

UNIVERSIDADE FEDERAL DE MINAS GERAIS
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
MINERAÇÃO DE DADOS (DCC-057)

**MINERAÇÃO DE PADRÕES FREQUENTES EM
DADOS FINANCEIRO INTRADAY: UMA ANÁLISE
ABRANGENTE DE PREÇOS DE AÇÕES, PARES DE
MOEDA E CRIPTOMOEDAS**

Jorge Augusto de Lima e Silva
(2021032005)
Vitor Emanuel Ferreira Vital
(2021032072)

Belo Horizonte, Outubro de 2023

RESUMO

No presente relatório apresentamos o desenvolvimento da primeira parte do primeiro trabalho da disciplina Mineração de Dados. Iremos documentar aqui as interações feitas com as inteligências artificiais ChatGPT e Bard para o desenvolvimento do trabalho, avaliando seu comportamento quando apresentado a problemas com um grau maior de complexidade.

Alunos:

Jorge Augusto de Lima e Silva
Vitor Emanuel Ferreira Vital

Orientador:

Prof. Wagner Meira Jr.
Departamento de Ciência da Computação (UFMG)

Palavras-chave: padrões, frequentes, ativos, financeiro, dados

Área do conhecimento: Ciência da Computação (Ciência de Dados)

ABSTRACT

In this report, we present the development of the first part of the first assignment for the Data Mining course. We will document here the interactions conducted with the artificial intelligences ChatGPT and Bard for the project's development, evaluating their behavior when exposed to problems with a higher degree of complexity.

Keywords: patterns, frequent, stocks, financial, data

SUMÁRIO

[RESUMO](#)

[ABSTRACT](#)

[SUMÁRIO](#)

[1. INTRODUÇÃO](#)

[2. ENTENDIMENTO DO NEGÓCIO](#)

[2.1. Background](#)

[2.2. Objetivos do Negócio](#)

[2.3. Critérios de Sucesso](#)

[2.4. Pressupostos](#)

[3. ENTENDIMENTO DOS DADOS](#)

[3.1. Oanda](#)

[3.2. HistData](#)

[3.3. Cripto](#)

[3.4. Análise exploratória](#)

[4. PREPARAÇÃO DOS DADOS](#)

[4.1. Discretização dos Dados](#)

[4.2. Remoção de Escala](#)

[5. MODELAGEM](#)

[6. VALIDAÇÃO](#)

[CONCLUSÕES](#)

[Apêndice A](#)

[REFERÊNCIAS](#)

1. INTRODUÇÃO

Este projeto de mineração de padrões frequentes tem como objetivo analisar e extrair padrões comuns de dados de mercado financeiro, permitindo assim um melhor entendimento do comportamento de diversos tipos de ativos, como ações, moedas e criptomoedas, com o objetivo final de observar e documentar certas tendências que podem ser observadas no mercado.

Ao analisar e extrair padrões frequentes do comportamento de ativos podemos ganhar insights valiosos sobre o comportamento do mercado financeiro de uma forma geral. O conhecimento destas tendências pode ajudar diferentes públicos a tomar decisões mais conscientes, seja ao auxiliar um investidor a identificar uma oportunidade de investimento quanto em permitir que um executivo desenvolva estratégias de uma maneira mais informada.

O objetivo do projeto é analisar um conjunto de sequências que representam os valores diários de ativos financeiros e identificar comportamentos recorrentes ou padrões comuns entre esses ativos ao longo do dia.

O principal desafio que podemos vir a enfrentar o reconhecimento de padrões espúrios, ou seja, padrões pequenos demais e que não apresentam nenhum valor semântico no contexto em questão ou padrões que podem ser gerados facilmente se supormos que os dados seguem alguma determinada distribuição de probabilidade, ou seja, dados aleatórios.

O que será desenvolvido nesta primeira parte do projeto é a solução feita baseada nas ideias das inteligências artificiais generativas ChatGPT, da OpenAI, e Bard, do Google. Além disso, analisaremos os erros encontrados durante a modelagem e as estratégias que encontramos para induzir a IA a contornar o problema, quando possível.

2. ENTENDIMENTO DO NEGÓCIO

2.1. Background

Perguntando ao ChatGPT sobre a natureza da fonte dos dados, foi retornada a seguinte resposta: O mercado financeiro é um ambiente onde ativos financeiros são comprados e vendidos, e os preços desses ativos são determinados pela interação entre oferta e demanda. Existem diversos tipos de ativos financeiros, e muitos deles são acompanhados por indicadores que ajudam os investidores e analistas a entender o desempenho e as tendências do mercado.

E quando perguntado sobre a natureza dos ativos presentes na base de dados, a seguinte resposta foi fornecida: Indicadores financeiros são valores numéricos que representam aspectos específicos do desempenho financeiro de

empresas, mercados ou economias. Eles são usados para avaliar a saúde financeira, tendências de mercado e outros aspectos importantes da atividade econômica. Indicadores financeiros são informações cruciais para investidores, analistas, empresas e governos. Moedas Fiat são moedas emitidas e respaldadas por governos, como o dólar dos Estados Unidos (USD) e o euro (EUR). Elas têm valor legal em seus respectivos países e são usadas para transações internacionais. Criptomoedas são ativos digitais descentralizados que usam criptografia para garantir transações e controlar a criação de novas unidades. O Bitcoin (BTC) é uma das criptomoedas mais conhecidas.

E quando questionado sobre a natureza das negociações de ativos financeiros, a resposta foi: A negociação de ativos financeiros envolve a compra e venda de instrumentos financeiros em mercados regulados ou plataformas eletrônicas. Os participantes incluem investidores individuais, instituições financeiras, traders profissionais, empresas e governos. As negociações ocorrem em bolsas de valores físicas ou plataformas eletrônicas. Os participantes emitem ordens de compra ou venda, registradas em um livro de ordens, e as negociações são correspondidas com base no preço e na hora de envio. O preço de mercado é determinado pela última ordem correspondida. A liquidez varia, afetando a facilidade de negociação. Os mercados têm horários específicos de negociação, e notícias e eventos podem impactar os preços. Custos de negociação incluem comissões, spreads e taxas de câmbio. Estratégias de negociação como análise técnica, análise fundamental, arbitragem e day trading são utilizadas pelos participantes. Compreender esses aspectos é fundamental antes de entrar no mercado financeiro.

2.2. Objetivos do Negócio

O ChatGPT já não apresentou resultados tão bons quanto as anteriores, sendo bem genéricas, mas ainda assim interessantes:

- **Identificação de Comportamentos Recorrentes:** este objetivo é fundamental, pois permite entender melhor o comportamento histórico dos ativos financeiros. A identificação de padrões frequentes pode ajudar na interpretação das tendências passadas e na compreensão das dinâmicas do mercado.
- **Deteção de Anomalias:** a detecção de anomalias é igualmente importante, pois ajuda a identificar eventos ou situações incomuns nos dados. Anomalias podem ser indicativas de eventos inesperados que podem afetar o mercado financeiro e, portanto, merecem atenção especial.
- **Predição de Tendências:** a capacidade de usar padrões passados para prever tendências futuras é um objetivo valioso, especialmente para investidores e traders. Essa previsão pode auxiliar na tomada de decisões informadas sobre compra, venda ou retenção de ativos financeiros. Esses três objetivos são cruciais para a análise de dados financeiros, pois proporcionam insights sobre

o passado, o presente e o potencial futuro do mercado financeiro, contribuindo para decisões mais informadas e uma melhor compreensão do comportamento dos ativos.

2.3. Critérios de Sucesso

Os critérios de sucesso que foram associados à esses objetivos são os seguintes:

- **Taxa de Identificação de Padrões:** Avalie a capacidade do modelo em identificar tanto padrões recorrentes quanto anomalias. Uma taxa alta de identificação de padrões relevantes é um indicador de sucesso, pois abrange a identificação de comportamentos recorrentes e a detecção de anomalias.
- **Relevância Financeira:** Avalie se os padrões identificados têm relevância financeira, ou seja, se estão relacionados a eventos financeiros importantes ou tendências que afetam o mercado. Isso inclui a interpretação qualitativa dos resultados.
- **Acurácia das Previsões:** Avalie a precisão das previsões baseadas nos padrões identificados. Uma alta precisão nas previsões indica o sucesso em prever tendências futuras.

Podemos ver agora que o ChatGPT começa a tentar fugir do escopo, mostrando uma tendência ao uso de algoritmos de aprendizado de máquina supervisionados, que são diferentes dos algoritmos usados em reconhecimento de padrões frequentes convencionais, i.e., Apriori, FPGrowth, etc.

Porém, mesmo tomando uma abordagem diferente, ainda temos muito potencial de encontrar resultados interessantes, e ainda é possível nas etapas de preparação dos dados e modelagem conduzi-lo à seguir um caminho mais alinhado com a proposta do projeto.

2.4. Pressupostos

Os pressupostos mais gerais a serem considerados ao realizar a análise de reconhecimento de subcadeias frequentes em dados financeiros incluem:

- **Qualidade dos Dados:** os dados financeiros são precisos, completos e confiáveis.
- **Estabilidade do Mercado:** o mercado financeiro opera em condições relativamente estáveis durante o período de análise.
- **Comportamento de Anomalias:** anomalias no mercado financeiro podem ocorrer devido a eventos inesperados, e a capacidade de detecção de anomalias depende da capacidade do modelo em distinguir entre mudanças anormais e flutuações normais.
- **Relevância dos Critérios de Sucesso:** os critérios de sucesso definidos são relevantes para avaliar o desempenho dos objetivos e refletem as metas desejadas para a análise.

Esses pressupostos gerais fornecem uma base importante para a condução da análise de dados financeiros, mas é fundamental reconhecer que eles podem ser afetados por mudanças nas condições do mercado e nos dados ao longo do projeto. Portanto, a monitorização contínua e a adaptação são necessárias para lidar com qualquer variação que possa surgir.

3. ENTENDIMENTO DOS DADOS

Para este projeto, tomou-se como base três conjuntos de dados financeiros dentre os quais se incluem ações, pares de moedas e criptomoedas. Todas as informações foram encontradas no repositório <https://github.com/FutureSharks/financial-data>.

Os dados utilizados são fornecidos em intervalos de tempo que variam de 1 minuto a uma semana, entretanto, para eliminar os impactos de dados ausentes foi definido um intervalo de tempo de 15 minutos, isso porque é um tempo no qual variações nos valores dos ativos não são tão acentuados, em sua maioria, e nos permite garantir que haverá informações presentes para todos os casos. Os arquivos em questão foram unidos, conforme técnicas que serão discutidas posteriormente, resultando em 8.218.011 medições que foram coletadas em diferentes sites.

A fim de facilitar a união dos conjuntos de dados, foram considerados seis atributos relevantes para cada um dos ativos analisados (instrument) (1) **date** data e hora da coleta (2) **close** preço de fechamento da análise realizada (3) **high** preço máximo do dia em questão (4) **low** preço mínimo do dia em questão (5) **open** preço de abertura da análise (6) **Price** preço médio obtido entre o início da análise (open) e o fechamento (close).

Os dados serão usados para desenvolver modelos de previsão de preços de forex, identificar tendências de mercado e avaliar o desempenho de estratégias de negociação.

3.1. Oanda

O conjunto de dados "oanda_data.csv" possui dados intradiários de preços de ações em formato de barra de minuto para backtesting ou análise.

O arquivo de dados comprimido em formato CSV contém informações de mercado de forex (câmbio) da Oanda, uma corretora de forex online. O conjunto utilizado possui 7.337.074 instâncias, sendo os ativos considerados 'AU200_AUD', 'AUD_JPY', 'AUD_USD', 'CORN_USD', 'DE10YB_EUR', 'EUR_JPY', 'EUR_USD', 'FR40_EUR', 'GBP_USD', 'JP225_USD', 'NAS100_USD', 'NATGAS_USD', 'SOYBN_USD', 'SPX500_USD', 'SUGAR_USD', 'UK100_GBP', 'UK10YB_GBP', 'US2000_USD', 'USB02Y_USD', 'USB10Y_USD', 'USD_CAD', 'WHEAT_USD', 'STICO USD', retiradas dos anos de 2005 a 2020.

3.2. HistData

O conjunto de dados "histdata_data.csv" possui dados de mercado históricos de moedas, sendo elas 'ETXEUR', 'GRXEUR', 'JPXJPY', 'SPXUSD'.

O arquivo de dados comprimido em formato CSV contém informações da HistData, uma empresa que fornece dados históricos de mercado para pesquisadores e analistas. A empresa oferece uma ampla gama de dados, incluindo dados de ações, forex, commodities, ETFs e criptomoedas. Os dados utilizados foram coletados de várias fontes, incluindo plataformas de negociação, bancos e corretoras. O conjunto utilizado possui 610.815 instâncias retiradas dos anos de 2010 a 2018.

3.3. Cripto

O conjunto de dados "cryptocurrency_data.csv" foi obtido com base em informações de dois sites BitStamp e Kraken.

O BitStamp é uma exchange de criptomoedas que oferece dados históricos de preços, volumes e taxas de negociação. Em paralelo, temos o Kraken, que é uma corretora de criptomoedas que oferece dados históricos de preços, volumes, taxas de negociação e dados de staking.

O conjunto utilizado possui 270.122 instâncias retiradas dos anos de 2012 a 2018 dentre as quais consideramos as seguintes criptomoedas 'BTC_EUR', 'BTC_USD' e 'ETH_EUR'.

3.4. Análise exploratória

Uma breve análise da correlação dos dados em questão nos permite verificar se os seis atributos date, close, high, low, open, e price estão correlacionados entre si.

Um experimento foi realizado utilizando-se a função pairplot da biblioteca Seaborn, no qual 50 mil amostras foram selecionadas aleatoriamente, a fim de verificar a correlação entre os atributos, para todos atributos analisados obtivemos uma curva linear, indicando correlação igual a um. Isso indica que esses atributos estão fortemente relacionados e tendem a se mover na mesma direção.

Tal conclusão é trivial, visto que cada atributo da amostra tem como base a medição do valor ativo em um intervalo de 15 minutos, sendo cada propriedade um instante da análise ou relação das análises feitas.

Portanto, ao trabalhar com esses dados, é importante estar ciente dessa forte correlação entre os atributos e considerar estratégias adequadas para lidar com ela, como a seleção de variáveis, a transformação de dados ou o uso de técnicas avançadas de modelagem que possam acomodar tal correlação.

4. PREPARAÇÃO DOS DADOS

O processo de preparação dos dados foi feito da seguinte forma. Primeiro, foram considerados que os dados provenientes dos vários provedores seriam tratados da mesma forma. Os dados então foram organizados como séries temporais, em que cada instância que temos é o comportamento de um ativo ao longo de um dia de negociação. Uma vez que os dados foram organizados desta forma, pedimos ao ChatGPT que tratasse duas tarefas de extrema importância neste cenário: a discretização dos dados e a remoção das diferenças de escala presentes nos dados.

4.1. Discretização dos Dados

Quando se trata de discretização de séries temporais para encontrar padrões frequentes usando uma árvore de sufixos e o objetivo é manter o valor semântico dos dados, uma abordagem eficaz é a discretização baseada em percentis. Essa abordagem divide os dados em categorias com base nos percentis dos valores, mantendo a relação relativa entre os valores originais.

Aqui estão os passos para realizar a discretização baseada em percentis em séries temporais:

- Passo 1: Preparação dos Dados: Comece com suas séries temporais de dados financeiros.
- Passo 2: Cálculo dos Percentis: Calcule os percentis dos valores em cada série temporal. Por exemplo, você pode calcular os percentis 25, 50 (mediana) e 75 para cada ponto de dados.
- Passo 3: Definição das Categorias: Categorizamos os dados com base nos percentis calculados, criando categorias como "Baixo", "Médio" e "Alto" para representar diferentes faixas de valores. Por exemplo, valores no 25º percentil ou abaixo são "Baixo", valores entre o 25º e 75º percentil são "Médio" e valores acima do 75º percentil são "Alto". Isso ajuda na análise dos dados.
- Passo 4: Aplicação da Discretização: Atribua a cada ponto de dados a categoria correspondente com base nos percentis calculados e nas categorias definidas.

4.2. Remoção de Escala

O uso da transformação logarítmica é interessante neste caso, pois ajuda a normalizar os dados e facilita a comparação das variações percentuais entre ativos financeiros com diferentes ordens de magnitude nos preços de fechamento.

Nesta transformação porém, fizemos uma alteração da ideia original proposta pelo ChatGPT, ao invés de usarmos

$$\log \left(\frac{close_t}{close_{t-1}} \right)$$

estamos usando

$$\log \left(\frac{close_t}{open_t} \right).$$

5. MODELAGEM

A parte de modelagem do problema foi uma que o ChatGPT apresentou muita dificuldade em responder, constantemente entrando em loop e respondendo formas genéricas e óbvias de modelagem, nunca fornecendo um modelo em específico a ser usado mas informado que passos devem ser seguidos, de forma genérica, para a escolha de um modelo.

O principal meio utilizado para contornar este problema foi solicitar que ele fornecesse algoritmos que poderiam ser usados para atingir os objetivos definidos na seção 2.2. Porém, temos que muitos dos algoritmos propostos por ele fugiam do escopo deste trabalho, utilizando algoritmos que tratam de problemas diferentes da mineração de padrões frequentes, como regressões e classificações. Desta forma, de modo a impedir que o escopo fosse violado, foi perguntado ao ChatGPT como os algoritmos Apriori, FPGrowth e como Árvore de Sufixos poderiam ser usadas para encontrar padrões frequentes nos dados, porém, os resultados ainda foram um pouco genéricos, não nos fornecendo um pipeline direto a ser seguido mas sim uma visão geral do poder do algoritmo.

A modelagem do problema foi feita então quando foi perguntado ao ChatGPT como uma árvore de sufixos poderia ser usada para reconhecer padrões frequentes em dados financeiros intraday, uma ideia muito alinhada com a que tivemos quando escolhemos este tema para o projeto, e sua resposta foi a seguinte: Uma árvore de sufixos é uma estrutura de dados poderosa para encontrar padrões frequentes em séries temporais de dados financeiros intraday. Aqui está como uma árvore de sufixos pode ser usada para reconhecer padrões frequentes:

- Passo 1: Preparação dos Dados: Comece com séries temporais intraday de dados financeiros para vários ativos, representando seus históricos de preços em intervalos regulares.
- Passo 2: Construção da Árvore de Sufixos: Crie uma árvore de sufixos a partir das séries temporais, onde cada nó representa um sufixo da série temporal. Insira sufixos na árvore, atualizando contagens se já existirem ou criando novos nós.
- Passo 3: Definição de Padrões Frequentes: Estabeleça critérios para determinar o que constitui um padrão frequente na árvore de sufixos, com base na frequência de ocorrência (por exemplo, em quantas séries temporais ou quantas vezes).

- Passo 4: Identificação de Padrões Frequentes: Percorra a árvore em busca de sufixos que atendam aos critérios definidos como padrões frequentes, coletando informações sobre quais ativos ou séries temporais os têm e com que frequência ocorrem.
- Passo 5: Interpretação dos Padrões: Analise os padrões frequentes para determinar sua relevância financeira, procurando associações com eventos ou comportamentos do mercado, como tendências de preços, reversões ou outros comportamentos financeiros.

Utilizar uma árvore de sufixos permite analisar as sequências temporais de dados financeiros de forma eficiente, identificando padrões frequentes que podem ser valiosos para a tomada de decisões no mercado financeiro. É importante adaptar os critérios de frequência e a interpretação dos padrões de acordo com os objetivos específicos da análise.

6. VALIDAÇÃO

Uma vez que seguimos a metodologia definida e encontramos subcadeias frequentes nos dados, avaliar a qualidade dos resultados é um passo imprescindível, uma vez que podemos ter encontrado muitos resultados espúrios ou sem significância nenhuma para um especialista do domínio. Desta forma, questionamos o ChatGPT sobre quais métricas poderiam ser usadas para avaliar os resultados, e sua resposta foi a seguinte: A avaliação da qualidade dos padrões frequentes em dados financeiros é essencial para uma análise eficaz. Aqui estão maneiras de fazer isso:

- Suporte: Avalie a frequência de ocorrência dos padrões em relação ao tamanho do conjunto de dados. Padrões com alto suporte são mais robustos e informativos.
- Confiança: Verifique a probabilidade condicional de ocorrência de padrões, indicando associações fortes entre itens. Isso é útil para identificar regras de associação financeira.
- Lift: Use o lift para comparar a probabilidade de ocorrência de um padrão em relação à probabilidade de ocorrência dos itens independentemente. $Lift > 1$ sugere associação positiva e é considerado interessante.
- Interpretação Financeira: Avalie se os padrões têm sentido do ponto de vista financeiro, relacionando-os a eventos de mercado conhecidos.
- Relevância de Negócios: Considere a relevância dos padrões para os objetivos de negócios ou análise, se eles fornecem insights valiosos para as partes interessadas.
- Visualização e Interpretação: Use gráficos, tabelas ou visualizações de séries temporais para facilitar a compreensão dos padrões.
- Validação: Se possível, valide os padrões com dados externos para

verificar sua consistência.

- Impacto nas Decisões: Pergunte como os padrões podem informar as decisões financeiras ou estratégicas, como estratégias de negociação ou investimento.
- Iteração e Aperfeiçoamento: Esteja disposto a ajustar critérios, parâmetros e repetir a análise para melhorar os resultados.
- Avaliação de Negócios: Consulte stakeholders ou especialistas financeiros para obter feedback sobre a relevância e utilidade dos padrões encontrados.

Lembre-se de que a avaliação da qualidade pode variar de acordo com os objetivos específicos da análise e deve estar alinhada com os requisitos do projeto financeiro.

CONCLUSÕES

Durante esse relatório discutimos as soluções implementadas para o ChatGPT quando deparado com o problema de Mineração de Padrões Frequentes em Dados Financeiros Intraday. Em muitos momentos nos deparamos com problemas ao interagir com o ChatGPT, em especial com sua falta de especificidade, i.e., a geração de respostas que não respondem nada. Porém, em todos os casos foi possível contornar o problema, tirando o chat do loop em que ele estava e conseguindo conduzi-lo à uma solução mais alinhada com a que desejávamos.

Assim, considerando as respostas em que o ChatGPT conseguiu responder de forma completa e coerente com a pergunta, os resultados foram muito bons. Foram produzidos métodos que realmente fazem sentido e que apresentam resultados interessantes quando implementados.

Apêndice A

O seguinte link leva ao chat que foi usado para que obter as respostas pelo ChatGPT: <https://chat.openai.com/share/d35d0507-5549-465c-9ad6-9f697db5af1f>. Seguindo esse link você terá acesso à toda a conversa feita durante a realização do trabalho.

Outras conversas menores foram feitas em paralelo à essa, porém, no geral, eram pedidos para que o ChatGPT resumisse respostas que ele deu no chat do link.

REFERÊNCIAS

- **Huge Stock Price Data: Intraday Minute Bar.** Kaggle. Disponível em: <https://www.kaggle.com/datasets/arashnic/stock-data-intraday-minute-bar>. Acesso de 13 de set. de 2023 à 3 de out. de 2023.
- **Análise de Dados Financeiros.** ChatGPT. Disponível em: <https://chat.openai.com/share/d35d0507-5549-465c-9ad6-9f697db5af1f>. Acesso em 2 de out. de 2023 e 3 de out. de 2023.