# Crypto Price Forecasting

Neel Shah AU1940055, Nipun Patel AU1940033,
Tirth Patel AU1940137, Vinay Kakkad AU1940012

***Abstract*—Predicting the expected returns of crypto-currencies would be of great industrial importance. We use machine learning expertise to predict the short-term return of 14 popular crypto-currencies. We use the data provided by G-Research which includes high-frequency market data dating back to 2018. We perform pre-processing and test several regression machine learning algorithms. Future work involves feature engineering, fine-tuning the model parameters, and testing other algorithms specialized for time series data.**

***Index Terms*—machine learning, time-series, crypto-currency, regression**

## I. Introduction

Crypto have been a major trading market that witnesses billions worth of transactions every day. Over $40 billion worth of crypto-currencies are traded every day. They are among the most popular assets for speculation and investment, yet have proven wildly volatile. Fast-fluctuating prices have made millionaires of a lucky few, and delivered crushing losses to others. So, we tried to solve the dilemma about predicting such a volatile quantity with an appreciable accuracy using the prevalent machine learning algorithms. We also aim to learn different approaches associated with forecasting problems along with their limitations.

## II. Literature Survey

Crypto currencies have become a poplar asset class in the recent years and thus forecasting it has also became an poplar problem in Machine Learning. In general data used for forecasting is called time series data as time is one of the most important feature while predicting. The paper by Bontemp G. describes the machine learning strategies used while dealing with the time series forecasting problem. It discussed linear statistical methods such as ARIMA, non-linear models such as threshold auto regressive model, Decision trees, support vector machines and Black box approaches such as ANNs.

Velankar S. gave some interesting insights that how crypto currencies can not be treated exactly as sales or stocks data. The reason being unlike sales and stocks the price of these assets are not directly affected by the business news or the ruling government.

Besides time series, an interesting approach to predict price was used by researchers at Southern Methodist University. They used twitter tweets volume and sentiments to predict the prices. Wołk, Krzysztof also used social media sentiments for sort-term price prediction.

## III. Implementations

### A. **Dataset**

We are using the data set provided by G-research on Kaggle for their crypto forecasting competition. The data set provides historic price and trade data for 14 assets from 2018 to 2021. The records have 1 minute gap between them.

### B. **Pre-Processing**

There were only two thing that we processed in our dataset before using it for EDA and model building.

- The difference between consecutive timestamp was not 60 second across whole data set. As we need the consistent gap between entries for modeling time series, We added missing entries by padding as the gaps were not that big.
- Some of the entries in the dataset we NA. Compared to the dataset size the number of entries with NA entries were very small and therefore we simply replaced NA with 0.

### C. **Exploratory Data Analysis**

- We plotted the graphs of close price vs time to for different assets to visualize the data set. By doing so we also got the basic idea of how the price change in BTC is reflected in prices of other assets.
- We plotted Close price P(t) with P(t-1), P(t-2) etc, to see if price is dependent on the past price of the assets. Here we saw strong correlation between the price at t to price at t-4 after that the correlation was not strong.
- We also tried to decompose the time-series of different assets into components such as trend, seasonality and residual/noise.

### D. **Models**

We filtered out our dataset for a particular assets and then implemented our model. we have split the data into train and test data. This separation of dataset is not done randomly. First 80% of the data is split into train data and last 20% of the data is split into test data.

1. Linear Regression
   We have implemented the linear regression model on our dataset. For that we have used LinearRegression from linearmodel of sklearn library. We trained our dataset from an object of LinearRegression class using train dataset. We predicted the output of test dataset using this trained model. Then we are finding the correlation between the actual values and predicted values for every assets.
2. ARIMA
   We have implemented the ARIMA on our dataset. For that we have used ARIIMA from statsmodel library. We chose the value of $p(regressivecomponent)$= 4, $q(laggederror)$=2 and d(moving average compo-

nent)=0. These hyper-parameters were found on the trail and error basis and might not be the optimal ones. We then fit the model on the training data and predicted the output for the test set. Then we are finding the correlation between the actual values and predicted values for every assets.

## IV. Results

### A. Linear Regression

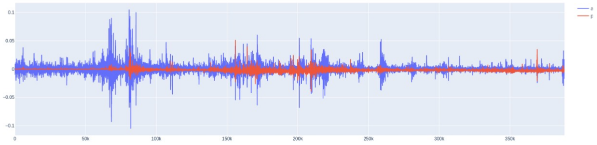


Fig. 1. Binance coin: Linear Regression Prediction

In the Fig 1. red lines shows the predicted value of model and blue line shows the actual values. Linear Regression was under-fitting and failed to capture the volatile nature of the crypto assets.

### B. ARIMA

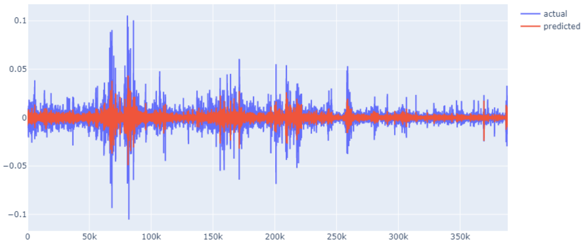
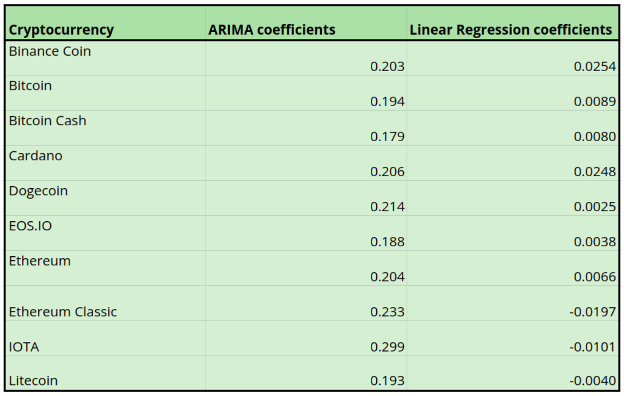


Fig. 2. Binance coin: Arima Prediction

Comparing the Fig. 1 and Fig. 2, we can see that the prediction lines i.e. red lines of ARIMA model are better than that of Linear Regression. Even though the ARIMA model produced better result it is still not a good model as we can see there signicant difference between actual and predicted values at some point in the plot.



| Cryptocurrency | ARIMA coefficients | Linear Regression coefficients |
|---|---|---|
| Binance Coin | 0.203 | 0.0254 |
| Bitcoin | 0.194 | 0.0089 |
| Bitcoin Cash | 0.179 | 0.0080 |
| Cardano | 0.206 | 0.0248 |
| Dogecoin | 0.214 | 0.0025 |
| EOS.IO | 0.188 | 0.0038 |
| Ethereum | 0.204 | 0.0066 |
| Ethereum Classic | 0.233 | -0.0197 |
| IOTA | 0.299 | -0.0101 |
| Litecoin | 0.193 | -0.0040 |

Fig. 3. Comparison of correlation for ARIMA and linear regression

The above table shows the correlation between the predicted values and actual values for both the models. We can clearly see that ARIMA is performing better the Linear Regression.

## V. Conclusions

We conclude that the ARIMA based model is capturing the right mapping compared to linear regression. At practical scenario current implementations of Linear Regression and ARIMA are not good at forecasting. But we do think that by doing the feature engineering for linear regression and finding the right hyper parameters of ARIMA will improve the results. We also plan to explore other models used for tackling the time series problem.

## References

[1] J. Fattah, L. Ezzine, Z. Aman, H. Moussami, and A. Lachhab, “Forecasting of demand using arima model,” *International Journal of Engineering Business Management*, vol. 10, p. 184 797 901 880 867, Oct. 2018. doi: 10.1177/1847979018808673.

[2] Konradb, *Ts-2: Linear vision*, Mar. 2022. [Online]. Available: https://www.kaggle.com/code/konradb/ts-2-linear-vision/notebook.

[3] S. Velankar, S. Valecha, and S. Maji, “Bitcoin price prediction using machine learning,” in *2018 20th International Conference on Advanced Communication Technology (ICACT)*, 2018, pp. 144–147. doi: 10.23919/ICACT.2018.8323676.

[4] J. Abraham, D. Higdon, J. Nelson, and J. Ibarra, *Cryptocurrency price prediction using tweet volumes and sentiment analysis*. [Online]. Available: https://scholar.smu.edu/datasciencereview/vol1/iss3/1.

[5] Iamleonie, *To the moon [g-research crypto forecasting eda]*, Dec. 2021. [Online]. Available: https://www.kaggle.com/code/iamleonie/to-the-moon-g-research-crypto-forecasting-eda/notebook.

[6] G. P. Zhang, “Time series forecasting using a hybrid arima and neural network model,” *Neurocomputing*, vol. 50, pp. 159–175, 2003.