G-Research Crypto Forecasting

Cao Bokai, Lai Fujie, Luo Zhuang, Shi Jie

6010Z

December 12, 2021

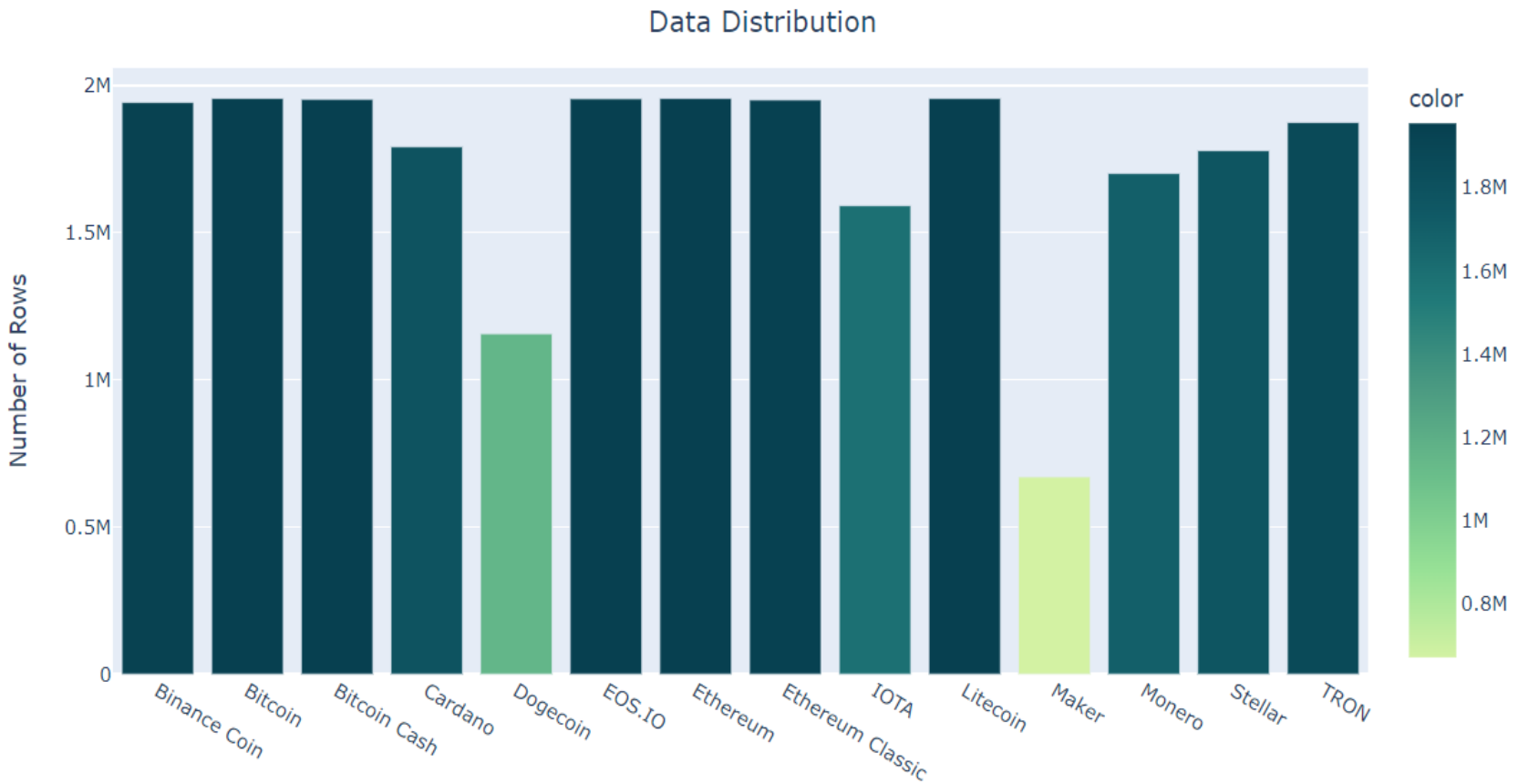
Abstract

- Over $40 billion worth of cryptocurrencies 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.
- In this competition, we machine learning expertise to forecast short term returns in some popular cryptocurrencies.
- In this report, we mainly focus on five parts. First is the data preparation. Second is feature engineering, where we design lots of features. The third part is brief introduction to our model and method used.
- After that, we give performance evaluation and interpretation of the result based on our model. We make our conclusion and prediction of the cryptocurrencies return trend.

Data preparation

Part I

- We use time series data of 14 cryptocurrencies and 5 datasets.
- These data dating back to 2018, and each row of dataset contains transaction information in 1 minute
- We performed some simple processing on the data.



Feature Engineering

Part II

- We made some factors with practical financial significance by simple transformation using price and volume data. Like spread(HIGH-LOW), upper/lower shadow, CLS. . .
- Besides above factors and basic price-volume data, we also take some outstanding alphas in conventional financial market (e.g. stock, futures market) as a reference.
- Our factors will directly promote the model performance to a more precise, robust, comprehensible one.
- For example, Alpha002 in the code, (-1*DELTA((((CLOSE-LOW) -(HIGH-CLOSE)/(HIGH-LOW)),1)), indicates long short power imbalance and how it moves since ((CLOSE-LOW)-(HIGH-CLOSE))/(HIGH-LOW) is clearly unbalancedness of long and short.

Model Selection

Part III

- our goal is to process cryptocurrency's time series data and do price predictions.
- In traditional finance research, ARIMA and GARCH are the most useful models and they are built with clear assumptions on time series data.
- But some research has been done and one of them suggests that LSTM outperforms traditional-based algorithms such as the ARIMA model. That's the reason for us to choose LSTM as our first model.

Model Selection

LSTM

LSTM refers to long short-term memory, is one of the recurrent network structures. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate.

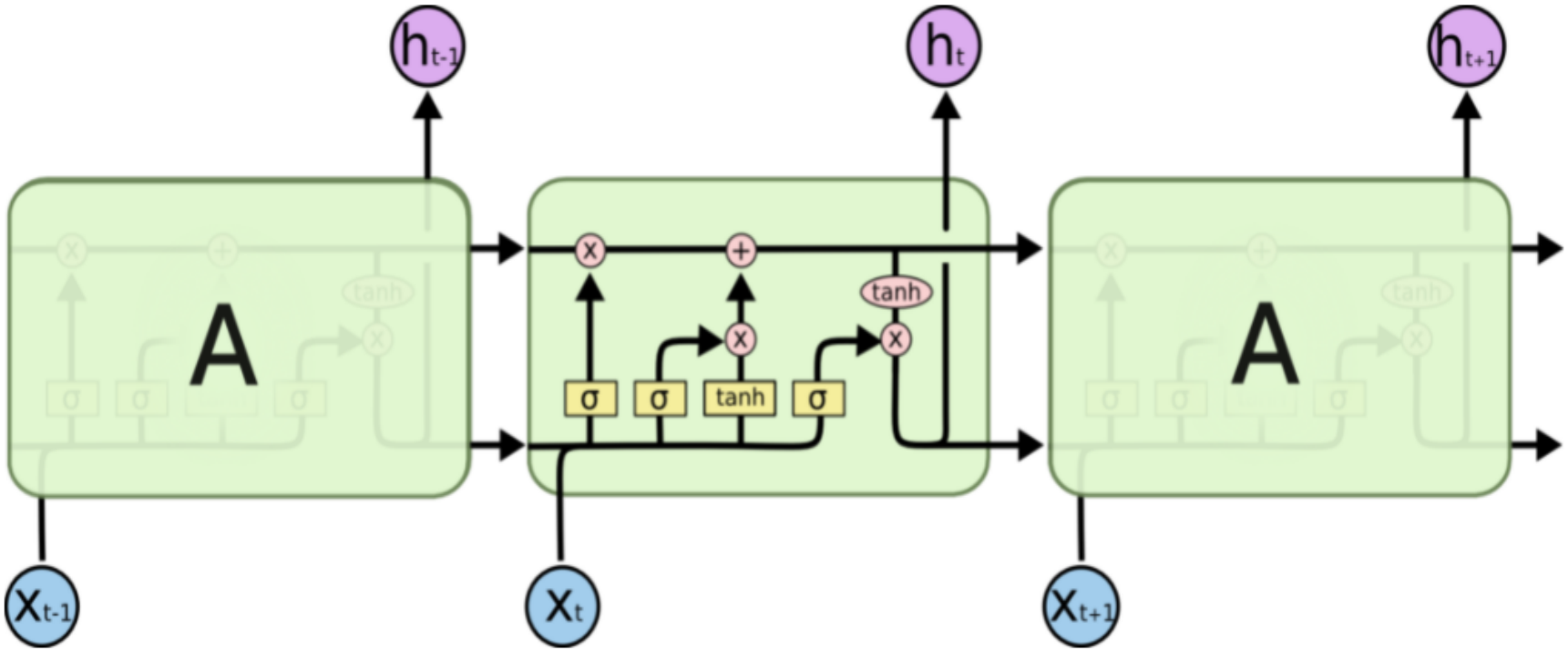


Figure: The repeating module in an LSTM contains four interacting layers.

Model Selection

Part III

- Because our object is to predict 15min future return, so we set out window size to 15 and train for 10 epochs.
- We use the Adam optimization algorithm and set the learning rate as 0.001. For the loss function, we compute the cosine similarity between labels and predictions.
- Firstly, we divide the dataset based on different cryptocurrencies. For each cell in the LSTM layer, we set a hidden state vector with size 32. After putting data into the LSTM network we do global average pooling for the output, we do these for each cryptocurrency.
- Second, we concat all the output data and put them into a 128 units linear layer, and finally get the output.

Model Selection

LSTM Workflow

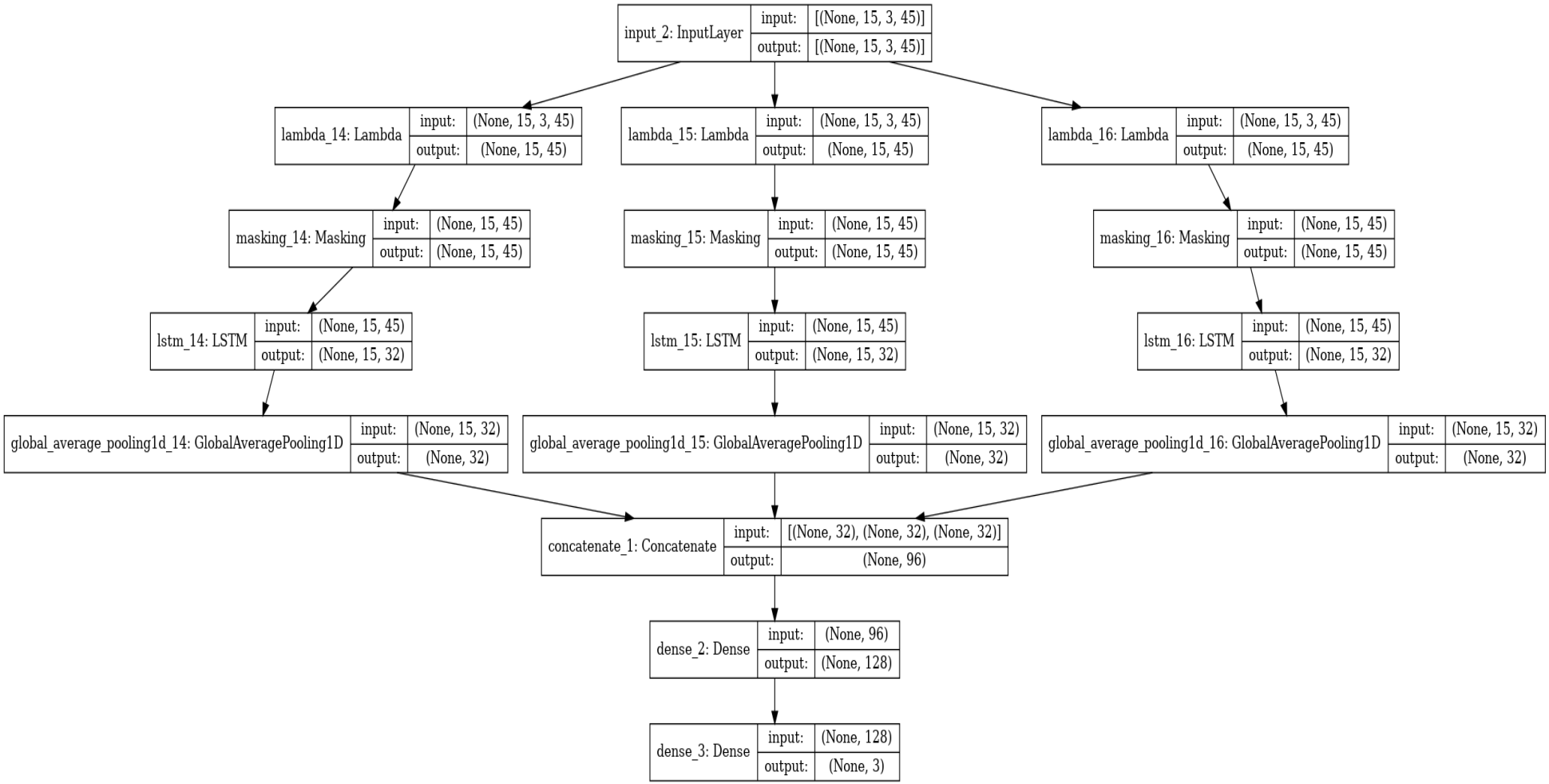


Figure: LSTM model

Model Selection

LSTM Performance

- We randomly selected 90% of the given train.csv for training and the remaining 10% for validation to judge the effectiveness of the model.
- In the competition, our predictions will be evaluated on a weighted version of the Pearson correlation coefficient. We calculate the correlation of our prediction values and true values and we get the Pearson value for each asset, shown below.

Table: LSTM's Correlation of Validation Data

Coin	Cardano	Bitcoin Cash	Binance Coin	Bitcoin	Dogecoin	EOS.IO	IOTA
Coef	0.0254	-0.0063	0.0052	0.009	-0.0009	0.0126	0.0024
Coin	Ethereum	Eth Classic	Litecoin	Maker	TRON	Stellar	Monero
Coef	0.0155	0.0111	0.0009	0.0105	0.0326	0.0070	-0.0123

Model Selection

LSTM Performance

One of the biggest problems we have met in this competition is overfitting. Even we increase training epochs the loss in the validation dataset won't decrease continuously.

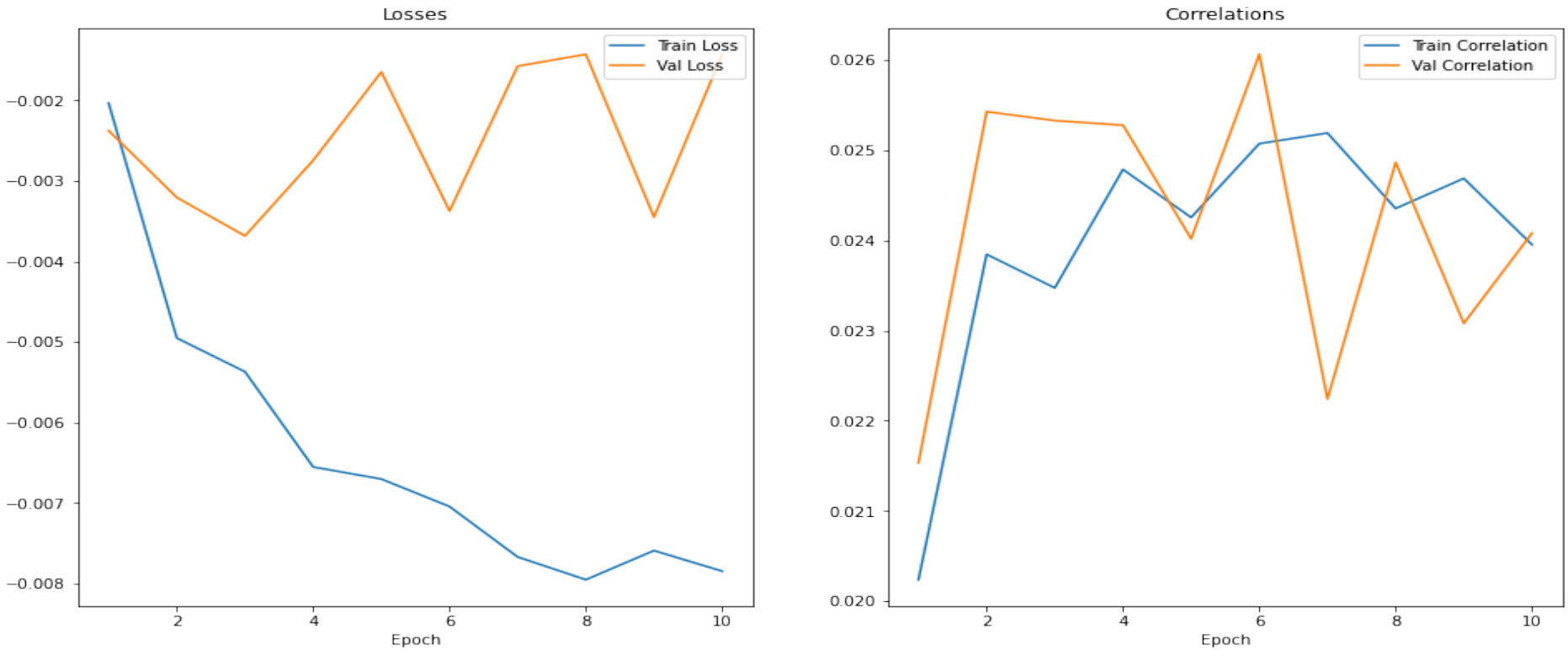


Figure: Loss & Correlation of LSTM model

Model Selection

LightGBM

- LightGBM is a fast, distributed, high performance gradient boosting framework based on decision tree algorithms open-sources by Microsoft.
- It uses a histogram-based algorithm to speed up the training process, reduce memory consumption and combine advanced network communication to optimize parallel learning, called parallel voting DT algorithm.
- Also, LightGBM uses the leaf-wise strategy to grow trees and find a leaf with the largest gain of variance to do the split.
- The library is used extensively in Kaggle competitions, and often forms part of the winning solution.

Model Selection

LightGBM Workflow

- Similarly to LSTM, we split train.csv to judge the model effect.
- We use the previously constructed factors to train the training set and use grid search for tuning the important parameter learning rate, feature fraction, num leaves and max depth.
- LightGBM can output the importance of the features, sorted in descending order as follows. It can be seen that Mean(Mean of 'Open', 'High', 'Low', 'Close') and LOGCNT(Trun Count into log value) are important for predicting cryptocurrency returns.

Model Selection

LightGBM Performance

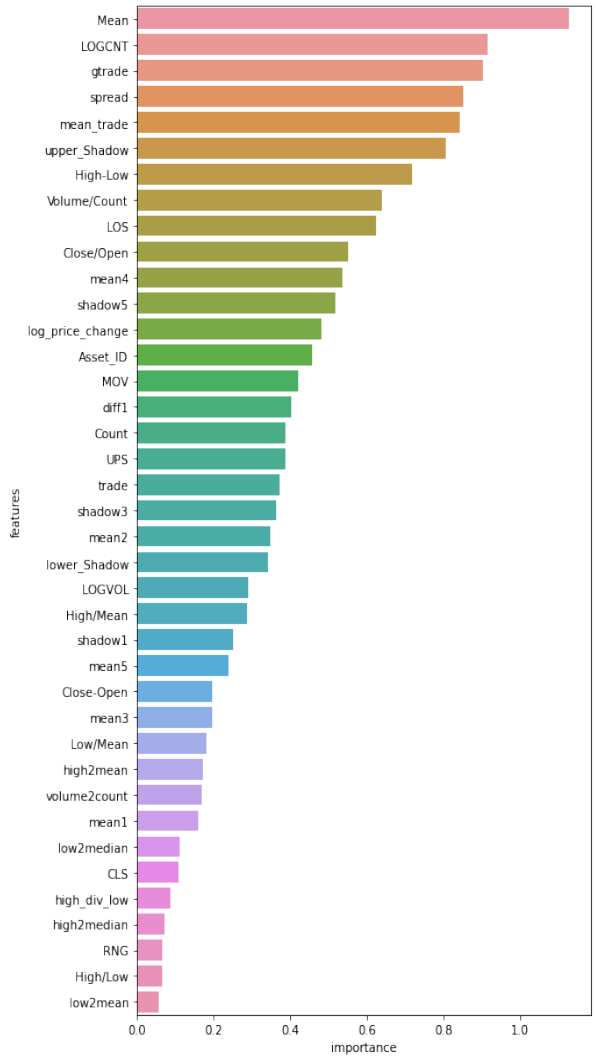


Figure: LightGBM Features Importance

Model Selection

LightGBM Performance

- We also calculate the correlation between the predicted and true values on the validation. It can be seen that for the same set of factors, LightGBM performs better on the validation set compared to LSTM.

Table: LightGBM's Correlation of Validation Data

Coin	Cardano	Bitcoin Cash	Binance Coin	Bitcoin	Dogecoin	EOS.IO	IOTA
Coef	0.0481	-0.0014	0.0174	-0.0018	0.0638	0.0006	-0.0029
Coin	Ethereum	Eth Classic	Litecoin	Maker	TRON	Stellar	Monero
Coef	0.0062	0.0104	-0.0331	0.0045	-0.0140	-0.0092	-0.0028

Conclusion

- ...

End

Thank you for watching !