Metodologia

1. Escolha das criptomoedas de estudo

As criptomoedas foram selecionadas tomando base os cinco maiores capitalizações de mercado - segundo a Coin Market Cap -, o raciocínio por trás dessa escolha foi focar nas moedas que mais importam e que podem explicar grande parte do mercado cripto.

2. Processamento das bases

Para a análise, utilizamos bases de alta frequência do histórico de pares de criptomoedas, com dados minuto a minuto e histórico diário dos grandes mercados financeiros mundiais. Nestas, foram realizados pré-processamentos para consistência dos dados, agrupamento dos dados em períodos semelhantes para análises comparativas e filtros para combinação de datas (em dados diários).

3. Modelagem

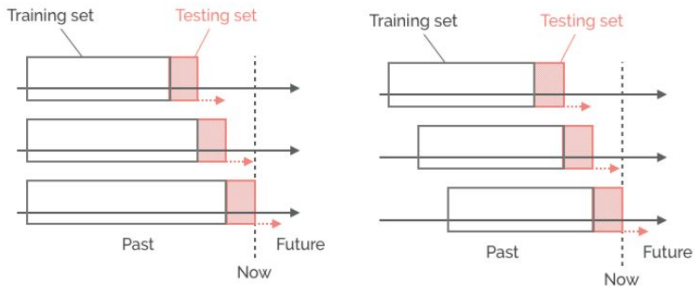
a. Modelos utilizados:

Para a modelagem, decidimos por utilizar desde modelos mais simples até alguns mais complexos, para entender como essas criptomoedas se comportam em cada uma deles e analisar o trade-off entre complexidade *vs* performance. Os modelos testados foram:

- Support Vector Machines for Regression (SVR);
- Moving Average;
- Linear Regression;
- Multi Layer Perceptron (MLP);
- Recurrent Neural Networks with LSTM blocks;
- KNN (K-Nearest Neighbors).

b. Variáveis:

Para as variáveis, há duas abordagens conhecidas, sendo elas:



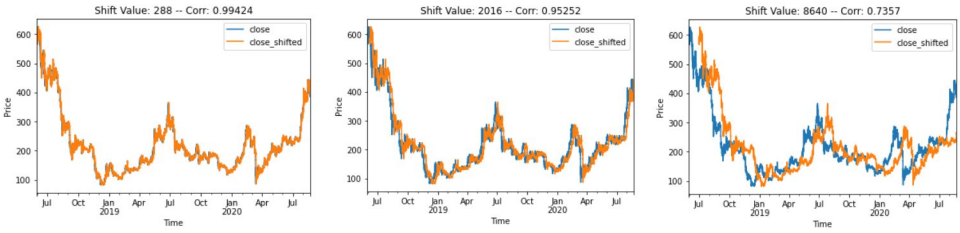
a) Janela expansiva b) Janela deslizante

Utilizamos a abordagem de janela deslizante para criação das variáveis, gerando dessa forma variáveis atrasadas como nosso conjunto de *features* e o valor atual como *target*.

4. Conjunto de teste

Para a criação do conjunto de teste nos baseamos no conceito de *backtest*, que nada mais é do que o isolamento do valor a ser predito em relação às variáveis de treino, ou seja, treinar no passado e realizar predições no futuro. É muito importante que, dada a sazonalidade de uma série temporal, seja respeitado o limite temporal de tal sazonalidade para realizar a divisão de treino e teste, para que o modelo utilizado, caso consiga, aprenda essa a sazonalidade de série. Quando quebramos tal padrão na hora de dividir nossa base podemos fazer com que o modelo não aprenda isso, ocasionando possíveis erros.

Realizamos uma análise de sazonalidade utilizando nossa série de 5-5min. Utilizamos o conceito de correlação após deslocamento temporal, ou seja, deslocamos a série em algumas possíveis sazonalidades (dia, semana e mês) e verificamos em qual desses deslocamentos a série apresentava maior correlação. Percebemos assim que a maior correlação existente era no deslocamento diário, o qual usamos para separar nossa base e também a processar.



a) Deslocamento diário b) Deslocamento semanal c) Deslocamento mensal

5. Métrica de avaliação e explicabilidade

Como métrica padrão para avaliar cada modelo foi utilizada a raiz da soma dos erros quadrados, ou RMSE, que é uma das métricas mais conhecidas para avaliação de modelos de regressão. Sua fórmula é a seguinte:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_t - y_t)^2}{N}}$$

Utilizamos a biblioteca Shap do python para realizar a interpretação das predições, a qual nos traz a importância de cada *feature* para a predição do modelo.