Crypto Price Forecasting Using Machine Learning Algorithms

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 currencies 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. Also the simultaneous activity of thousands of traders ensures that most signals will be transitory.

Our aim is to answer the question that Can we forecast short term (15 min) returns of highly volatile crypto assets?. We have used the Machine Learning techniques to tackle the above question. Our contribution can se summarized as

1) Found the suitable data set for the given problem.
2) Performed Exploratory Data Analysis on the data set.
3) We applied several variants of Regression algorithm for the given problem.
4) We also looked at statistical methods for time series forecasting such as ARIMA.
5) We tuned the hyper parameters of each model to attain best possible results and compared them.

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 information for 14 different cryptocurrencies from 2018 to 2021. The dataset includes features like open, close, high, low, VWAP(volume weighted average price) volume and Target. The records have 1 minute gap between them.

TABLE I
Assets in the Dataset

Asset ID	Asset Name
0	Binance Coin
1	Bitcoin
2	Bitcoin Cash
3	Cardano
4	Dogecoin
5	EOS.IO
6	Ethereum
7	Ethereum Classic
8	IOTA
9	Litecoin
10	Maker
11	Monero
12	Stellar
13	TRON

B. **Exploratory Data Analysis** In exploratory data analysis, we plotted various visualisation for building intution about the data.

1. On plotting the closing prices of different assets, we could observe that the all the asset follow a similar trend. But on plotting the closing the price against the lagged closing price, we could observe a strong relation between them.

2. On plotting a correlation matrix for the closing price of all the assets, we could observed that various cryptocurrencies have a very high correlation.
3. On plotting the correlation matrix for different predictors, we could observe that columns like close, high, low and VWAP are highly correlated with the open price.
4. We plotted the $correlation\ against\ time$ plots of all the assets by keeping bitcoin as the base asset. From this we could observe that assets show a high but variable correlation.
5. On plotting the autocorrlation plots for the target variable of all the assets, we could observe very low correlation, which tells us that data is performing a random walk.

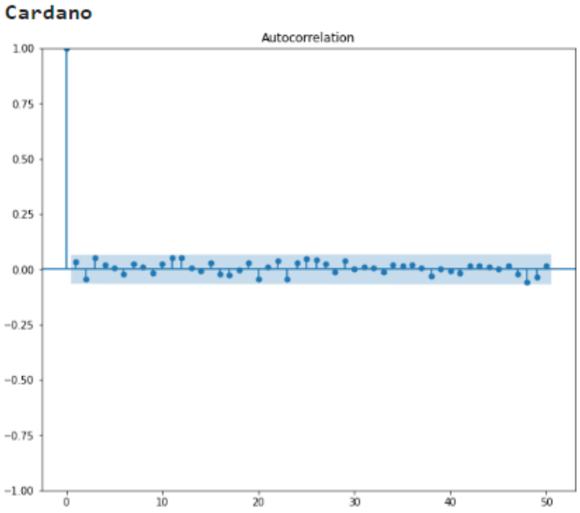

Fig. 1. Autocorrelation Plot for Cardano

6. Time series data contains various trends and seasonal patterns. We decomposed the target of all assets into such trends and seasonal components and plotted them.
7. Through ADF(Augmented Dickey Fuller) test, we could observe that data is stationary(i.e. it does not contain a strong seasonal component)

C. Pre-Processing and Feature Engineering

1. 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 the gaps with last available data point.
2. Some of the entries in the target column of the data were NA. Compared to the dataset size the number of entries with NA entries were very small and therefore we simply replaced NA with 0.
3. Some of the predictor were highly correlated thus we replaced with two new features - upper shadow and lower shadow
4. The past information is very important for time series forecasting, thus we added lag features.

D. Regression Models

TABLE II
FEATURE ENGINEERING

Feature	Value
upper shadow	High - max(Close, Open)
lower shadow	min(Close, Open) - Low

TABLE III
SELECTING NUMBER OF LAGGED FEATURES

Number of Lag Features	Performance(Correlation)
0	0.0231
2	0.0319
4	0.0336
8	0.0352
16	0.0365

1. *Simple Linear Regression*
   We performed linear regression through ordinary least squares to create a baseline model with all the default predictors. We fit the model using the train data and calculated the performance by finding the correlation between the actual and predicted targets.
2. *Multioutput Linear Regression*
   As we observed that the prices of different cryptocurrencies show a high correlation, we created a multi-output regressor that extracts information from other assets.
   Based on the observation from EDA, we performed feature engineering (adding the $upper\ shadow$, $lower\ shadow$, and $lag\ features$). We added up to 20 lag features by iterative feature selection, after which we started observing saturation. Feature engineering helped us beat the baseline scores obtained from multioutput and simple linear regression.

E. Statistical Models

Statistical models such as MA, AR, ARIMA uses past values and past errors to for estimating the future values for time series forecasting. We have for our problem used ARIMA as it can model/learn both moving average and auto regressive part in a single model.

1. ARIMA
   The findings of exploratory data analysis suggested that the data for all the coins is stationary and is doing random walk. Usually ARIMA(p,d,q) with hyperparameters (0,1,0) is used for modelling the random walk. However through experimentation with different settings we tried to find the optimal hyper-parameters. The tuned hyper-parameters are mentioned in Table-1.

TABLE IV
TUNED HYPERPARAMETERS FOR ARIMA.

Hyperparameters	Value
p(order of the autoregressive part)	4
d(degree of first differencing involved)	1
q(order of the moving average part)	2

F. Ensemble Learning

Ensemble methods use multiple learning algorithms to achieve better performance than what could be obtained from one class alone. For our problem we have used LightGBM a gradient boosting technique for tree based learning algorithms.

1. LGBM Regressor
   We trained the different assets on separate models and found optimal hyper-parameters for each asset rather than training a common model for all the assets. The tuned hyperparameters for two assets namely Bitcoin and Stellar are shown in Table-2 ans Table-3.

TABLE V
TUNED HYPERPARAMETERS OF LGBM FOR BITCOIN.

Hyperparameters	Value
Learning Rate	0.01
No. of Leaves	111

TABLE VI
TUNED HYPERPARAMETERS OF LGBM FOR STELLAR.

Hyperparameters	Value
Learning Rate	0.05
No. of Leaves	41

IV. Results

A. Linear Regression

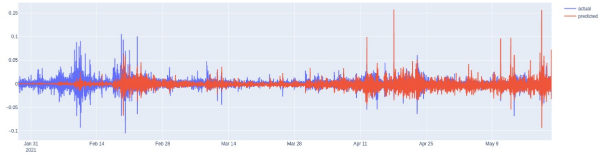

Fig. 2. Binance coin: Linear Regression Prediction

In the Fig 1. red lines shows the predicted value of model and blue line shows the actual values.

B. ARIMA

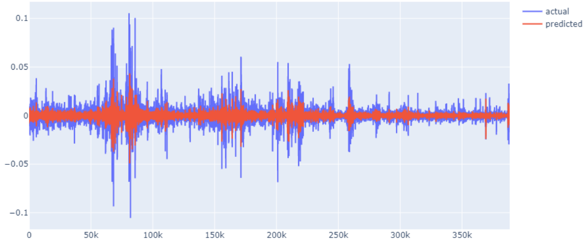

Fig. 3. Binance coin: Arima Prediction

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.

C. LGBM

After tuning the hyperparameter, the LGBM model gave us a correlation of score of 0.0372. The results of all the models are summarized as below:

TABLE VII
FINAL RESULTS

Model	Performance(Correlation)
Simple Linear Regression	0.013
Multioutput Linear Regression	0.0231
Regression(after feature Engineering)	0.0365
ARIMA	0.0423
LGBM	0.0372

V. Conclusions

We conclude that the classical machine learning algorithms are unable to capture completely the volatile nature of the short term returns of cryptocurrencies. As the evaluating metric is correlation coefficient, we expected to see the performance of models to be greater than 0.6 which is way higher than what we obtained. We can say that in practical scenario performance of the models discussed here are not up to the mark. The failings of classical machine learning can be justified by the fact that assets returns are highly stationary and as found from the EDA are doing a random walk. We also observed that through rigorous EDA and feature engeineering, we were able to significantly improve the score of Linear Regression. We believe that the performance can be improved by some extend by looking at Deep learning based approaches such as ANNs and LSTMs and and combining them with the classical machine learning models.

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.