Agrupamento de Empresas da B3 utilizando k-means

Pedro Arthur Ramos Oliveira ITA

[1]

1 Introdução

A bolsa de valores tem atraído um número crescente de investidores no Brasil nos últimos anos. Esse aumento de investidores no mercado acionário tem gerado dúvidas frequentes sobre como investir e em quais empresas negociar ações. É comum que investidores que atuam por conta própria acabem enfrentando prejuízos, especialmente se não realizarem um estudo adequado das ações em que pretendem investir. A mineração de dados tem sido cada vez mais utilizada para auxiliar na tomada de decisões por parte dos investidores que desejam apostar em ações, considerando as mais diversas condições do mercado. Muitos acionistas confiam nos resultados derivados da mineração de dados para guiar suas negociações na bolsa, tanto em investimentos de curto quanto de longo prazo. Este projeto busca apresentar uma forma de identificar empresas cujas ações sejam atrativas para investimentos de longo prazo, ou seja, que ofereçam boas perspectivas de retorno ao longo dos anos. O objetivo principal é identificar boas ações para este tipo de investimento.

De acordo com os consultores da ToroInvestimentos.com, um dos maiores obstáculos que impedem as pessoas de entrarem nesse mercado é a dificuldade de interpretar as ações e compreender o funcionamento do próprio negócio. Diante das inúmeras variáveis que podem afetar a qualidade de um investimento, muitos investidores acabam enfrentando dificuldades que resultam em perdas, especialmente aqueles que não têm um conhecimento sólido sobre o mercado acionário. Este trabalho visa auxiliar investidores na identificação de ações promissoras para investimentos de longo prazo, com o objetivo de se tornarem sócios das empresas e colherem os lucros gerados ao longo do tempo. Isso proporciona maior segurança e tranquilidade para investidores menos experientes. Para a execução deste trabalho, foi necessário estudar o funcionamento desse modelo de negócios, assim como as mudanças que podem ocorrer no mercado e os fatores que as influenciam. Por ser um ambiente repleto de particularidades, é essencial um entendimento aprofundado das variáveis envolvidas. A extração de informações confiáveis também é crucial para que essas variáveis possam ser posteriormente utilizadas no processo de mineração de dados. A próxima etapa envolve a montagem de um banco de dados, que será usado em testes e para a obtenção de resultados que atendam aos objetivos do projeto.

2 Revisão Bibliográfica

A mineração de dados tem sido amplamente utilizada para diversas finalidades, auxiliando na previsão e extração de informações importantes para a tomada de decisões em diferentes áreas, como medicina, marketing, telecomunicações, educação e finanças (GOLDSCHMIDT; BEZERRA, 2016). No contexto financeiro, sua aplicação pode trazer benefícios significativos, permitindo identificar variáveis que influenciam os resultados de uma empresa ou ação, tanto de forma positiva quanto negativa (GOLDSCHMIDT; BEZERRA, 2016).

No mercado financeiro, o uso de técnicas de mineração de dados possibilita que profissionais capacitados obtenham informações valiosas, permitindo a antecipação na tomada de decisões e, consequentemente, uma vantagem competitiva (VELOSO; MOREIRA; SILVA; SILVA, 2011). Além disso, a mineração de dados facilita a análise de grandes volumes de informação, que são comuns em mercados como o de ações, onde muitas variáveis podem interferir diretamente nos resultados (VELOSO; MOREIRA; SILVA; SILVA, 2011). Ao aplicar técnicas de mineração de dados no mercado acionário, é possível identificar informações que orientam a tomada de decisão dos investidores. Através do estudo de relatórios gerados por esses métodos, torna-se viável comparar dados históricos de empresas e avaliar o desempenho financeiro dessas no mercado (CORTES; PORCARO; LIFSCHITS, 2002). Isso permite ao investidor, especialmente aquele interessado em investimentos a longo prazo, tomar decisões mais seguras e fundamentadas.

3 Técnicas de Mineração de Dados

Nesta seção, abordam-se algumas técnicas de mineração de dados comumente utilizadas.

3.1 Classificação

O método de classificação consiste em examinar um conjunto de dados e atribuir a cada um deles uma classe previamente definida, sendo amplamente utilizado em diversos setores (CORTES; PORCARO; LIFSCHITZ, 2002).

3.2 Agrupamento

Este método tem como objetivo dividir um conjunto de dados em grupos que sejam homogêneos internamente. A ideia é criar grupos em que os dados dentro de cada grupo sejam mais semelhantes entre si e diferentes dos dados de outros grupos. A formação desses grupos se distingue por não contar com classes predefinidas para categorizar os dados analisados. Os grupos são estabelecidos com base na similaridade dos dados, agrupando-os conforme variáveis comuns. Assim, os dados são separados em dois ou mais grupos, dependendo das características do banco de dados analisado (CORTES; PORCARO; LIFSCHITZ, 2002).

4 Objetivos

4.1 Objetivo Geral

Desenvolver e avaliar uma solução de mineração de dados com o objetivo de selecionar as melhores empresas para investimento a longo prazo no mercado de ações.

4.2 Objetivos Específicos

- Compreender os indicadores da análise fundamentalista para a construção da base de dados.
- Criar uma base de conhecimento que inclua todas as variáveis necessárias para a análise.
- Avaliar e testar diferentes algoritmos de classificação e métodos estatísticos.
- Executar os algoritmos de classificação e os métodos estatísticos mais relevantes, coletando os resultados para verificar a eficácia da abordagem proposta na classificação dos perfis das empresas.

5 Metodologia

5.1 Análise Fundamentalista

A análise fundamentalista tem como objetivo avaliar a saúde financeira de uma empresa, projetando o valor ou a significância dos seus parâmetros com o intuito de ajudar na tomada de decisões relacionadas ao valor de suas ações. Esta análise leva em consideração fatores macro e microeconômicos que influenciam o desempenho das empresas no mercado de ações, permitindo uma avaliação minuciosa que pode prever resultados a longo prazo, geralmente entre cinco e dez anos (EXAME, 2013). Este processo envolve tanto a análise de fatores quantitativos, como inflação, taxas de juros, valores de caixa e câmbio, quanto qualitativos, como a gerência, controladores e composição do conselho administrativo (EXAME, 2013). Com base nessas informações, a análise fundamentalista auxilia o investidor na definição do valor de suas ações, permitindo uma gestão mais segura de sua carteira.

5.2 Variáveis Extraídas para Avaliação

Para compreender o valor de uma empresa, são analisadas diversas variáveis financeiras. Todas essas informações foram extraídas do site Bastter.com, uma plataforma que oferece dados sobre empresas listadas na BM&FBOVESPA (BASTTER, 2017).

Entre as variáveis mais relevantes para a análise estão:

- **Valor de Mercado**: Lucro Líquido por Ação
- **EV**: Enterprise Value = Valor de mercado + Dívida bruta – Caixa.
- **Pessoas Fisicas**: Total de acionistas das ações da empresa.
- **LPA Descontado**: Lucro líquido (descontado dos não recorrentes) por ação.
- **VPA**: Valor Patrimonial por ação.
- **P/L descontado**:Preço da ação dividido pelo lucro líquido (descontado dos não recorrentes) por ação.
- **Margem segurança**: Diferença positiva entre o potencial de ganho na ação e a taxa de juros praticada pelo mercado ou a perspectiva de ganho investido em títulos do governo.

- **Luc. Liquido**: A partir da receita líquida se diminui os custos e as despesas das vendas para se chegar ao lucro líquido.
- **Divida Bruta/PL**: É a dívida bruta dividida pelo patrimônio líquido. Uma das formas de avaliar o endividamento de uma empresa.
- **Divida Liquida**: Dívida Líquida é igual a dívida bruta – caixa. Demonstra a dívida da empresa diminuindo o caixa.
- **EM**:Significa "Equity Multiplier" e é igual a (Ativo total / Patrimônio líquido). Quanto maior, mais alavancada está a empresa. Demonstra quantos reais a empresa está operando para cada real do dinheiro do acionista.

5.3 Classificação das Empresas

Para classificar as empresas em boas ou ruins, o estudo considera a evolução dos lucros e retornos aos acionistas. Empresas boas apresentam crescimento contínuo, e as empresas ruins mostram declínio nos resultados ou consistência em valores baixos.

5.4 Coleta e Carregamento dos Dados

Os dados financeiros das empresas foram carregados a partir de um arquivo Excel utilizando a biblioteca `pandas` Os dados foram retirados da plataforma Bastter. A utilização de um `DataFrame` facilita a manipulação e preparação dos dados para análise.

5.5 Definição dos Pesos para as Variáveis

Foram atribuídos pesos a cada métrica financeira com base em sua relevância para a análise. Esses pesos foram usados para ponderar as variáveis no cálculo do score de cada empresa, de forma a refletir diretamente na classificação final.

5.6 Pré-processamento dos Dados

- Substituição de valores ausentes nas variáveis numéricas pela média de cada coluna.
- Aplicação de transformações logarítmicas para suavizar a variabilidade das variáveis.
- Ajuste dos valores para garantir que todas as variáveis sejam positivas, facilitando as transformações.
- Normalização das variáveis numéricas usando `MinMaxScaler`, garantindo que estejam na mesma escala.

5.7 Salvamento dos Dados

Utilizou-se a biblioteca `pickle` para salvar os dados pré-processados, incluindo o `DataFrame`, os dados de entrada (`X`), os rótulos (`y`) e o objeto `scaler`. Isso permite que a análise seja retomada sem a necessidade de reprocessar os dados.

5.8 Otimização de Hiperparâmetros

A otimização dos hiperparâmetros foi realizada com `GridSearchCV`, buscando a melhor combinação de parâmetros para o `RandomForestCl`

- Parâmetros avaliados: `criterion` (`gini/entropy`), `min_samples_split`, e `n_estimators`.
- A combinação com melhor desempenho foi utilizada para treinar o modelo final.

5.9 Treinamento e Avaliação do Modelo

Um modelo `RandomForestClassifier` foi treinado usando os dados de entrada (`X`) e rótulos (`y`). A performance do modelo foi avaliada através de técnicas de validação cruzada, calculando-se a acurácia, matriz de confusão e outras métricas de desempenho.

5.10 Agrupamento com K-Means

Após a classificação, aplicou-se o algoritmo de agrupamento `KMeans` para segmentar as empresas em grupos com características semelhantes:

- Determinação do número ótimo de clusters utilizando o método do cotovelo, baseado na métrica *Within-Cluster Sum of Squares* (WCSS).
- Ajuste do `KMeans` com 6 clusters, atribuindo rótulos a cada empresa de acordo com o cluster ao qual pertence.

5.11 Análise dos Centróides dos Clusters

Analisamos as características médias de cada cluster utilizando os centróides do `KMeans`:

- Reversão da normalização dos centróides para que os valores sejam interpretáveis no contexto original das variáveis.
- Criação de um `DataFrame` para facilitar a análise do perfil médio de cada grupo de empresas.

5.12 Integração dos Clusters ao Dataset Original

Os rótulos dos clusters foram combinados com o `dataset` original, permitindo uma análise detalhada das empresas dentro de cada grupo. Essa integração facilita a análise de padrões comuns entre empresas do mesmo segmento e cluster.

5.13 Redução de Dimensionalidade com PCA

Para visualizar os clusters em um espaço bidimensional, aplicou-se a Análise de Componentes Principais (PCA):

- A redução de dimensionalidade foi realizada para 2 componentes principais, que retêm a maior parte da variabilidade dos dados.
- A redução facilita a visualização dos clusters em um gráfico 2D.

5.14 Visualização dos Clusters

Gráficos de dispersão foram gerados utilizando `seaborn`, onde cada ponto representa uma empresa no espaço dos componentes principais, e os clusters são diferenciados por cores. Isso permite uma visualização clara da separação entre os grupos e sua proximidade no espaço reduzido.

Appendix

Appendix content

Referências

Goldschmidt; Bezerra. (2016). Exemplos de aplicações de data mining no mercado brasileiro. Disponível em: http://computerworld.com.br/exemplos-de-aplicacoes-dedata-mining-no-mercado-brasileiro

Veloso; Moreira; Silva; Silva. (2011). Data mining, seus benefícios, utilizações, metodologia, campo de atuação dentro de grandes e pequenas empresas. Disponível em: http://periodicos.unifacef.com.br/index.php/resiget/article/download/154/12

CORTES, Sergio da Costa; PORCARO, Rosa Maria; LIFSCHITZ, Sergio. Mineração de Dados – Funcionalidades, Técnicas e Abordagens. 2002. Monografia apresentada Universidade PUC-Rio para a obtenção de Doutorado em Ciência da Informação.

GRANATYR, Jones. 7 Técnicas de inteligência artificial para profissionais de TI ganharem mais dinheiro. 2017. (Doutor em Ciência da Computação).

TORORADAR. (2017) O que é análise fundamentalista. Disponível em: https://www.tororadar.com.br/investimento/analise-fundamentalista/o-que-e

EXAME. (2013) Como funciona a análise fundamentalista de ações. Disponível em: https://exame.abril.com.br/seu-dinheiro/como-funciona-analise-fundamentalistaacoes-576374/

FRANCISCON, Eduardo Alexandre. Análise de empresas para investidores a longo prazo como sócio dentro da BM&FBOVESPA utilizando mineração de dados. 2017. Trabalho de Conclusão de Curso (Graduação em Sistemas de Informação) – União de Ensino do Sudoeste do Paraná, 2017.