TRANSFERO SWISS AG

Transfero Swiss AG
Data Scientist Challenge

Julho de 2021
Tempo permitido: 1 semana

INSTRUÇÕES

1. O candidato poderá fazer as questões computacionais na linguagem que preferir, com respostas em um PDF. Também é possível entregar um Jupyter Notebook ao invés do PDF, ou um programa de visualizacão análogo.
2. O candidato pode e é incentivado a usar periódicos/journals para desenvolver o trabalho. Se for o caso, cite as referências.
3. Não é preciso fazer todas as questões, mas fazer o máximo que for possível.

---

1. Especialmente após a eletronificação dos mercados, o fenômeno da decentralização é de central relevância para os investidores institucionais presentes no mercado. Considere a seguinte situação: um pesquisador precisa negociar um determinado ativo, o que o faz buscar um preço de referência para o mesmo (dentre todos os disponíveis — possivelmente uma combinação deles). Sua estimativa resultante deve, em algum sentido preciso, traduzir de maneira adequada a entrada de informação de mercado que dirige o ativo.
   (a) Na condição de *Data Scientist*, o pesquisador delega a você a tarefa de abordar o problema acima. Qual seria sua metodologia para fazê-lo, e quais seriam as métricas que você utilizaria para avaliar seus resultados?
   (b) Na literatura ecônomica, existe uma fértil pesquisa na linha de "*price discovery*" — descoberta de preço. Descreva a abordagem tomada pelos autores no artigo em anexo, bem como as conclusões por eles alcançadas.

(c) [**Bônus**] Utilize os *datasets* fornecidos para ilustrar como você implementaria algum(s) dos procedimentos discutidos no item (a) ou (b).

2. Uma das mais relevantes características dos ativos digitais é a alta volatilidade observada nos preços dos mesmos. Neste contexto, é mister que tenhamos uma estimativa adequada deste parâmetro. Além disso, há grande valor em modelos que buscam prever valores futuros da volatilidade dos ativos, o que se verifica pela extensa literatura neste tema.

   (a) Focando em uma escala de tempo de alta frequência[1], cite pelo menos uma metodologia para estimativa da volatilidade. Explique como tal metodologia lida com o problema de alto nível de ruído microestrutural.

   (b) Cite pelo menos uma técnica que você utilizaria para *volatility forecasting*.

   (c) [**Bônus**] Utilize algum dos *datasets* disponibilizados a fim de ilustrar suas respostas aos itens (a) e (b).

3. Para as questões seguintes foram enviados dois (2) datasets. O código utilizado para fazê-las deve ser anexado junto às respostas no e-mail.

   Dataset trades : Dataset de todos os trades realizados na exchange determinada, do período do dia 6 de Junho de 2021 até o dia 13 de Junho de 2021 (7 dias). Seguem maiores informações sobre o dataset (*dataset_trades.zip*):

exchange	a exchange de onde o dado foi obtido
symbol	qual instrumento financeiro estava sendo monitorado
timestamp	timestamp UTC
local_timestamp	ignorar
id	código único do trade
side	tipo do trade (compra ou venda)
price	em qual preço foi realizado
amount	volume do trade

[1] Digamos, considerando séries temporais contendo todos os negócios de um par de criptomoedas em um certo período de tempo, ou de todos os *snapshots* do livro de ofertas deste ativo neste mesmo intervalo de tempo.

Dataset blocos minerados: Dataset de blocos minerados na blockchain do Bitcoin com periodicidade de 1 dia. Seguem maiores informações sobre o dataset (*dataset_blocks.zip*):

date	datetime UTC
blocks	número de blocos minerados

Faça a questão na ordem descrita abaixo:

(a) **(feature engineering)** Faça uma série temporal **OHLCV** (Open-High-Low-Close-Volume) com periodicidade de 1 minuto (*timeframe* de 1M), utilizando o dataset de trades.

Exemplo de colunas da série temporal ao término de (a):

datastamp (index)	open	high	low	close	volume

Dimensão da matrix: $[m \times 6]$

(b) **(asof-join)** Junte a série temporal originada da (a) com a série temporal de blocos minerados, obtendo uma nova série temporal que contemple em uma de suas colunas o número de blocos minerados.

Exemplo de colunas da série temporal ao término de (b):

datastamp (index)	open	...	close	volume	**blocks**

Dimensão da matrix: $[m \times 7]$

(c) **(feature enrichment)** Adicione na série temporal oriunda de (b) a feature MMS, definida a seguir em dois períodos (construir, portanto, duas features):

Média Móvel Simples do *close* da série temporal, nos seguintes períodos $n = \{3, 21\}$.

$$MMS(X = close, n) = \frac{1}{n} \sum_{i=0}^{n-1} X_{t-i}$$

Exemplo de colunas da série temporal ao término de (c):

datastamp (index)	...	volume	blocks	**MMS_3**	**MMS_21**

Dimensão da matrix: $[m \times 9]$

(d) **(feature engineering)** Crie o vetor da variável dependente utilizando as features de *open* e *close* da série temporal oriunda dos itens anteriores (itens a-c), e aplicando a função de classificação binária abaixo:

$$target(open_{t+1}, close_{t+1}) = \begin{cases} 1, & close_{t+1} \geq open_{t+1} \\ 0, & \text{Caso contrário} \end{cases}$$

(e) **(feature importance)** Utilize as séries temporais construídas nas questões anteriores, as variáveis independentes (itens a-c) e a variável dependente (item d), e faça análise de *feature importance* com o algoritmo de preferência e discorra sobre o descarte ou não das features contidas - podendo inclusive ser todas caso achar apropriado. Use os métodos, modelos e os pacotes que preferir.

(f) **(MLops)** Supondo que todos os dados utilizados pudessem ser obtidos em um certo intervalo de tempo por uma API. Como você implementaria um serviço que usa esses dados para construir e armazenar as features acima em uma infraestrutura cloud based? Nessa questão, é possível citar serviços da plataforma cloud de sua preferência (Azure, AWS, Google Cloud, etc).