Proposta Monografia em Sistemas de Informação II

Vinícius Assis Neves[1]
Orientador: Cristiano Arbex Valle[2]

Departamento de Ciências da Computação - Universidade Federal de Minas Gerais (UFMG)

[1]vassisn2001@gmail.com
[2]arbex@dcc.ufmg.br

Um modelo de Aprendizado de Máquina para a classificação de ativos da bolsa de valores quanto a sua atual subvalorização e potencial de valorização futura

10 de setembro de 2023
Belo Horizonte - MG, Brasil

1. Introdução

Esta Monografia, de forma geral, busca desenvolver um modelo de Aprendizado de Máquina eficiente para que, alimentado com dados de diferentes demonstrativos financeiros e valores de mercado de empresas da bolsa de valores americana, seja capaz de identificar de forma eficiente ativos que possuem um valor atual abaixo de um valor abstrato justo. Diz-se abstrato pois é impossível calcular um valor realmente justo para um ativo, já que isso seria uma análise muito subjetiva e irrealista, pois um valor realmente justo de uma empresa refletiria diversos fatores do contexto socioeconômico e, principalmente, dados de performance futuros da entidade, ou seja, um valor totalmente conceitual por ser fisicamente impossível calculá-lo.

O papel do algoritmo de machine learning a ser desenvolvido é o de aproximar esse valor justo teórico da melhor forma possível, e isso será feito partindo do pressuposto que, em média, o mercado é precificado de forma justa, ou seja, se o modelo for capaz de entender como os indicadores financeiros das empresas impactam no seu valor de mercado, ele será capaz de estimar um valor justo aproximado para todas as empresas.

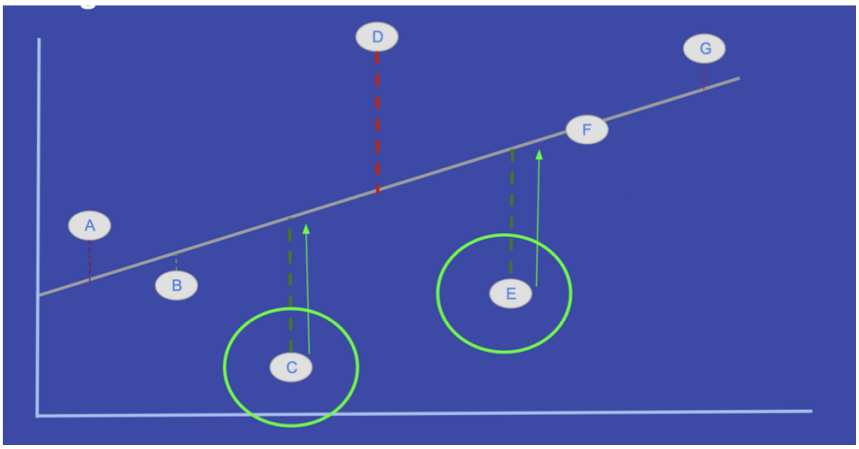

Figura 1 - Gráfico Ilustrativo resíduos do modelo

A figura 1 contém um gráfico que ilustra como os ativos subvalorizados seriam escolhidos. Este é apenas um gráfico demonstrativo, já que nele existe apenas uma variável, no eixo x. No experimento real, várias variáveis (contas

dos demonstrativos) são utilizadas. Seja a linha cinza a reta que demonstra o valor que assume ser o justo para o ativo; e seja os círculos onde cada ativo se encontra no momento estudado; caso o círculo, que representa um ativo, esteja abaixo da reta, significa que seu valor real está abaixo do que o modelo assume ser o "justo", logo, este ativo pode possuir um potencial grande de valorização.

1.1. Trabalho já realizado (MSI I)

Na primeira parte deste projeto (Monografia em Sistemas de Informação I), o objetivo traçado foi exatamente o mesmo: criar um modelo capaz de identificar ativos subvalorizados e utilizar deste para criar portfólios de ativos com alto potencial de lucratividade. Nessa primeira parte do projeto, apenas experimentos introdutórios foram realizados, e os dados utilizados foram diferentes dos que provavelmente serão utilizados durante todo o desenvolvimento da MSI II.

Na primeira parte, a principal fonte de dados utilizada foi a página "Financial Statements Data Sets", da U.S. Securities and Exchange Commision [1], para obtenção de datasets de demonstrativos financeiros. Também foram utilizados dados do Yahoo Finance [2], para histórico de preços e informações que foram usadas para o cálculo do valor de mercado histórico das empresas analisadas. Os dados de demonstrativos financeiros compreendiam o período de 2009 a 2022, contendo demonstrativos trimestrais, e estavam em formato XBRL, o que fez necessário um processamento de dados substancial para organização desses em um formato mais fácil de utilizar no modelo de aprendizado de máquina.

Uma grande limitação dos dados foi que eram poucas as contas dos demonstrativos financeiros que eram comuns entre várias empresas. Por exemplo, para o final de 2015, tinha-se: 4492 Balance Sheets, 4490 Income Statements e 4463 Cash Flow Statements. Selecionando as 6 colunas mais frequentes de cada demonstrativo, foram selecionadas apenas 299 empresas que possuíam um valor não nulo para todas as 18 colunas.

Nessa parte do projeto, porém, o banco de dados utilizado provavelmente será o Compustat da WRDS (Wharton Research Data Services), a análise exploratória ainda dirá quantas empresas serão selecionadas se o mesmo critério de seleção for seguido, porém o Compustat possui dados de demonstrativos financeiros de um período consideravelmente maior (dados a

partir de 1950 são encontrados), e possivelmente será uma fonte de dados mais adequada para este projeto.

1.1.1. Resultado Anterior

Poucos experimentos foram realizados (por não ser o objetivo da MSI I), porém, um experimento realizado com os demonstrativos do final de 2016, com 197 empresas selecionadas pelo critério explicado, apresentou relativamente bons resultados.

Para realizar tal experimento, foi empregada uma estratégia de trading (explicada na seção Metodologia), onde basicamente aplica-se capital em um conjunto de ativos apenas uma vez, na data inicial do experimento. Os ativos escolhidos são os considerados abaixo do valor justo pelos diferentes modelos utilizados. Foram esses:

1. Regressão Linear: Utilizado o algoritmo "LinearRegressor" da biblioteca sckit-learn (sklearn) com os hiperparâmetros padrões.
2. Regressão Multi-Layer Perceptron : Utilizado o algoritmo "MLPRegressor" também da biblioteca sckit-learn (sklearn) com os hiperparâmetros padrões.

Quatro períodos foram testados, de 2017 a 2018, 2017 a 2019, 2017 a 2020 e 2017 a 2021.

Os resultados podem ser visualizados na Figura 2.

	Modelo Reg. Linear	Modelo Regressão Rede Neural	Desempenho do mercado no período (S&P 500)
01/2017 - 01/2018:	18%	320%	20%
01/2017 - 01/2019:	48%	129%	17%
01/2017 - 01/2020:	89%	136%	42%
01/2017 - 01/2021:	17%	25%	65%

Figura 2 - Resultados da estratégia de trading empregada

É importante notar que os resultados acima, apesar de muito interessantes, não representam o projeto como um todo por serem limitados e não terem sido consistentes com outros experimentos realizados até então. O

intuito da MSI II é explorar estratégias diferentes e alcançar resultados bons e consistentes.

2. Referencial Teórico

São comuns os estudos em Ciências de Dados que utilizam algoritmos de aprendizado de máquina para prever e buscar um retorno positivo do investimento em ativos, por meio de estratégias relacionadas com a análise técnica, e com o mesmo objetivo, operações frequentes e realização de lucro no curto prazo. Estudos que buscam explorar o potencial da Ciência de Dados para realizar predições e indicações de investimentos com metodologias da análise fundamentalista já não são tão frequentes, possivelmente pelo desafio de verificar resultados, já que o lucro apenas se realiza depois de alguns anos.

Previsão do mercado financeiro, usando principalmente dados temporais, como séries de preços e volumes de negociação, tem sido uma técnica explorada por pesquisadores buscando realizar lucros na bolsa de valores. Como este não é o objetivo deste trabalho, as pesquisas listadas serão tais que exploram tópicos relacionados, como a análise de demonstrativos financeiros ou de índices fundamentalistas.

O artigo [3] apresenta uma Rede Probabilística Artificial que leva em conta dados históricos e índices fundamentalistas para realizar classificações de tendências do mercado.

A referência [4] segue uma abordagem mais parecida com a que será explorada neste estudo, utilizando-se de aprendizado de máquina para fazer a análise de demonstrativos financeiros e prefer a magnitude e o sinal do retornos significativos de ativos. O estudo atingiu uma acurácia entre 53% e 59% para a previsão da magnitude e sinal dos retornos. Apesar de não ser o principal objetivo desse, também analisou o retorno de investimentos baseados nos ativos que o algoritmo definiu que teriam retornos extraordinários. Seus experimentos resultaram em retornos que decaíram conforme o período analisado se aproximava da atualidade. Na simulação realizada de 1991-2000, com base em dados disponíveis no ano anterior, o portfólio conseguiu um retorno de aproximadamente 28% a.a, enquanto na simulação mais atual, 2010 - 2019, o retorno foi de aproximadamente 8% a.a. A hipótese levantada é que o

mercado foi se otimizando no sentido de aprender a precificar melhor os ativos e evitar ações extremamente desvalorizadas.

3. Metodologia

Como já abordado na introdução, a metodologia utilizada na primeira parte do projeto é parecida com a que será desenvolvida agora. Um modelo para aproximar o valor justo de empresas será desenvolvido (melhorado a partir do que foi anteriormente trabalhado), com base nesse modelo empresas consideradas subvalorizadas serão selecionadas e utilizadas para montagem de portfólios de trading com intuito de lucro no médio/longo prazo. Os passos a serem seguidos para os principais experimentos a serem realizados, utilizado diferentes algoritmos de aprendizados de máquina, são:

1. Definição de um período de tempo analisado
2. Seleção de todas as empresas com demonstrativos disponíveis do exercício anterior ao começo de período, e que ainda estejam listadas no final do período.
3. Exclusão de empresas que não possuem features suficientes
    a. Ao contrário do trabalho realizado até agora, busca-se completar com outros algoritmos de aprendizado de máquina alguns valores faltantes para otimização do conjunto de dados realizados.
4. Treino do modelo para aproximar o valor de mercado justo dos ativos no início do período.
5. Definição de um threshold relativo como um filtro de quais empresas são mais prováveis de estarem desvalorizadas. Exemplo, se um threshold de 500% for selecionado, apenas empresas em que o valor justo estimado for maior que 500% do valor real serão selecionadas para a carteira.
6. Computação da variação de mercado de cada empresa selecionada no período analisado.
7. Simulação de quanto uma carteira de investimentos igualmente distribuída entre as empresas selecionadas teria lucrado caso o investimento hovesse sido feito no início do período.
8. Comparação do resultado do item anterior com a variação do mercado no período.

Estratégias adicionais que podem ser utilizadas nessa fase do projeto são: Usar regressões para preencher dados faltantes e aumentar o conjunto de dados; Combinar demonstrativos de diferentes anos utilizando esses como variáveis categóricas; Abordagem com séries temporais; Utilizar o setor das empresas como feature; Selecionar uma empresa para o portfólio apenas se foi considerada subvalorizada por períodos consecutivos; Selecionar uma empresa para o portfólio apenas se foi considerada subvalorizada por períodos consecutivos e se possui atual tendência de alta;

O intuito dessa fase do projeto é alcançar resultados, portanto, as estratégias acima são apenas algumas das que poderão ser exploradas para alcançar os resultados esperados.

4. Resultados esperados

Sendo o objetivo do projeto encontrar ações subvalorizadas que tendem a se valorizar no futuro, nenhum esforço é empenhado para prever ou simular variações de mercado, já que tais mudanças dependem principalmente de fatores externos e futuros, tais com que os dados de um período anterior aos experimentos realizados tendem a não possuir nenhuma correlação. Sendo assim, considera-se como resultado esperado que os experimentos realizados superem a variação geral do mercado, ou seja, se o mercado, de forma geral, se valorizou em x% no período analisado, espera-se que o portfólio construído no experimento tenha uma valorização de y%, tal que $y > x$. Dessa forma, pode-se isolar, até certo ponto, se a empresa analisada realmente teve uma deslocação de mercado em direção ao seu preço justo por motivos de que estava subvalorizada.

Espera-se também, que com a realização de vários experimentos, os quais explorem períodos diferentes, durações diferentes, thresholds diferentes, entre outros, note-se que a seleção realizada constantemente supera o mercado, de forma a fazer a ferramenta eficiente para a construção de um portfólio com bom desempenho, com chances de realização de lucros futuros.

5. Cronograma

- **15/09 - 29/09 -** Análise exploratória do banco de dados Compustat
- **30/09 - 16/10 -** Organização dos dados e primeiros experimentos com o conjunto de dados utilizado
- **16/10 -** Apresentação Parcial - Pitch
- **17/10 - 30/10 -** Experimentos mais detalhados e esforço para completar features faltantes e aumentar o conjunto de dados
- **31/10 - 26/11 -** Exploração de técnicas e estratégias diferentes com objetivo de alcançar melhores resultados; Conclusão do projeto
- **27/11 -** Apresentação Final - Pitch
- **28/11 - 04/12 -** Elaboração do Relatório Final
- **04/12 -** Entrega do Relatório final

Referências

[1] U.S. Securities and Exchange Commision, Financial Statement Data Sets. Disponível em https://www.sec.gov/dera/data/financial-statement-data-sets.

[2] Yahoo Finance. Disponível em "https://finance.yahoo.com/".

[3] S. H. Kim nd S. H. Chun, ‘‘Graded forecasting using an array of bipolar predictions: Application of probabilistic neural networks to a stock market index,’’ *Int. J. Forecasting*, vol. 14, no. 3, pp. 323–337, Sep. 1998.

[4] Amel-Zadeh, Amir and Calliess, Jan-Peter and Kaiser, Daniel and Roberts, Stephen, Machine Learning-Based Financial Statement Analysis (November 25, 2020). Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3520684