Ciência de Dados e Analytics

Sprint: Análise de Dados e Boas Práticas

Aluno: Roberto Ritschel da Silva

PUC - RJ
09/2024

1. Definição do Problema:

1.1 Objetivo:

Analisar os dados de negociação do mercado futuro de Bitcoin, afim de identificar padrões e tendências que possam ser utilizados através de modelo de Machine Learning para previsão de tendência do preço.

1.2 Método Supervisionado:

Dado o objetivo citado acima, foi escolhido a utilizado do Método de Árvore de Decisão, devido ao problema a ser resolvido é de um movimento de "Alta" ou "Baixa" no preço de ativo, e não da previsão do valor do ativo, portanto, trata-se de um modelo de Classificação.

1.3 Catálogo de Dados:

Os dados são reais e foram obtidos na Sprint de Engenharia de Dados, entre os dias 11/06/2024 e 26/06/2024, através do uso da API pública da Binance. Além dos dados iniciais da 1ª sprint, foram calculadas também as médias móveis exponenciais de 7,21 e 100 períodos, muito utilizadas no mercado financeiro.

Nome da Coluna	Tipo do Dado	Descrição
Open Time	Datetime	Indica qual é o horário de abertura do candle
Open	Float	Indica qual é o preço de abertura do candle
High	Float	Indica qual é o preço máximo do candle
Low	Float	Indica qual é o preço mínimo do candle
Close	Float	Indica qual é o preço de fechamento do candle
Volume	Float	Indica qual é o volume financeiro do candle
sumOpenInterest	Float	Indica a quantidade de contratos ativos no momento de fechamento do candle
sumOpenInterestValue	Float	Indica o valor financeiro desses contratos ativos
Long/Short Ratio	Float	% que indica a relação entre o volume de posições compradas e vendidas

Tabela 1: Catálogo de dados da base price_data

2. Análise dos Dados

2.1 Qualidade dos dados

Para a obtenção e tratamento dos dados, não havia problema com dados faltantes (seja nulos/vazios), visto que negociação de criptomoedas, diferentemente das bolsas de valores, funcionam 24hrs. Os tipos de dados foram os quesitos mais desafiadores, visto que os dados obtidos de preço, open interest e long/short ratio são separados por ".", enquanto no csv havia a necessidade de trabalhar com ",". Após essa conversão, houve nova necessidade de conversão para "." Afim de transformar o tipo de dado de string para float.

2.2 Correlação entre Long/Short Ratio e Preço de Fechamento

O coeficiente de correlação obtido para a análise dessas features foi de -0,92, o que indica uma forte relação de correlação entre as features. Por ser negativo, indica que quanto mais uma variável aumenta, mais a outra variável diminui e, portanto, para aplicação no modelo, quanto maior for o long/short ratio, maior a tendência de queda no preço do ativo:

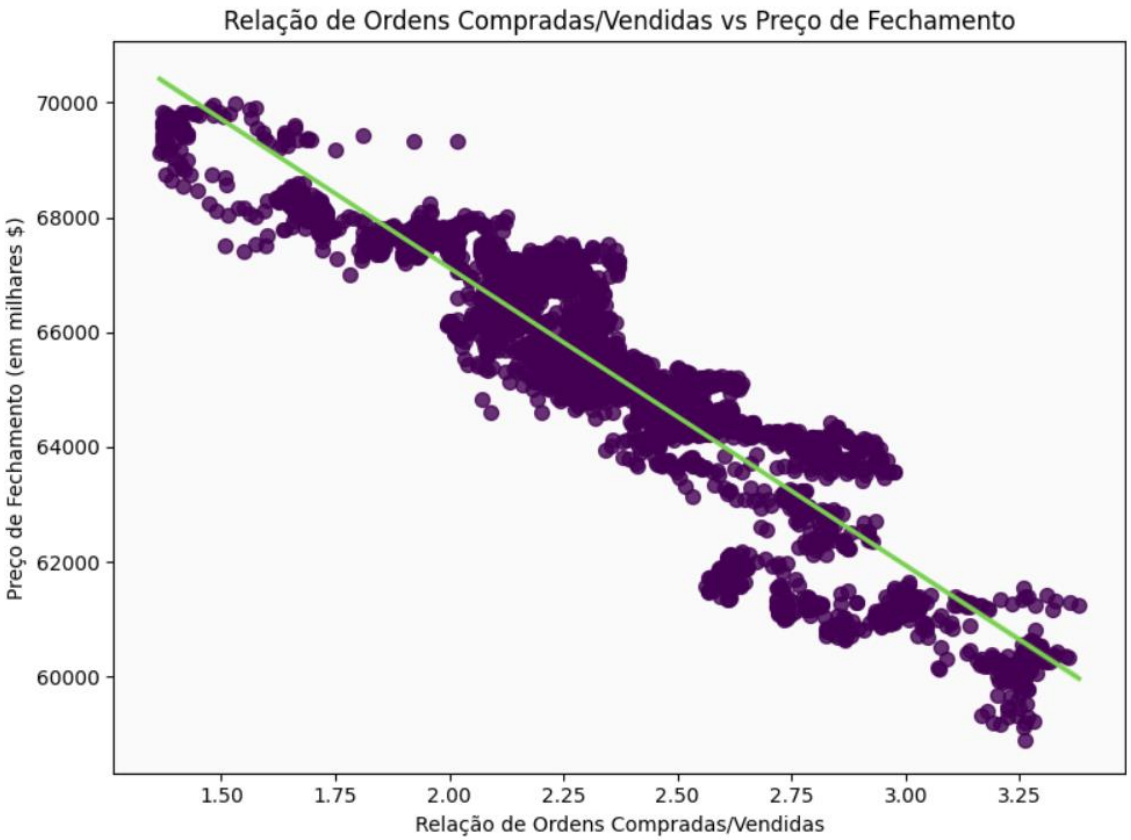

Figura 1: Gráfico de dispersão do L/S Ratio x Preço de Fechamento

```
#Cálculo da correlação
correlation = df['Long/Short Ratio'].corr(df['Close'])
print(f"Coeficiente de correlação: {correlation}")
```

✓ 0.0s

```
Coeficiente de correlação: -0.9202778983610352
```

Figura 2: Cálculo do coeficiente de correlação

2.3 Correlação entre Open Interest e Preço de Fechamento

O coeficiente de correlação obtido para a análise dessas features foi de 0,64, o que indica que há uma correlação entre as features. Por ser positivo, indica que quanto mais uma variável aumenta, a outra também aumenta e portanto, para aplicação no modelo, quanto maior for a quantidade de contratos em aberto, maior a tendência de alta no preço do ativo:

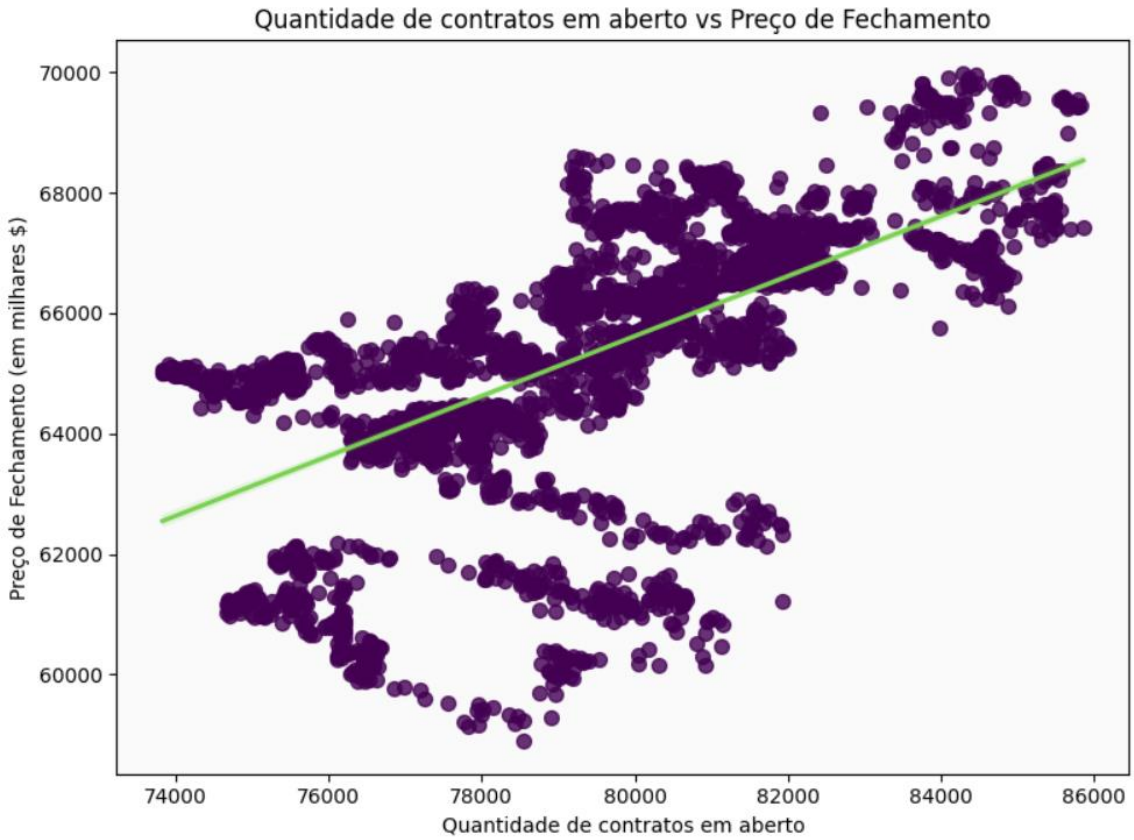

Figura 3: Gráfico de dispersão do L/S Ratio x Preço de Fechamento

```
#Cálculo da correlação
correlation = df['sumOpenInterest'].corr(df['Close'])
print(f"Coeficiente de correlação: {correlation}")
```

✓ 0.0s

```
Coeficiente de correlação: 0.6439449755290904
```

Figura 4: Cálculo do coeficiente de correlação

2.4 Árvore de Decisão

O modelo escolhido apresentou boa acurácia (54%), podendo assim ser utilizado na criação de automações de trading financeiro. É importante ressaltar

que esse valor pode variar devido à arbitrariedade dos dados no momento de treinamento do modelo.

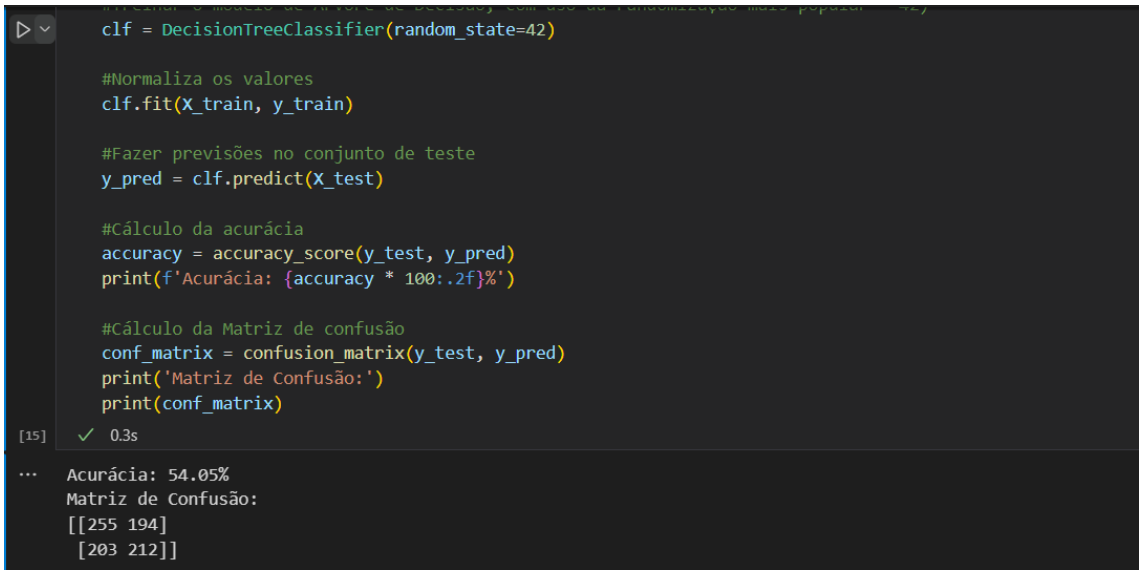

Figura 3: Acurácia e Matriz de Confusão Obtidas

Os resultados da matriz de confusão indicam que o modelo acertou 255 candles para a previsão de queda no preço, enquanto 203 casos havia sido previsto queda porém houve alta. Já para a alta, o modelo acertou 212 vezes, enquanto 194 previsões foram de alta e na realidade houve queda no ativo.