Crypto Forecasting

Final 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. Various models like XGBoost, LGBM, KNN, Ridge and Lasso Regression, etc. are used for model fitting and the final results are generated using (model finalized) as it gives the minimum MSE.

II. Keywords

1) Crypto Currency
2) Regression
3) Time Series
4) Vectorizing
5) Pandas
6) Apache Spark
7) Exploratory Data Analysis
8) Feature Engineering
9) Model Fitting
10) Hyperparameter Optimization
11) Heatmaps
12) Grid Search

III. Libraries And Frameworks Used

1) Pandas
2) Numpy
3) Scikit-learn
4) Matplotlib
5) Seaborn
6) Optuna
7) lightbgm
8) xgboost
9) Multiprocessing

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 discussion of local learning techniques as an effective tool for 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 powerful 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 have proved to be effective. Along with these, lazy learning algorithms, multi-step time series forecasting (iterative and recursive strategies) have also been developed.

VI. Implementation

1) Data preprocessing to handle missing values using Numpy and temporal/time dependent features by converting them into human readable format using pandas so that visualization can be done. Also, the data is normalized using Scikit-learn library to organize the data and minimize redundancy.
2) Creating a base model - Linear Regression and evaluate it on validation dataset based on which other models will be compared.

3) Using seaborn library along with matplotlib to 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) Plot every 'feature' vs 'Target' to analyze the dependencies of target on individual features.
5) Outlier detection has been done in case of any. Moreover, plotting correlation between all the features and target for every Asset ID individually and together as well.
6) Based on the results of EDA, implementing feature engineering methods to delete features that do not contribute to predict the target and to create new features that might be helpful.
7) Futher implement other regression models including KNN, Random Forest, Decision tree, XGBoost, LGBM, Lasso and Ridge Regression and compare them with the base model - Linear Regression.
8) Select best 3 models that perform best using validation or cross-validation approaches. Here, MSE is used as the performance metrics.
9) Lastly use Grid Sreach and Optuna for the hyperparameter optimization on these models.
10) Finally, train the tuned model and make predictions on the test/unseen/future data.

VII. Results

Initially, we created a base model - Linear Regression. 'Timestamp' vs 'Target' graphs are generated. Fig. 1 and Fig. 2 shows graphs of Asset ID 0 and 1 respectively. The MSE for the linear regression is approximately 3×10^{-5} which is already quite low. However, after pre-processing the data and carrying out feature engineering, we found that Linear regression was not the best fit model for our data.

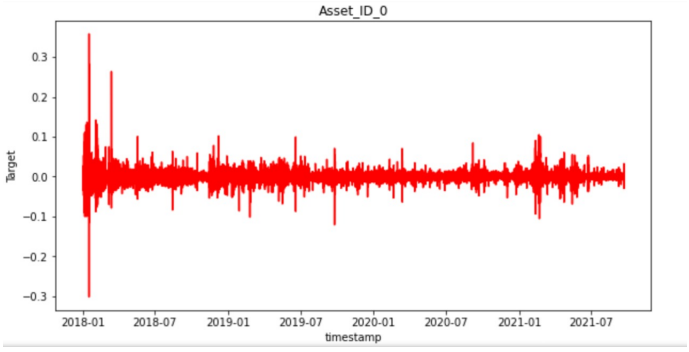


Fig. 1. Asset ID 0: Binance Coin

On carrying out Exploratory Data Analysis by using a correlation heatmap, we found that the correlation between features including 'Open', 'Close', 'High', 'Low' and 'VWPA' is 1 and hence, they are perfectly correlated. Due to the linear dependencies between these features, we can drop all except one of these columns.The correlation graph is shown in Fig. 3.

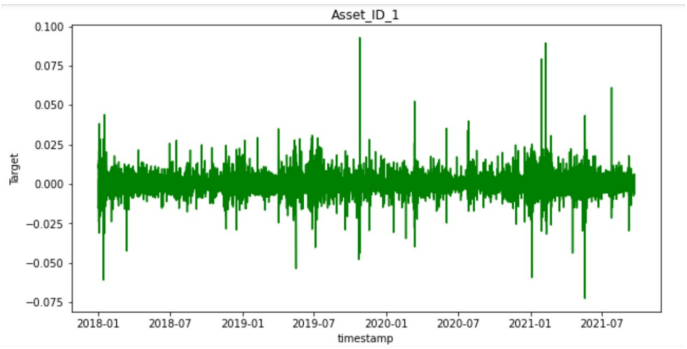


Fig. 2. Asset ID 1: Bitcoin

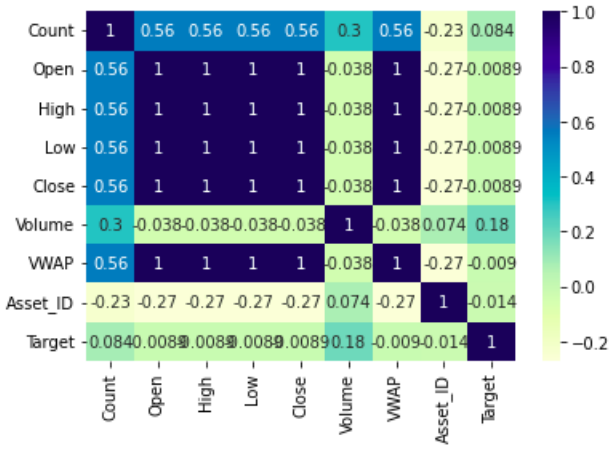


Fig. 3. Correlation Heatmap

Moreover, to understand the relation and correlation between the previous values of the 'Target' feature, we implemented Lag Features. By creating new columns in the data set, we shift the 'Target' Values by a specific number of rows. Then a correlation is found using the in-built functions of Pandas. However, upon implementation and plotting of lag features over Target values very low correlation was found between the current value and the previous values.

Fig. 4. Heatmap for Lag features

We have also generated box plots to check for outliers. It

turns out that there are not many outliers that affect the data set and hence there was no point of removing them. Also, plots for seasonal trends was generated. The graph represented monthly change in Target values for every year for every Asset ID. It was seen that the Target values shows spike for every start of the month every year but there is no clear trend within the month which is why there is no conclusion for the same.

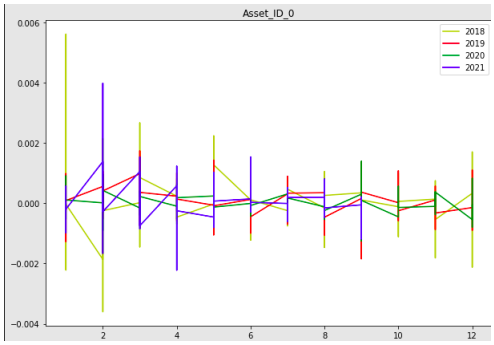


Fig. 5. Seasonality Graph

The distribution of each feature is plotted. It turns out that every feature is distributed as skewed Gaussian distribution. The variance is low for every feature and the mean is nearly 0. Every graph has a peak thus has high kurtosis.

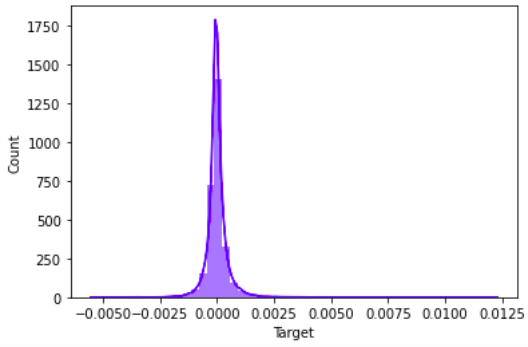


Fig. 6. Distribution of Features

Later we performed model fitting using different Regression models to find which model best fits the data. We used models including KNN Regressor, Decision Tree Regressor, Random Forest Regressor, XGBoost Regressor, LGBM Regressor, Lasso Regressor and Ridge Regressor. The MSE for each, after model fitting are shown below.

```
Linear Regression:  0.27847453553821033
KNN:  0.2750872586068413
Decision Tree:  0.6172909686943512
Random Forest:  0.242716706232181
XGB:  0.23056290071649205
LGBM:  0.20484413908640328
LassoCV:  0.23906921550956728
RidgeCV:  0.27293100941055715
```

Fig. 7. MSE for different ML models

According to this, the three best models that we get are XGBoost, LGBM and Lasso.The hyperparameter optimization

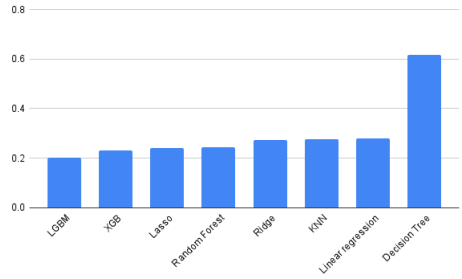
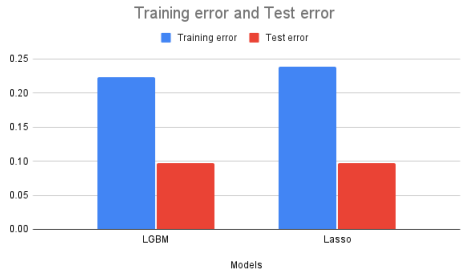


Fig. 8. MSE for different ML models

will be carried out for these three models so as to tune the model and as a result get lowest MSE.

For the Hyperparameter optimization, we used Grid Search and Optuna for tuning our models. We tested the samples on the testing supplement On training using the optuna, the MSE for Lasso Regression changed from 2.39069 to 0.098, for LGBM changed from 2.04844 to 0.097 while that of XGBoost changed from 0.2144 to 200 which is very high as compared to the other two(the numbers are in order of 1×10^{-6}).



(1).png

Fig. 9. MSE on supplement train data and the train data

VIII. Conclusion

We had three ML models that gave us the least MSE for our data, namely,Lasso regression, XGBoost and Light Gradient Boosting Machine on model fitting. These three model after tuning and testing on the supplement train gave us different MSE values. The test errors of both LGBM and Lasso are found to be the same. However, the train error for both were different and that of LGBM is lower and so the generalization gap for the LGBM turns to be lower than than that of Lasso. Hence, we can safely choose LGBM as the best model and for predicting future returns.

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].