Crypto Forecasting

Mid-Semester Report

Sameep Vani AU1940049 | Kashvi Gandhi AU1940175 | Kavya Patel AU1940144 | Khushi Shah AU1920171

I. Abstract

This project intends to forecast the short term returns of 14 popular crypto currencies using various Machine Learning techniques. We have a high-frequency market data dating back to 2018 which will be used for building an efficient model with maximum accuracy and minimum error. The dataset of crypto forecasting is a standard time-series data.

II. Keywords

1) Crypto Currency
2) Regression
3) Time Series
4) Vectorizing
5) Pandas
6) Apache Spark

III. Libraries Used

1) Pandas
2) PySpark
3) Numpy
4) Scikit-learn
5) Matplotlib
6) Seaborn

IV. Introduction

CRYPTO CURRENCY has revolutionized the idea of future finance. It is thought to be the next internet revolution where transactions will be done utilizing peer-to-peer network creating a blockchain of users. It is ought to make the world a better place by reducing the risks of fraud, making e-commerce stronger, making transactions safer, keeping individuals and companies accountable and on top of all encouraging the scientific advancements.

Crypto currency is a digitally/virtually existing currency that uses cryptography to secure transactions. They are one of the most popular assets for speculation and investments but are highly volatile. The fluctuating nature causes both hype and risk.

Crypto Forecasting is a process of utilizing the time-series data of given cypto currencies along with other features to predict the future value of these crypto currencies by analysing the past trends and data. It is a challenge to forecast crypto currency prices using all the trading features like price, volume, open, high, low values present in the dataset.

V. Literature Survey

Crypto currency price prediction is one of the trending topic amongst the researchers. For time series forecasting, the researchers have focused on three aspects :

1) Formalization of one-step forecasting problems as supervised learning tasks
2) The duscussion of local learning techniques as an effective tool fro dealing with temporal data
3) The role of the forecasting strategy while moving from one-step to multi-step forecasting.

Further, recent papers on this topic has led to development and comparison of using many algorithms such as Bayesian Regression, Linear Regression, SVM, ANN, Deep Learning and reinforcement learning. Not only this, each algorithm is used with a suitable and diversified performance metric such as RMSE, R2, f1-statistics, Relative RMSE, Correlation Coefficients, adjusted R2, MSE, Accuracy etc. Out of all these, the results have proven DN-ANN pair along with LSTM (Long Short Term Memory) proved to be the best models. Further, time series analysis has also grown to be efficient especially with powerfull models such as ARIMA and SARIMA combined with Random Forests algorithm for Machine Learning. Further, there are numerous supervised learning approaches to solve time-series analysis. Many traditional statistical/Machine Learning models such as KNN and Random Forest Regression. Along with these, laxy learning algorithms, multi-step time series forecasting (iterative and recursive strategies) have also been developed.

VI. Implementation

1) Use PySpark to process the data which includes handling missing values, outliers and temporal/time dependent features by converting them into human readable format so that visualization can be done.
2) Create a base model (Linear Regression) and evaluate it on validation data based on which other models will be compared.
3) Using seaborn library along with matplotlib, perform Exploratory Data Analysis for time series data. This step includes creation and exploration of lag features, trends and seasonality for the given dataset.
4) Based on the results of EDA, implement feature engineering methods such to create new features that might be helpful.
5) Implement different models such as Polynomial Regression, Moving Average, ARMA, ARIMA, SARIMA etc.
6) Select the model that is performing best using validation or cross-validation approaches. Further use Pearson Correlation Coefficient along with MSE as performance

metrics. We are using Pearson Correlation Coefficient because of the fact that Kaggle is evaluation the submissions based on this attribute.

7) Lastly use Optuna Library (in python) to tune the hyperparameters of that model.
8) Finally, again train the tuned model and make predictions on the test/unseen/future data.

VII. Results

We have created a base model (linear regression) based on the graphs generated by plotting time vs target. Below shown are only 3 examples of the 14 graphs that were generated. Based on the graphs given below, we can conclude that Linear Regression Model can be a good fit although it is still left for final conclusion which will compare other algorithms as well. Further as expected, the linear regression MSE is approximately 0.00003 which is already quite less. However, it is still left to conclude that Linear Regression model is the best fit or not.

Fig. 1. Asset ID 0: Binance Coin

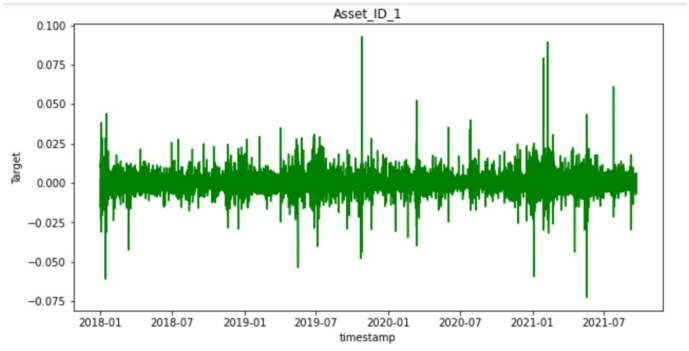

Fig. 2. Asset ID 1: Bitcoin

VIII. Conclusion

Linear Regression model is not the only model that fits time series data. Even though the model has a very low Test MSE (0.00003) but it is highly probable that since the overall actual value and predictions are so small that the overall MSE is small. Thus, there is a need to compare with other more powerfull time-series models that can help to see which

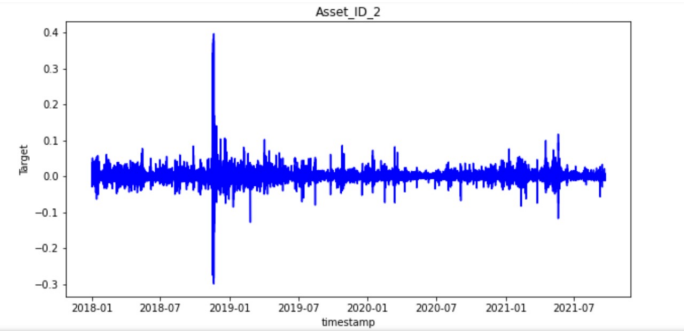

Fig. 3. Asset ID 2: Bitcoin csh

```
In [8]: from sklearn.model_selection import train_test_split
        X_train_1, X_test, y_train_1, y_test = train_test_split(X_train_pd, y_train, random_state=SEED_VALUE, train_size=0.8

In [9]: from sklearn.linear_model import LinearRegression

        lr_model = LinearRegression()

        fitted_lr_model = lr_model.fit(X_train_1, y_train_1)

In [13]: fitted_lr_model.coef_
Out[13]: array([-2.35027333e-06,  1.94206659e-08,  9.34639626e-07, -5.12755579e-07,
                 8.32016720e-08, -5.06421801e-07,  2.45604199e-11])

In [17]: fitted_lr_model.intercept_
Out[17]: 1.249994415045429 7e-05

In [10]: preds = fitted_lr_model.predict(X_test)
         preds
Out[10]: array([ 2.11456602e-05, -1.08168385e-05,  1.46151472e-05, ...,
                 6.59229985e-06,  3.46810487e-05, -4.55057467e-06])

In [11]: from sklearn.metrics import mean_absolute_error, mean_squared_error

         mse = mean_squared_error(y_test, preds)
         mse
Out[11]: 3.228664238435009e-05
```

Fig. 4. Base Model : Linear Regression

performs the best on test data.

To conclude, the final data manipulation will be done with pyspark (Apache Spark API for python). Finally, the model will be trained using Scikit-Learn and tuned using Optuna.

IX. References

[1] Mudassir, M., Bennbaia, S., Unal, D. and Hammoudeh, M., 2020. Time-series forecasting of Bitcoin prices using high-dimensional features: a machine learning approach. [online] SpringerLink. Available at: ¡https://link.springer.com/article/10.1007/s00521-020-05129-6¿ [Accessed 3 March 2022].

[2] Sebastião, H. and Godinho, P., 2021. Forecasting and trading cryptocurrencies with machine learning under changing market conditions. [online] SpringerOpen. Available at: ¡https://jfin-swufe.springeropen.com/articles/10.1186/s40854-020-00217-x¿ [Accessed 25 February 2022].

[3] Khedr, A., Raj P V, P., El-Bannany, M. and Arif, I., 2021. Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey. [online] Available at: ¡https://www.researchgate.net/publication/350347505_Cryptocurrency_price_prediction_using_traditional_statistical_and_machine-learning_techniques_A_survey¿ [Accessed 25 February 2022].

[4] Bontempi, G., Taieb, S. and Le Borgne, Y., 2013. Machine Learning Strategies for Time Series Forecasting. [online] Available at: ¡https://www.researchgate.net/publication/236941795_Machine_Learning_Strategies_for_Time_Series_Forecasting¿ [Accessed 22 February 2022].

[5] Bontempi, G., Taieb, S. and Le Borgne, Y., 2013. Machine Learning Strategies for Time Series Forecasting. [online] Available at: ¡https://www.researchgate.net/publication/236941795_Machine_Learning_Strategies_for_Time_Series_Forecasting¿ [Accessed 22 February 2022].

[6] Youtube.com. 2022. Great Learning. [online] Available at: ¡https://www.youtube.com/watch?v=FPM6it4v8MY¿ [Accessed 13 February 2022].

[7] Peixeiro, M., 2019. The Complete Guide to Time Series Analysis and Forecasting. [online] Medium. Available at: ¡https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775¿ [Accessed 10 February 2022].

[8] Raman, V., 2020. Time-Series Forecasting with Spark ML: Part—1. [online] Medium. Available at: ¡https://medium.com/analytics-vidhya/time-series-forecasting-with-spark-ml-part-1-4e888144ad27¿ [Accessed 10 February 2022].

[9] Kaggle.com. 2022. G-Research Crypto Forecasting — Kaggle. [online] Available at: ¡https://www.kaggle.com/c/g-research-crypto-forecasting/data¿ [Accessed 7 February 2022].