicos

Como os indicadores possuem valores numéricos e não categóricos optou-se pela abordagem de aprendizado supervisionado para construção do modelo. O qual consiste no treinamento do modelo com a apresentação das variáveis quantitativas e como resultado tem-se a previsão da variável alvo (target). A finalidade do algoritmo é encontrar correlações e reconhecer padrões nas variáveis de entrada e assim fazer uma estimativa aproximada do resultado desejado.

A partir dos balanços padronizados realiza-se a construção dos indicadores de liquidez e dos indicadores de rentabilidade com o auxílio de planilha eletrônica. O número de indicadores calculados referente aos períodos trimestrais é de 360 no total.

Emprega-se o software livre RStudio, um ambiente de desenvolvimento integrado para R, que é uma linguagem de programação para gráficos e cálculos estatísticos, assim como da linguagem de programação Python através do ambiente *Júpiter notebook* e recursos da biblioteca scikit Learning assim como outras bibliotecas de manipulação de dados e visualização dos dados (Pandas, Numpy, Seaborn e Matplotlib) para a construção do data frame e implementação das instruções necessárias para o aprendizado. Como variável *target*, ou seja, a variável que se deseja efetuar a previsão, é utilizada a coluna do lucros líquidos trimestrais e como dados de entrada os indicadores de liquidez e de rentabilidade.

Os dados obtidos são divididos entre dados de treino e dados de teste, sendo que 70% dos dados são utilizados como dados de treino e 30% para dados de teste do modelo. Após o treinamento com os dados de treino é efetuado o *fit* do modelo, que consiste em apresentar os 30% dos dados que o algoritmo não conhece e verificar a acurácia e precisão do modelo com base nesses dados de teste, sendo possível quantificar a previsão efetuada pelo modelo (James et al., 2017).

A precisão do modelo deste artigo tem como base o coeficiente de de determinação ajustado que penaliza a inclusão de regressores pouco explicativos. Na hipótese de se verificar um valor do coeficiente de determinação ajustado superior a 70% é considerado que o modelo será capaz de realizar previsões do lucro líquido futuro ao se deparar com novos conjuntos de dados. Para essa finalidade é realizada a troca de variáveis preditoras e inclusão de outros indicadores mais relevantes.

4 RESULTADOS

4.1 O modelo estimado

São apresentados os resultados da regressão do modelo em função das variáveis explicativas, estimativas dos parâmetros, análise de variância e análise de resíduos no apêndice 1. Na linha de liquidez corrente o valor *t* de 1.066 refere-se ao *teste t* (Interceptação) da estimativa 347.678 dividido pelo desvio padrão 326.090. Pr (> | t |) fornece o valor p para esse teste t (a proporção da distribuição t que é maior do que o valor absoluto de sua estatística t).

Os asteriscos após Pr(>| t |) fornecem uma maneira visualmente acessível de avaliar se as variáveis do modelo foram significativas, quanto mais asteriscos mais relevantes. Verifica-se a maior relevância para os indicadores de Liquidez Geral (1.74e-11 ***), ROE(1.72e-09 ***) e ROA(3.54e-05 ***). Através da análise dos resultados obtidos e expostos no apêndice 1 é apresentada a formulação matemática do modelo de regressão obtido:

Equação 6 - Cálculo do Modelo de regressão linear

$$LL' = -5{,}32 \times 10^{6} + 3{,}48 \times 10^{5}(LC) - 1{,}08 \times 10^{5}(LO) - 5{,}23 \times 10^{5}(LS) + 1{,}63 \times 10^{5}(LI) + 4{,}32 \times 10^{6}(LG) + 2{,}15 \times 10^{5}(ROE) - 3{,}25 \times 10^{7}(ROA).$$ [2]

Fonte: Elaborado pelo autor

[2] Valores dos estimadores na íntegra: $\propto$ = -5.319.424, $\beta 1$ = 347.678, $\beta 2$ = - 108.609, $\beta 3$ = - 522.747, $\beta 4$ = 162.751, $\beta 5$ = 4319.923, $\beta 6$ = 215.313, $\beta 7$ = - 32.531.472.

Quanto ao coeficiente de correlação (r) entre duas variáveis ele flutua entre -1 e 1. Quando r < 0 a relação entre as variáveis é inversa, por outro lado um r > 0 descreve uma relação direta entre os indicadores sendo o resultado mais comum, porém com magnitudes diferentes. A correlação perfeita ocorre quando r = -1 ou 1, apresentando uma condição em que os indicadores são iguais. A partir da matriz mostrada no gráfico 1, é possível observar que os indicadores que possuem maior correlação dispõem valores próximos a |1|, e as menos correlacionadas apresentam valores menores que |0.5|.

4.2 Correlação entre os indicadores

Conforme observado no gráfico 1 é digna de ênfase é a constatação da alta correlação entre as seguintes variáveis: Liquidez seca e liquidez corrente (0,97), da liquidez geral com o lucro líquido (0,91) e do ROA com a liquidez geral (0,91). Evidenciando o caráter das grandezas serem diretamente proporcionais e a relevância desses indicadores para a precisão do modelo.

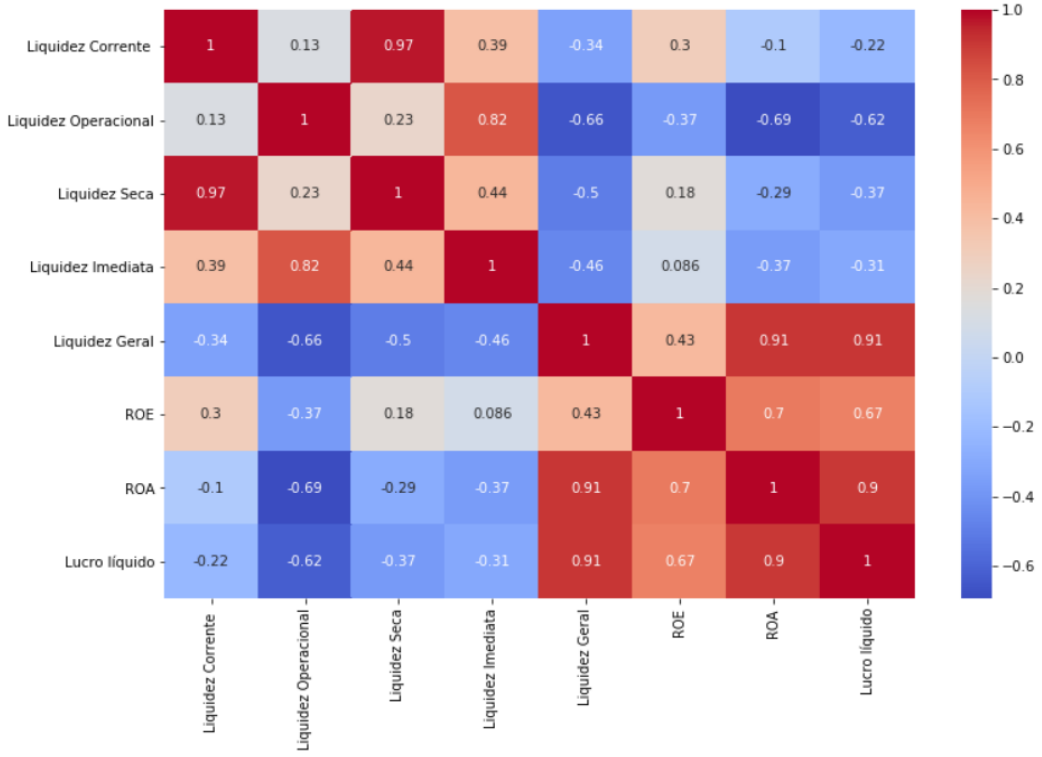

Gráfico 1: Correlação entre as variáveis

Fonte - Elaborada pelo autor

De acordo com o gráfico 2 e conforme constatado pela análise dos dados obtidos é verificado que a maior concentração dos indicadores de liquidez seca (ILS) encontram-se na faixa de 1.5 a 2 e de liquidez corrente (ILC) de 2 a 2.5 mostrando que a empresa opera com o uso de recursos de rápida conversibilidade para pagamento de dívidas de curto prazo.

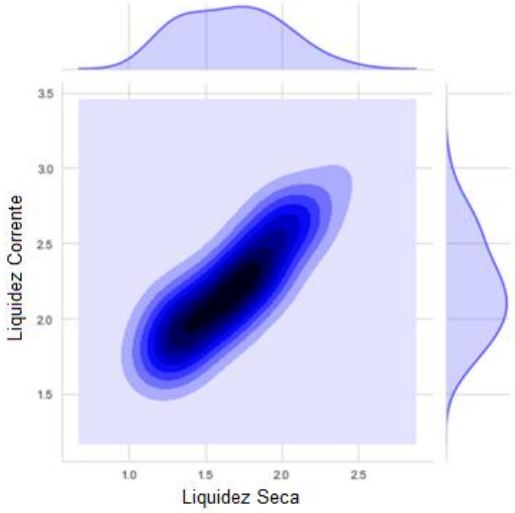

Gráfico 2 - Liquidez seca x Liquidez corrente

Fonte: Elaborado pelo autor

O índice de liquidez geral (ILG) evidencia a liquidez ampla e revela a existência de margem de segurança da empresa em honrar com suas obrigações. Os valores positivos e maiores que 1 revelam que os recursos do patrimônio líquido financiam os ativos circulantes. Percebe-se que a situação econômico-financeira da empresa é sólida pois apresenta os valores de liquidez geral entre 1.2 e 1.4 (Gráfico 3). Também é possível verificar em sua maioria o comportamento diretamente proporcional entre a liquidez geral e o lucro líquido.

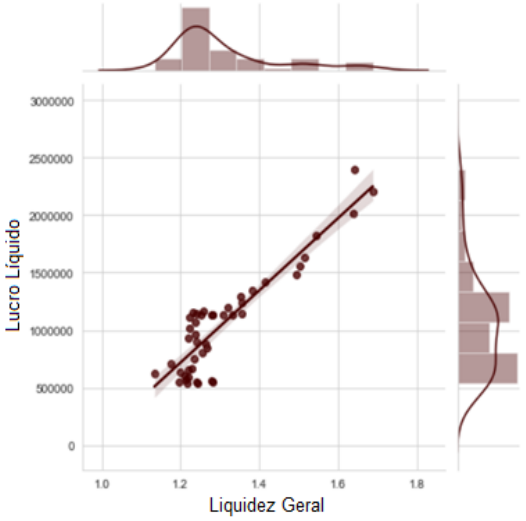

Gráfico 3: Liquidez geral x Lucro líquido

Fonte: Elaborado pelo autor

O retorno sobre ativos (ROA) mostra o quão rentável são os ativos de uma empresa em relação a geração de receita. No período observado a maior concentração de valores foi próximo a 0.09 e a liquidez geral manteve-se na faixa positiva e menor que 1.4 de acordo com o observado no gráfico 4.

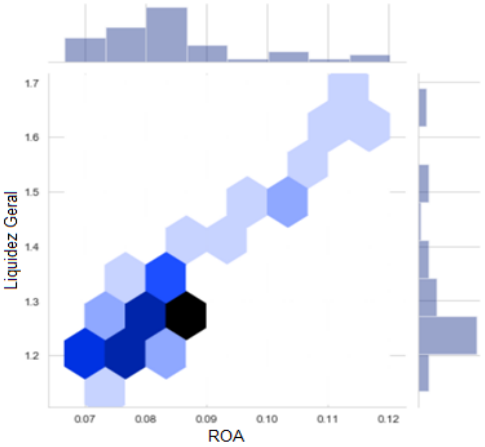

Gráfico 4: Liquidez geral x ROA

Fonte: Elaborado pelo autor

4.3 Validação do modelo

De acordo com o valor calculado pelo modelo, o erro médio absoluto (SSE) foi de 111.309,20 e o resultado obtido pelo cálculo da raiz quadrada da média das diferenças quadradas (RMSE) foi de 136.934,46 que significa a diferença entre a previsão e a observação real. Quanto mais próximo de zero, maior a qualidade dos valores medidos ou estimados. Todas as estimativas compartilharam o objetivo básico de minimizar o erro quadrático médio das previsões (RMSE) e obter um coeficiente de determinação apropriado.

O coeficiente de determinação (R^2) expressa a quantidade da variância dos dados que é explicada pelo modelo linear. Dessa maneira é possível afirmar que o modelo explica 92.05% da variabilidade dos dados antes do ajuste e 90,55% considerando o coeficiente de determinação ajustado ($\bar{R}^2$). O normal é que o experimentador selecione o modelo de regressão que tenha o valor máximo de$\bar{R}^2$. Indicando que o modelo proposto ultrapassa a hipótese inicial de alcançar um valor superior a 70% do coeficiente de determinação ajustado ($\bar{R}^2$). Evidenciando a capacidade do modelo de realizar previsões com uma precisão adequada. Contudo, a maneira mais eficaz de testar a capacidade do modelo é comparar os resultados previstos com os realizados efetivamente.

4.4 Resultado previstos x realizados

No gráfico 5 observa-se através da visualização dos valores reais do lucro líquido do período e dos valores previstos pelo modelo um comportamento linear que é esperado em um modelo de regressão linear eficaz. A qualidade do ajuste é verificada com a comparação entre os valores observados na amostra e os valores obtidos a partir da aplicação do modelo; quanto mais próximos forem esses valores, maior será a qualidade da equação construída.

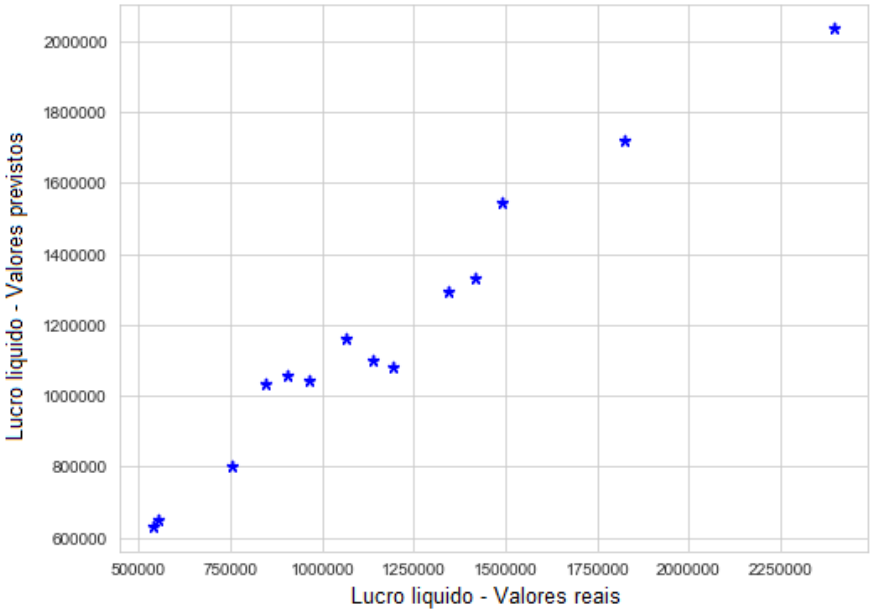

Gráfico 5: Valores previstos x valores reais do lucro líquido

Fonte: Elaborado pelo autor

No apêndice 2 são apresentados na íntegra os valores previstos pelo modelo do lucro líquido em comparação com os realizados efetivamente. A maior aproximação constatada teve uma lacuna de 75.543 [3] entre o rendimento real da WEG no trimestre e o que havia sido previsto pelo modelo baseado nos indicadores de liquidez e rentabilidade anteriores.

5 CONSIDERAÇÕES FINAIS

O problema de pesquisa do presente estudo é verificar qual a eficácia e precisão da previsibilidade do modelo em relação aos valores reais do lucro líquido. Por intermédio da aplicação de um modelo de *machine learning* que utiliza a regressão linear, constatou-se a previsão do lucro líquido com precisão de 90,55%, ou seja, uma margem de erro eficaz de previsibilidade por meio de dados históricos. Assim, constata-se que o modelo é eficaz e preciso.

Os objetivos específicos do artigo são atendidos por meio do levantamento de pesquisas relacionadas a *machine learning* aplicada a finanças. Como ferramenta para alcançar o objetivo prático de construção e avaliação do modelo, são utilizados recursos da linguagem Python e R.

Mediante a análise exploratória dos indicadores de liquidez e de rentabilidade da empresa estudada, verifica-se a correlação entre as variáveis independentes e sua relevância para o modelo. Essas mesmas regras podem ser utilizadas para outros conjuntos de dados de outras empresas ou para qualquer problema que envolva previsões com variáveis contínuas.

A precisão do modelo e as métricas de avaliação devem ser definidas pelo problema de negócio em que estiver sendo trabalhado. Para medir a adequação do modelo, podem ser empregadas diversas técnicas.

Neste artigo são realizadas a verificação da correlação entre as variáveis, a visualização dos indicadores mais relevantes, averiguação do comportamento linear entre os valores reais e dos previstos pelo modelo, e ainda a constatação de que os métodos dos

[3] (Em milhares de reais)

mínimos quadrados e o coeficiente de determinação ajustado mostram-se satisfatórios. O que evidencia que o modelo é capaz de realizar previsões para novos conjuntos de dados.

Os dados podem sofrer diversas transformações antes de serem expostos ao algoritmo e pode haver a necessidade de obter mais dados ou utilizar outros algoritmos para a previsibilidade desejada, pois a regressão linear é um entre muitos algoritmos capazes de realizar essa tarefa. Por isso, todas as decisões do processo de escolha das variáveis, da análise exploratória e escolha do algoritmo devem ser guiadas pelo objeto investigado.

Outras ferramentas podem ser utilizadas para elaboração de um modelo de *machine learning* que seja aplicado a finanças. Como recomendações para futuras pesquisas, o modelo pode ser desenvolvido com a utilização da técnica de Redes Neurais Artificiais para a estimação dos valores projetados. As variáveis preditoras podem ser alteradas e indicadores mais relevantes podem ser incluídos. Além disso, o modelo proposto neste trabalho pode ser aplicado a outras empresas para comparação dos resultados.

REFERÊNCIAS

AIKAWA, R., MARONI NETO, R. (2020). Algumas Considerações sobre o Indicador EconomicValueAdded (EVA®): O Valor na Petrobras. In: LINHARES, W. L. (Org.) As ciências sociais aplicadas e a interface com vários saberes. Ponta Grossa – PR. Atena Editora, p.11-26.

ANDRADE, M. M. (2010). Introdução à metodologia do trabalho científico: elaboração de trabalhos na graduação. 10. ed. São Paulo. Editora Atlas.

ATHEY, S. (2018). *The impact of machine learning on economics. In The economics of artificial intelligence: An agenda* (pp. 507-547). *University of Chicago Press.*

SASSI, C. P. et al. (2011). Modelos de regressão linear múltipla utilizando os softwares R e STATISTICA: uma aplicação a dados de conservação de frutas. ICMC – USP.

COSTA, R. B. L. et al. (2013). A influência da gestão do capital de giro no desempenho financeiro de empresas listadas na BM &FBovespa (2001-2010). Revista de Contabilidade e Controladoria. UFPR, Curitiba, v. 5, n.1, p. 65-81, jan./abr.

FERRAZ, P. S., SOUSA, E. F., NOVAES, P. V. C. (2017). Relação entre liquidez e rentabilidade das empresas listadas na BM &FBovespa.ConTexto, Porto Alegre, v. 17, n. 35, p. 55-67, jan./abr.

FINKLER, A. C. (2017). Aprendizagem de máquina aplicada à previsão dos movimentos do Ibovespa. Dissertação de mestrado. Universidade Federal do Paraná. Curitiba.

GÉRON, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems.* O'Reilly Media, Inc.

JAMES, G., WITTEN, D., HASTIE, T., TIBSHIRANI, R. (2017). *An Introduction to Statistical Learning with Applications in R.* New York: Springer.

LUKOSIUNAS, A. (2018). Aplicação de técnicas de *machine learning* em modelos de escore de crédito. INSPER - Instituto de Ensino e Pesquisa.

MALHOTRA, N. (2019). Pesquisa de marketing: uma orientação aplicada. 7. ed. Porto Alegre: Bookman.

MARRA, V. N. (2019). Previsão de dificuldades financeiras em empresas latino-americanas via aprendizagem de máquina. Dissertação de mestrado. Universidade Federal de Uberlândia.

MATARAZZO, D. C. (2010). Análise financeira de balanços: abordagem básica e gerencial. 7 ed. São Paulo, Atlas.

NASCIMENTO, C. A. (2020). Precificação de ativos via *machine learning*: uma extensão de métodos lineares esparsos. Dissertação de mestrado - Programa de Pós Graduação em Economia. Universidade de São Paulo. Ribeirão Preto.

PRODANOV, C. C., Freitas, E. C. (2013). Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico. 2ª ed. Novo Hamburgo, Editora Feevale.

RODRIGUES, S. C. A., NUNES, C. M. P. (2012). Modelo de Regressão Linear e suas Aplicações. Tese de mestrado. Universidade da Beira Interior. Covilhã.

SANTOS, G. C. (2020). Algoritmos de *machine learning* para previsão de ações da B3. Dissertação de mestrado. Universidade Federal de Uberlândia. Uberlândia – MG.

SILVA, A. F., ALMEIDA, A. T. C., RAMALHO, H. M. B. (2020). Predição do Risco de Reprovação no Ensino Superior Usando Algoritmos de *Machine Learning*. Teoria e Prática em Administração, Jul.-Dez.

SINHORIGNO, S., VICENTE, R. (2007). Previsão de inadimplência de transações com cartão de crédito: um estudo comparativo. Dissertação de mestrado. Universidade de São Paulo, São Paulo.

SOARES, P. H. S., FARIA, J. A., OLIVEIRA, J. J. (2019). Análise das demonstrações contábeis: Uma proposta de referência de índices de liquidez para empresas brasileiras.ConTexto, Porto Alegre, v. 19, n. 43, p. 44-57, set./dez.

SOUZA, D. L. (2020). *Machine learning* e economia comportamental: mental accounting e teoria do prospecto no comportamento do consumidor de um e-commerce brasileiro. Tese de mestrado. Universidade Federal do Paraná. Curitiba.

VIEIRA, P. C. S. (2004). Geração de superfícies de interação pelo método da regressão linear múltipla com o modelo de dano em vigas de timoshenko 3d. Tese de doutorado em estruturas e construção civil publicação e.td - 006a/04. UNB. Brasília.

APÊNDICE I - Resultados da regressão do modelo em função das variáveis explicativas, estimativas dos parâmetros, análise de variância e análise de resíduos.

Coeficientes:	Estimativa	Desvio Padrão	t value	Pr(>| t |)
(Intercept)	-5.319.424	634.888	-8.379	4.51e-10 ***
Liquidez Corrente	347.678	326.090	1.066	0.2932
Liquidez Operacional	-108.609	54.556	-1.991	0.0539 .
Liquidez Seca	-522.747	359.230	-1.455	0.1541
Liquidez Imediata	162.751	119.036	1.367	0.1798
Liquidez Geral	4319.923	453.911	9.517	1.74e-11 ***
ROE	215.313	27.164	7.927	1.72e-09 ***
ROA	-32.531.472	6.919.774	-4.701	3.54e-05 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. Erro padrão residual: 99.840 em 37 graus de liberdade. R-quadrado múltiplo: 0,9205 R-quadrado ajustado: 0,9055. Estatística F: 124.9 em 7 e 37 graus de liberdade, p-value: < 2.2e-16.				

Fonte: Elaboração própria com software R a partir de Sassi, Perez, Myazato, Ye, Silva e Louzada (2011).

APÊNDICE II - Comparação entre o lucro líquido previsto e o real. [4]

Lucro líquido previsto	Lucro líquido real	Diferença Previsão x Real
546.930,8	546.785,5	145,30
691.785,6	635.942,5	55.843,10
640.107,6	650.403,2	-10.295,60
633.282,8	664.864	-31.581,20
616.456,5	709.974	-93.517,50
1.013.233,6	845.304	167929,60
1.033.766,3	962.316	71.450,30
1.088.732,5	1.013.189,5	75.543
1.156.900,5	1.165.810	-8.909,50
954.229,5	1.156.315,5	-202.086
1.243.130,8	1.293.346,5	-50.215,70
1.390.853,4	1.416.224,8	-25.371,40
1.640.326,7	1.560.378,2	79.948,50
2.059.287,8	2.014.206	45.081,80

Fonte: Elaborada pelo autor.

[4] (Em milhares de reais)