# Numerical Analysis for Machine Learning Project

# Bitcoin Price Prediction

**A.Y. 2022/2023**

Lorenzo Benzoni, Giacomo Cartechini

**Abstract**

The aim of this project is to try to forecast the Bitcoin price by using a Machine Learning approach. We tested several classification and regression algorithms and we compared their performances. We also tried to improve the results by using a Neural Network. The results are not satisfactory, but we think that this is due to the fact that the Bitcoin price is not a deterministic function of time, but it is influenced by many other factors. We trained the classification models to predict if the price of Bitcoin will raise or decrease the following day, and the regression models to predict the logarithmic returns of the price the following day.

## 1 Data Wrangling

We first gathered historical data about Bitcoin from coinmarketcap coinmarketcap and coinmetrics, and then we applied some data wrangling techniques to the data. We used the following features:

- Logarithmic returns at lags 1-7 days
- Volume
- Market Cap
- Relative price change
- Parkinson Volatility at lags 1-7 days
- Median Value
- Transaction Count
- New coins issued
- Total fees
- Median fees
- Active addresses
- Average difficulty

- Number of blocks
- Block size
- Number of payements
- On-chain volume
- Adjusted on-chain volume

We also scaled the data to avoid numerical instabilities in the algorithms.
Parkinson Volatility is a measure of the volatility of the price of a security, and it is defined as

$$\sigma_P = \sqrt{\frac{1}{4\ln 2}\sum_{i=1}^{4}\ln\left(\frac{H_i}{L_i}\right)^2} \tag{1}$$

With $H_i$ and $L_i$ being the highest and lowest price of the security in the i-th period.
The relative price change is defined as

$$R_{pc} = 2 \cdot \frac{high - low}{high + low} \tag{2}$$

# 2 Classification

For the classification part we have taken the positive class as the days in which the price of Bitcoin has increased, calculated as the difference between the closing price of the day and the closing price of the previous day. In this case we have used the log returns of the price at lag 1 as features. So, the positive class is defined as

$$price_{t+1} - price_t > 0 \tag{3}$$

For this reason, we have removed the last day from the dataset, because we don't have the price of the following day.

The first model we used is a logistic regression model, which yielded a pretty low accuracy on the validation set (46.4%), so we decided to try more complex models. We then tested Support Vector Machines with various kernels and parameters, and then we picked the best model on the validation set. The results of the validation procedure are shown below.

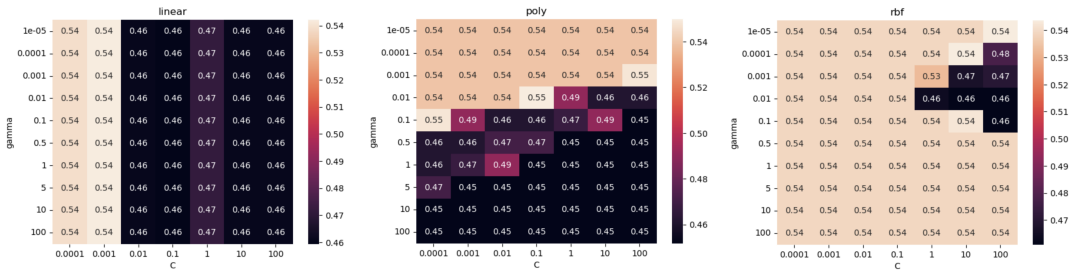


Figure 1: Linear SVM with radial basis function, polynomial and sigmoid kernels

Then we also tried a random forest classifier with different parameters, and the results of the validation are shown below.

We chose as the best model the polynomial kernel SVM with $C = 0.1$ and $\gamma = 0.01$. The estimated accuracy of this model on the test set is 48.16%.

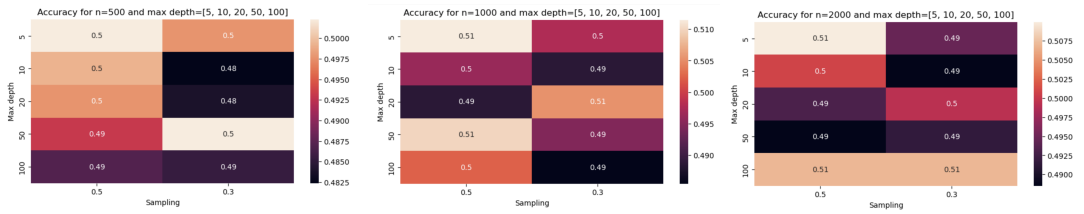


Figure 2: Random Forest Classifier

# 3 Regression

For the regression task we tried to use also different models. The first model we tried is a kernel ridge regression model, and the results on the validation set are shown below.

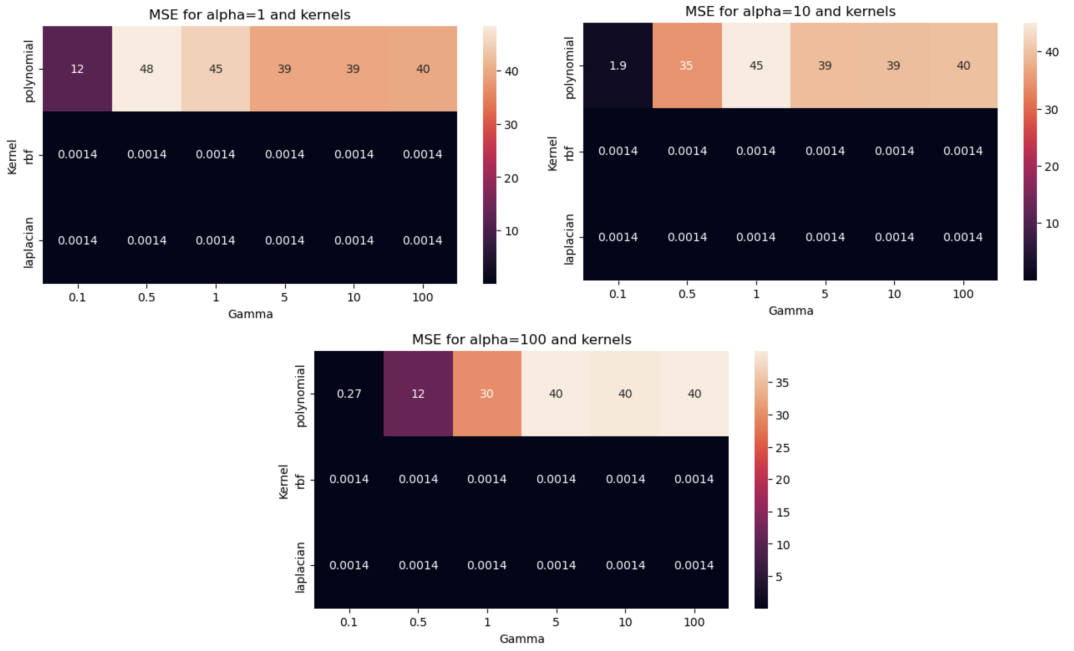


Figure 3: Kernel Ridge Regression with linear, polynomial and radial basis function

We also tried support vector machines for regression with different parameters, the results of the validation are shown below.

Finally, we tried to use a random forest regressor with different parameters, the results of the validation are shown below.

We chose as the best regression model the kernel regression with gaussian kernel, $\gamma = 1$, $\alpha = 1$, which yielded a mean squared error of 0.00179 on the test set, using the logarithmic returns as targets.

# 4 Conclusions

The final graph for the prediction of the best regression model is shown.

It looks like the model is trying to predict the mean, therefore we conclude that the Bitcoin price is not a deterministic function of time, but it is influenced by many other factors. It is possible that we did not take into consideration useful data, such as sentiment analysis data or news data which could have improved the results.

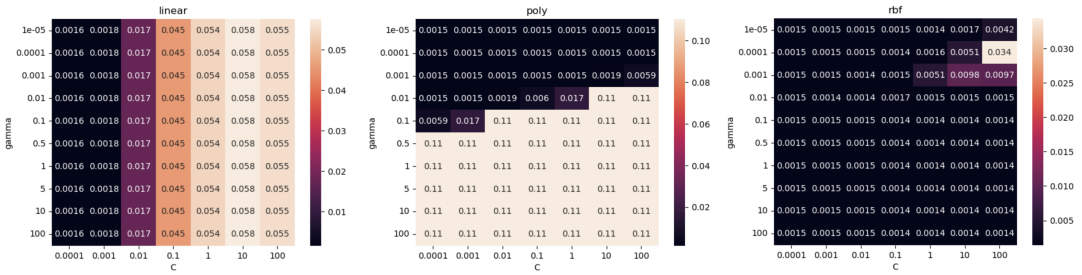


Figure 4: Linear SVM with radial basis function, polynomial and linear kernels

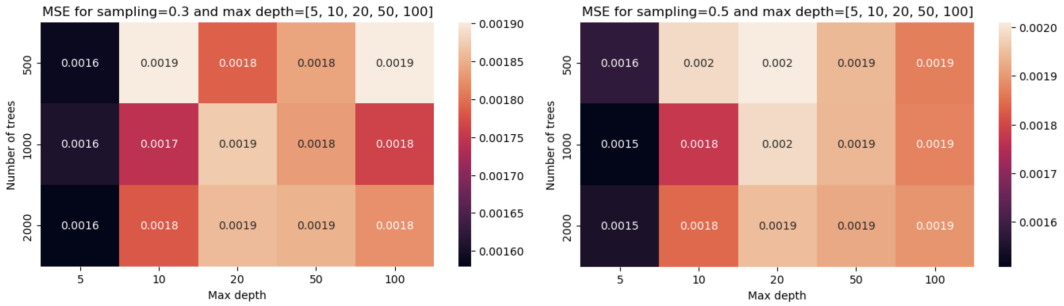


Figure 5: Random Forest Regressor

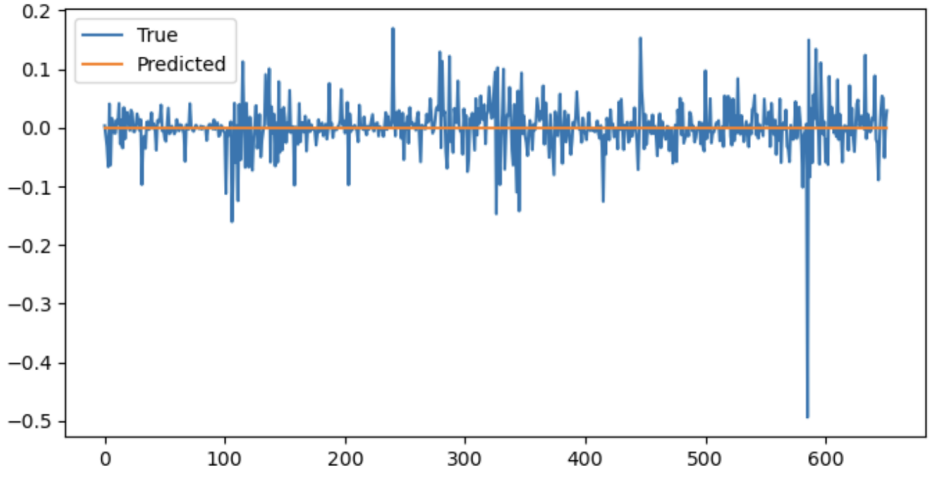


Figure 6: Final prediction