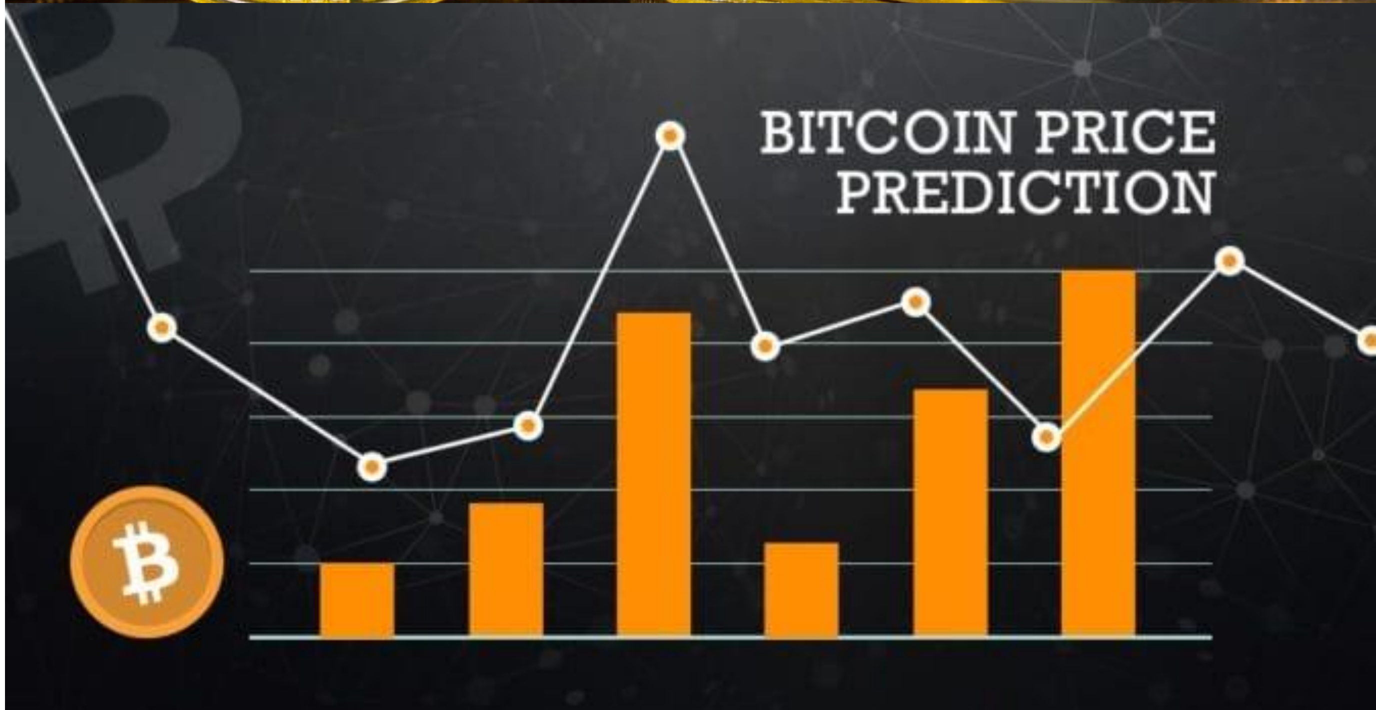
BITCOIN PRICE PREDICTION

Group Name: DataEureka

/Agenda

Bitcoin Price Prediction

- Data collection
- Data cleaning and preprocessing
- Basic LSTM Model baseline
- Features Engineering
- Modeling and optimization



/Data Input

Collected from:
Downloading open-source database
Web crawler

Collated to be: structured
sequenced in time series.

Dated Between:
2020/9/16 ~ 2021/9/17

1 Bitcoin property and network
Bitcoin Daily Average Prices, Hash Rate, Miner Rewards, Miner Reserves,...

2 Bitcoin Marketing and trading
Number of Large Transactions, Average Transaction Size, Average Balance, Average Time Between Transactions, …

3 Global economic indicators
Gold price, US dollar index, Dow Jones Commodity index, …

4 Investors and Media Attention
Google trend, Twitter positive, Twitter negative, …

5 Prices of Other Cryptocurrencies and BTC Index
Ethereum, Dogecoin, CCI30*

Data Preprocessing

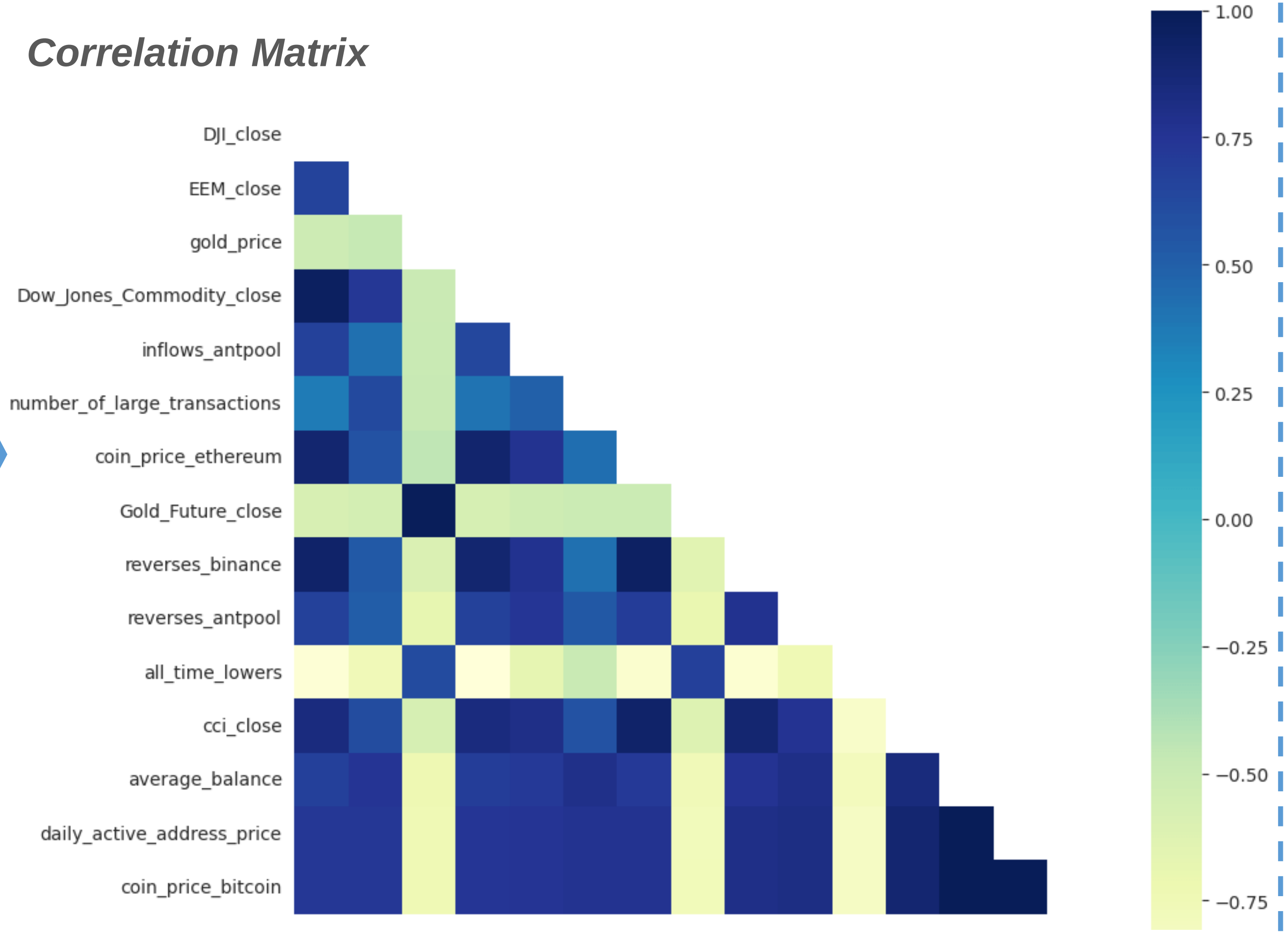
Description Analysis

Basic Statistics

Table 2
Summary statistics of features used for bitcoin daily data prediction.

Feature	Count	Mean	SD	Min	Max
Block size	740	819710.5	157884.5	361626.6	998175.2
Hash rate	740	23412238	17352387	2917084	61866256
Mining difficulty	740	3.18E+12	2.41E+12	4.22E+11	7.45E+12
Number of transactions	740	255215.8	57038.29	131875	490644
Confirmed transactions per Day	740	255694.5	57012.15	131875	490644
Mempool transaction Count	740	255215.8	57038.29	131875	490644
Mempool size	740	26489513	35357624	35369.5	1.37E+08
Market capitalization	740	9.87E+10	6.1E+10	1.53E+10	3.23E+11
Estimated transaction value	740	193095	96494.26	37558.21	629491.3
Total transaction fees	740	165.9894	192.5094	11.2287	1495.946
Google trend search volume index	740	8.452703	10.53606	2	100
Gold spot price	740	1268.682	43.20863	1174.2	1357.91

Correlation Matrix



Data Preprocessing

Data Cleaning

Imputation:

Google trend, Titter trend

Too many missing values - dropped

US index, Gold price, etc,

Impute by Friday numbers

Redundancy:

Miners related features

Drop similar miner pools

BTC price similarity

High correlation but no contribution - dropped

Data Normalization

Min-Max Scaler to normalize the data.

Data outcome

1 Date + 61 features

365 days

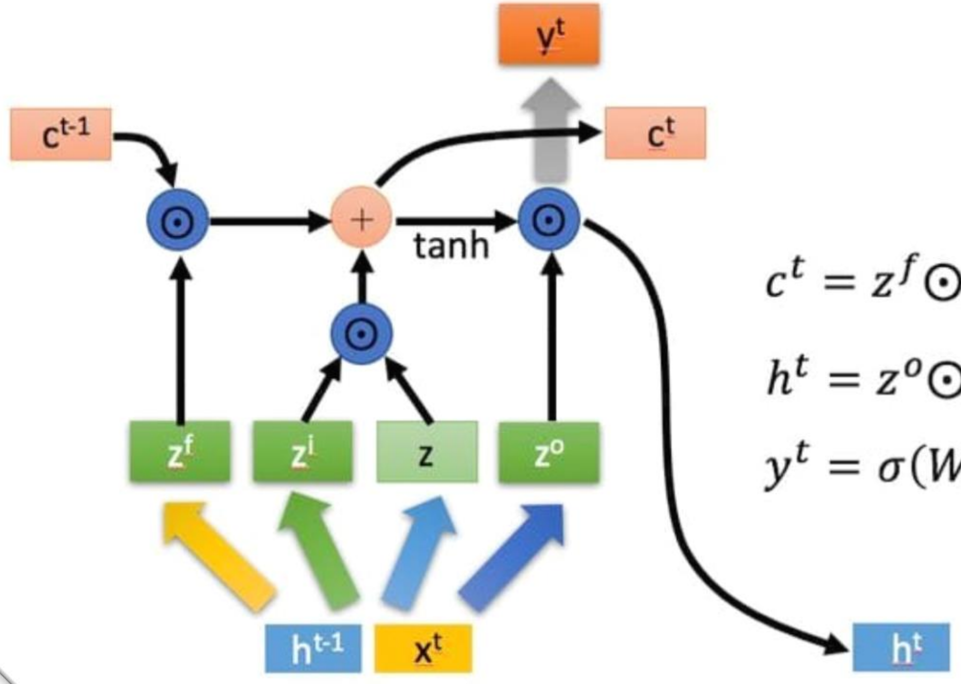
date	coin_price	coin_price	coin_price	gold_price	itb_btc_co	all_time_h	all_time_lc	average_ba	seconds	average_tr	new_addre	active_add	zero_bala
2020/9/14	0.001356	0.009181	0.000392	0.988841	0.710623	0.000547	1	0.0056	0.061118	0.011497	0.664999	0.610892	0.68680
2020/9/15	0.006322	0.009409	0.00045	0.955364	0.670118	0.000542	0.991071	0.011379	0.120176	0.017194	0.647389	0.544633	0.53474
2020/9/16	0.008669	0.007594	0.00045	1	0.711456	0.000536	0.982143	0.013402	0.133448	0.019472	0.6847	0.659941	0.67154
2020/9/17	0.008614	0.012067	0.000421	0.908207	0.720404	0.000528	0.973214	0.01426	0.05226	0.036946	0.693398	0.70955	0.80784
2020/9/18	0.009204	0.012717	0.000406	0.960763	0.637174	0.000522	0.973214	0.01462	0.089609	0.035739	0.602966	0.549028	0.57523
2020/9/19	0.010855	0.012416	0.000406	0.960763	0.511119	0.000517	0.964286	0.016945	0.156823	0.024626	0.508367	0.467749	0.52769
2020/9/20	0.009163	0.010657	0.000392	0.960763	0.379195	0.000513	0.964286	0.014599	0.173299	0.020891	0.374923	0.35941	0.45829
2020/9/21	0.004658	0.005598	0.00029	0.811375	0.540552	0.000509	0.964286	0.007973	0.170616	0.033825	0.518255	0.516646	0.5381
2020/9/22	0.000655	0.001963	0.000174	0.799136	0.579298	0.000504	0.964286	0.003234	0.250938	0.017428	0.587203	0.543495	0.53673
2020/9/23	0	0.00043	0.00016	0.681785	0.697201	0.0005	0.955357	0	0.179636	0.009643	0.651294	0.596635	0.57335
2020/9/24	0.000142	0	0.000145	0.640029	0.591855	0.000496	0.955357	0.002888	0.199046	0.010404	0.588934	0.514867	0.50629
2020/9/25	0.004474	0.003116	0.000247	0.632469	0.711966	0.000491	0.946429	0.007112	0.085222	0.013491	0.686604	0.751532	0.81521
2020/9/26	0.005392	0.004504	0.00029	0.632469	0.437533	0.000487	0.919643	0.009596	0.146375	0.017273	0.467656	0.52495	0.63746
2020/9/27	0.005436	0.005169	0.000276	0.632469	0.371016	0.000482	0.919643	0.008865	0.135582	0.013419	0.380915	0.33488	0.41899
2020/9/28	0.008466	0.006639	0.000276	0.649028	0.506284	0.000478	0.910714	0.010377	0.234674	0.012491	0.50202	0.431259	0.43798
2020/9/29	0.005877	0.005475	0.000247	0.719942	0.685825	0.000471	0.910714	0.008471	0.198352	0.01043	0.683722	0.582928	0.5490
2020/9/30	0.005973	0.005571	0.000203	0.730382	0.721326	0.000465	0.910714	0.008782	0.12623	0.006756	0.679639	0.716747	0.75612
2020/10/1	0.00612	0.006483	0.000203	0.784737	0.570968	0.00046	0.901786	0.006543	0.273161	0.004758	0.589387	0.496573	0.48293
2020/10/2	0.001964	0.002619	0.000116	0.788697	0.603142	0.000456	0.901786	0.002451	0.200644	0.016	0.585198	0.562241	0.55454
2020/10/3	0.00222	0.003201	0.000145	0.788697	0.536489	0.000452	0.892857	0.003585	0.148429	0.007137	0.536164	0.594588	0.66958
2020/10/4	0.003445	0.003598	0.000174	0.788697	0.406186	0.000444	0.892857	0.004501	0.118195	0.007412	0.426324	0.486353	0.59671
2020/10/5	0.005113	0.004481	0.000174	0.812095	0.505466	0.00044	0.892857	0.007824	0.190222	0.007319	0.515128	0.515302	0.54391
2020/10/6	0.005115	0.003528	0.000145	0.825774	0.595834	0.000435	0.892857	0.006272	0.242991	0.006708	0.596814	0.551526	0.53783
2020/10/7	0.00361	0.001308	0.000116	0.721742	0.743837	0.000429	0.883929	0.004767	0.132894	0	0.702004	0.652144	0.66467
2020/10/8	0.006005	0.002476	0.000131	0.732541	0.617524	0.000425	0.883929	0.007841	0.112618	0.002869	0.630755	0.624564	0.65402
2020/10/9	0.010335	0.005686	0.000189	0.861411	0.663186	0.00042	0.883929	0.012778	0.183096	0.015369	0.635882	0.585608	0.61011
020/10/10	0.016917	0.009907	0.000247	0.861411	0.497164	0.000416	0.883929	0.01959	0.176657	0.005338	0.531724	0.481136	0.54248
020/10/11	0.017411	0.009821	0.000247	0.861411	0.359694	0.000412	0.883929	0.021528	0.190111	0.018265	0.395151	0.343165	0.41988
020/10/12	0.019042	0.011345	0.000232	0.86933	0.552677	0.000404	0.883929	0.024039	0.138839	0.010086	0.549632	0.510616	0.57321
020/10/13	0.019179	0.01219	0.000218	0.74622	0.666387	0.000393	0.883929	0.023116	0.10008	0.009595	0.67607	0.647587	0.69936
020/10/14	0.01833	0.011644	0.000189	0.813895	0.586465	0.000388	0.883929	0.022305	0.136119	0.014247	0.574315	0.507193	0.53225
020/10/15	0.018452	0.010777	0.000145	0.74838	0.686771	0.000381	0.883929	0.022308	0.153186	0.014845	0.678822	0.626231	0.65819
020/10/16	0.017529	0.008936	0.000102	0.795896	0.642938	0.000376	0.883929	0.020666	0.13207	0.011896	0.630752	0.565477	0.58912
020/10/17	0.017145	0.008361	0.000102	0.795896	0.432153	0.000371	0.883929	0.019475	0.121282	0.021079	0.455306	0.402591	0.46358
020/10/18	0.018707	0.009838	0.000116	0.795896	0.277694	0.000368	0.883929	0.02137	0.279954	0.023837	0.33182	0.264805	0.32376
020/10/19	0.021526	0.010941	0.000116	0.797696	0.550715	0.000359	0.883929	0.024847	0.281394	0.037398	0.583848	0.499682	0.47781

Long short-term memory(LSTM)

Purpose

capable of learning long-term dependencies

use prior experience to inform future outcomes



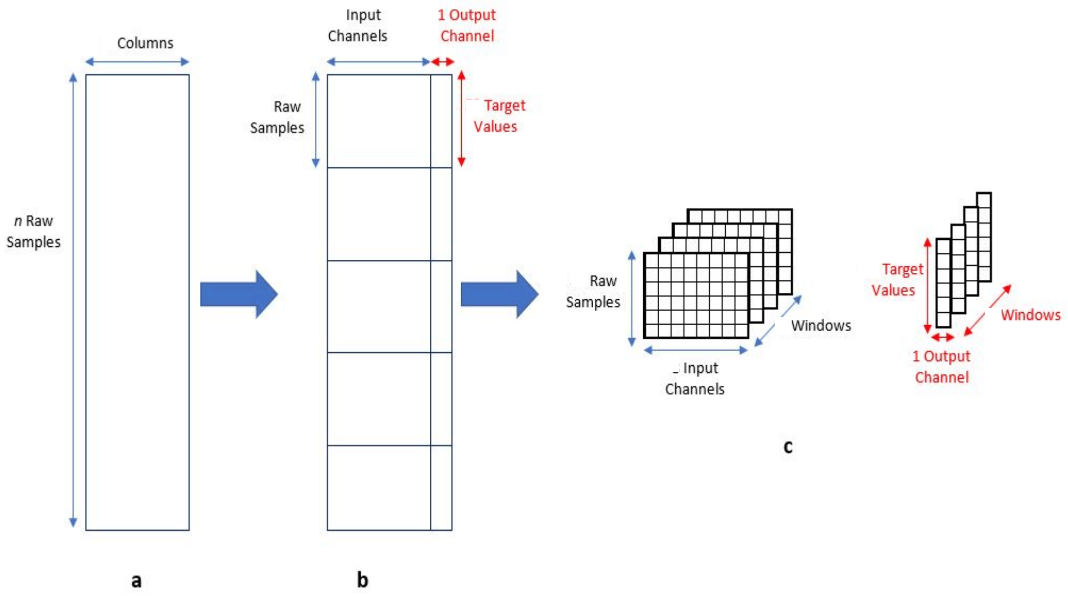
$$c^t = z^f \odot c^{t-1} + z^i \odot z$$

$$h^t = z^o \odot tanh(c^t)$$

$$y^t = \sigma(W' h^t)$$

Data Preprocessing

Extract data windows

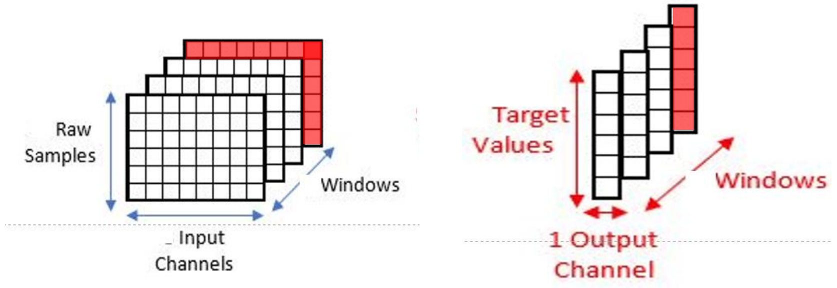


We decided to combine data of a certain time window as input to predict bitcoin price of next day.

After experiments, we found that the time window of 10 would be a reasonable choice.

Train Test Split

We used a **ratio of 0.2** to split train and test sets. So the data of last 60+ days was used for testing.



X_train, X_test   *_train, Y_test*

Original data shape
X: (366 * 61)   Y: (366 * 1)
After concatenating
X: (366 * 10 * 61)   Y: (366 * 1)
After splitting
X_train: (300 * 10 * 61)   X_test: (66 * 10 * 61)
Y:_train: (300 * 1)   Y_test: (66 * 1)

Model Evaluation

train and test the model 5 rounds

calculate their average RMSE as a indicator of its performance

fixed the random seed in each round to ensure it can be reproduced

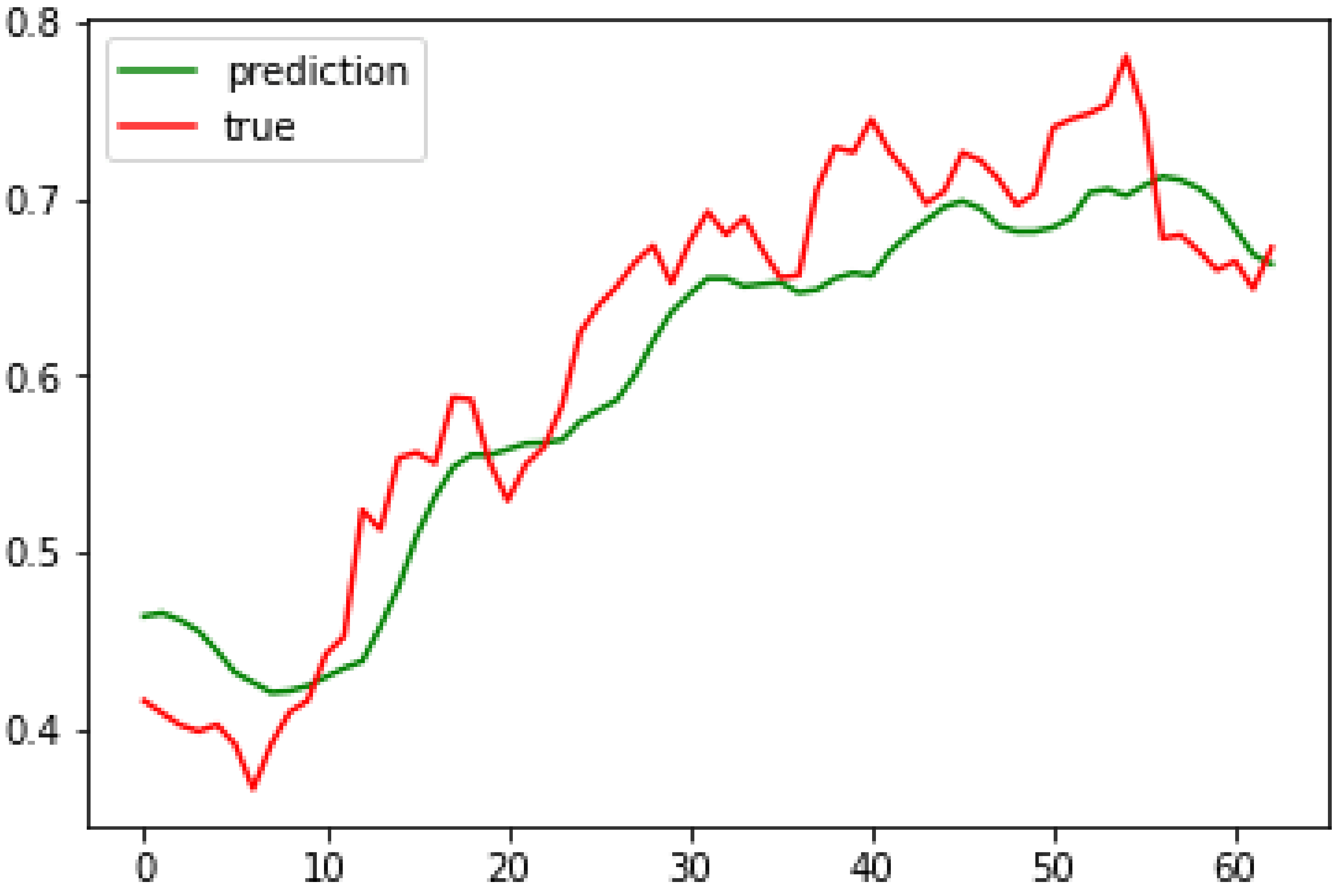
```
Training epoch  0 MSE:  0.3618645668029785
Training epoch  20 MSE:  0.018359770998358727
Training epoch  40 MSE:  0.0027389577589929104
Training epoch  60 MSE:  0.003316698130220175
Training epoch  80 MSE:  0.0028881970793008804
train 5 rounds in total, this is the 0 th round, RMSE is0.043214183300733566

Training epoch  0 MSE:  0.34949031472206116
Training epoch  20 MSE:  0.009240277111530304
Training epoch  40 MSE:  0.0054707834497094154
Training epoch  60 MSE:  0.0029721681494265795
Training epoch  80 MSE:  0.002721790922805667
train 5 rounds in total, this is the 1 th round, RMSE is0.04568817466497421

Training epoch  0 MSE:  0.33345314860343933
Training epoch  20 MSE:  0.015268449671566486
Training epoch  40 MSE:  0.004078330937772989
Training epoch  60 MSE:  0.0033113863319158554
Training epoch  80 MSE:  0.0027421447448432446
train 5 rounds in total, this is the 2 th round, RMSE is0.047878626734018326

Training epoch  0 MSE:  0.2620851397514343
Training epoch  20 MSE:  0.014127722941339016
Training epoch  40 MSE:  0.0031828766223043203
Training epoch  60 MSE:  0.002784544602036476
Training epoch  80 MSE:  0.002620649291202426
train 5 rounds in total, this is the 3 th round, RMSE is0.04254891350865364

Training epoch  0 MSE:  0.22142408788204193
Training epoch  20 MSE:  0.006789344362914562
Training epoch  40 MSE:  0.0036627124063670635
Training epoch  60 MSE:  0.002691515488550067
Training epoch  80 MSE:  0.00261324574239552
train 5 rounds in total, this is the 4 th round, RMSE is0.04782906174659729

average RMSE is 0.04543179199099541
```

Baseline Model

First Attempt

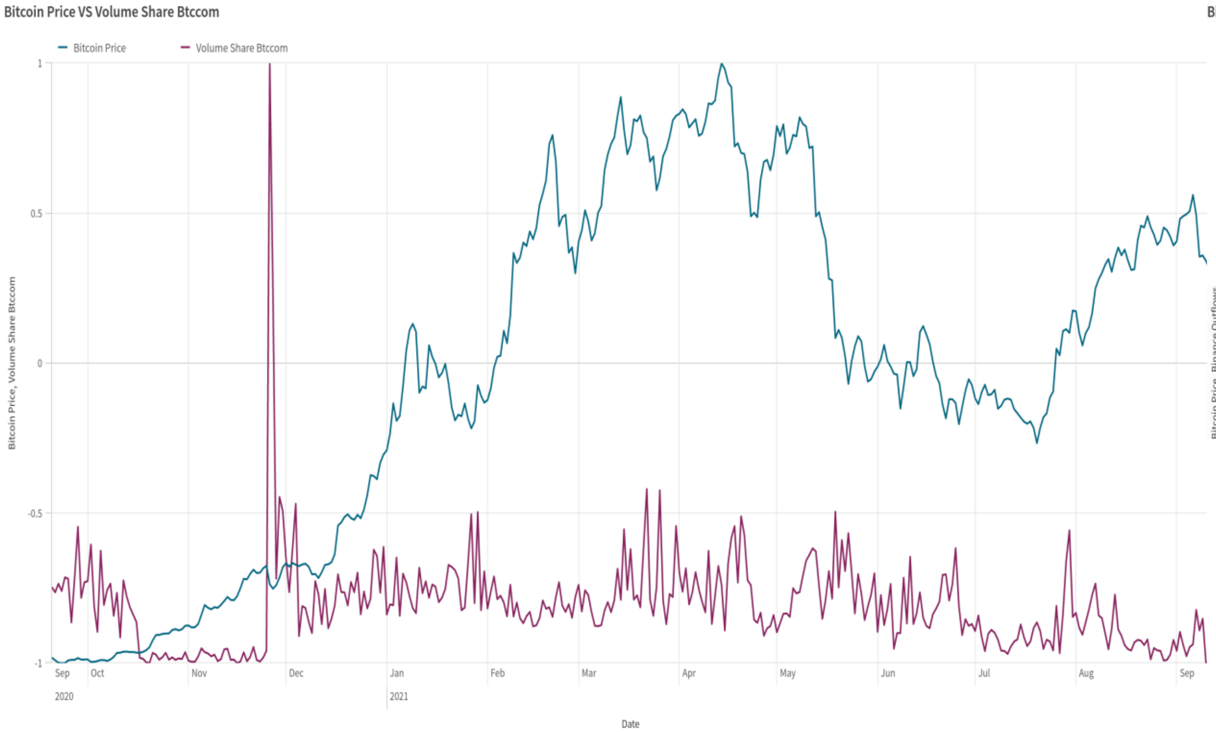
All features are used for modeling

AvgMSE=0.0454

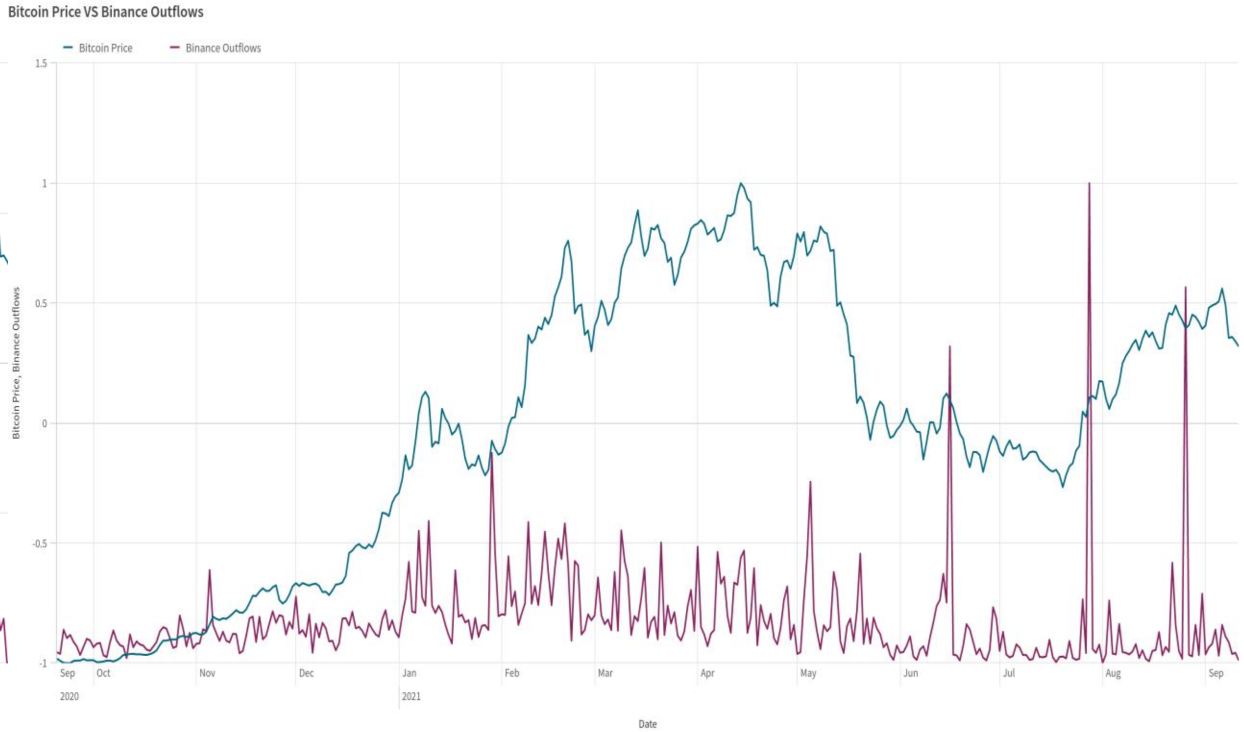


Features Engineering

1. Noise → Need criteria to drop **columns** with low significance!



BTC Price V.S. Volume share Btccom



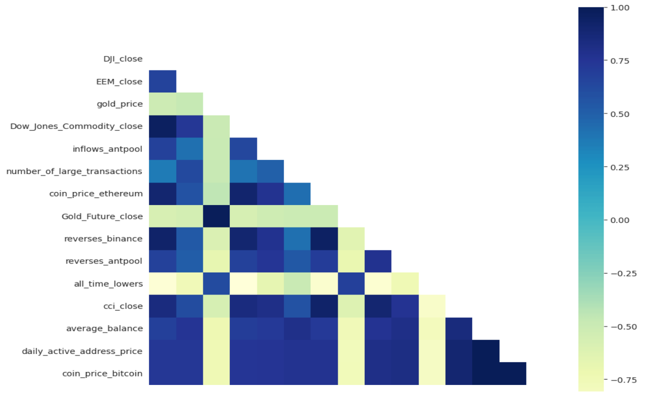
BTC Price V.S. Binance outflows

Features Engineering

Feature selection

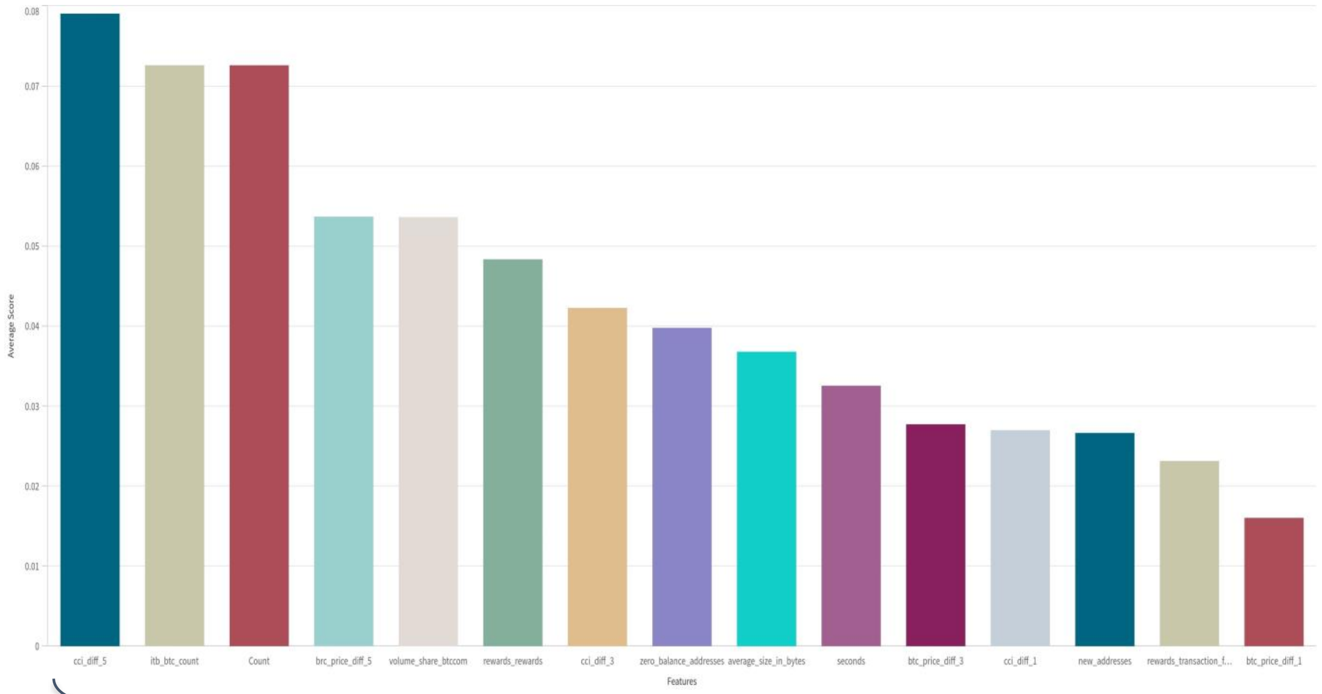
3 criteria

1. Ridge regression for feature selection
2. Random forest regressor feature importance
3. Correlation with bitcoin price



Features scoring

Calculate the average of three scores results after scaling.



...

Drop features with lowest scores ['rewards_transaction_fees', 'new_addresses','seconds','average_size_in_bytes']

Features Engineering

2. Add Lagging Effect Features

We added the 1-day, 3-day, 5-day return rate data into our features to enrich our feature set.

- 1-day lag Bitcoin Price return rate:

$$Return_{1-day} = \frac{Price_t - Price_{t-1}}{Price_{t-1}}$$

- 3-day lag Bitcoin Price return rate:

$$Return_{1-day} = \frac{Price_t - Price_{t-3}}{Price_{t-3}}$$

- 5-day lag Bitcoin Price return rate:

$$Return_{1-day} = \frac{Price_t - Price_{t-5}}{Price_{t-5}}$$

Features Engineering

3. Add Lagging Effect Features

We also found that Crypto Currency Index (CCI) is an important feature of our model, so we added its change rate for 1-day, 3-day, 5-day.

- 1-day lag CCI increase rate:

$$\widehat{CCI}_{1-day} = \frac{CCI_t - CCI_{t-1}}{CCI_{t-1}}$$

- 3-day lag CCI increase rate:

$$\widehat{CCI}_{3-day} = \frac{CCI_t - CCI_{t-3}}{CCI_{t-3}}$$

- 5-day lag CCI increase rate:

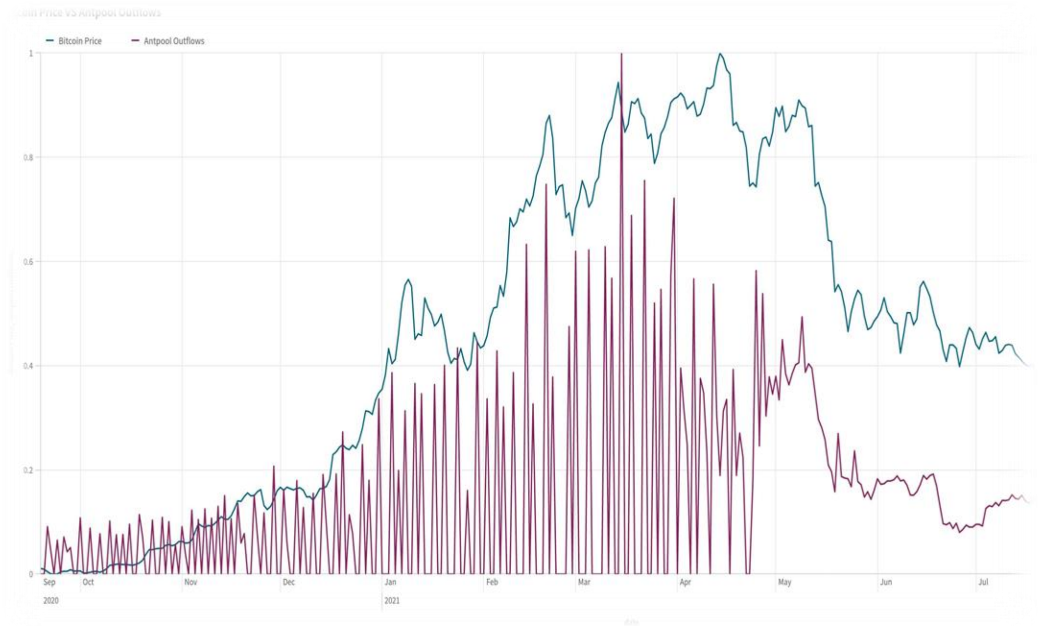
$$\widehat{CCI}_{5-day} = \frac{CCI_t - CCI_{t-5}}{CCI_{t-5}}$$

https://cci30.com/

/Features Engineering

High Volatility Features

- Visualized High Volatility Features



Bitcoin Price VS Antpool Outflows

Bitcoin Price VS Volume Share Antpool

Good trend!

But too high Volatility!!!

Need Smoothing

Features Engineering

After Smoothing…

Bitcoin Price VS Antpool Outflows After Smoothing

Model Tuning

Tuning Model Architecture	Memory Cells
	Hidden Layers
	Weight Initialization
Tuning Learning Behavior	Learning Rate
	Optimization Algorithm
	Regularization

In Our Case：

- 1. Hidden size and epochs size are set low to prevent over-fitting
- 2. Used GridSearchCV to find the best number of layers, learning rate, and regularization.

- Hidden Size:64,
- Epochs Size:100
- Learning Rate:0.01
- weight decay for Adam: 0.01
- Regularization:0.01

Final Model Performance

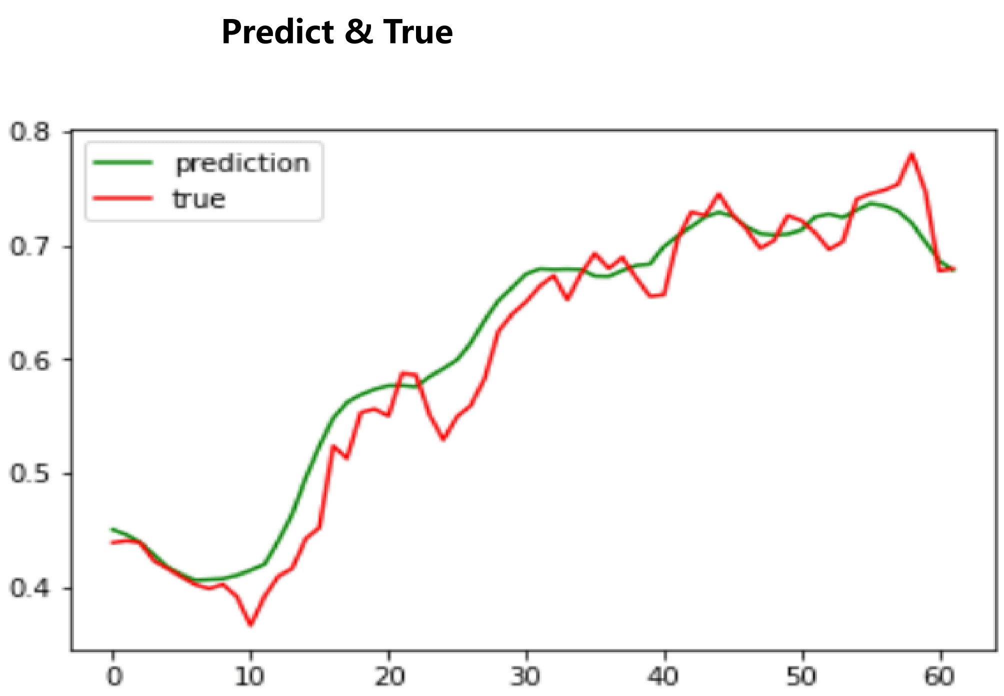
- **1、 Performance:**

RMSE = **0.02631**.
not over-fitted

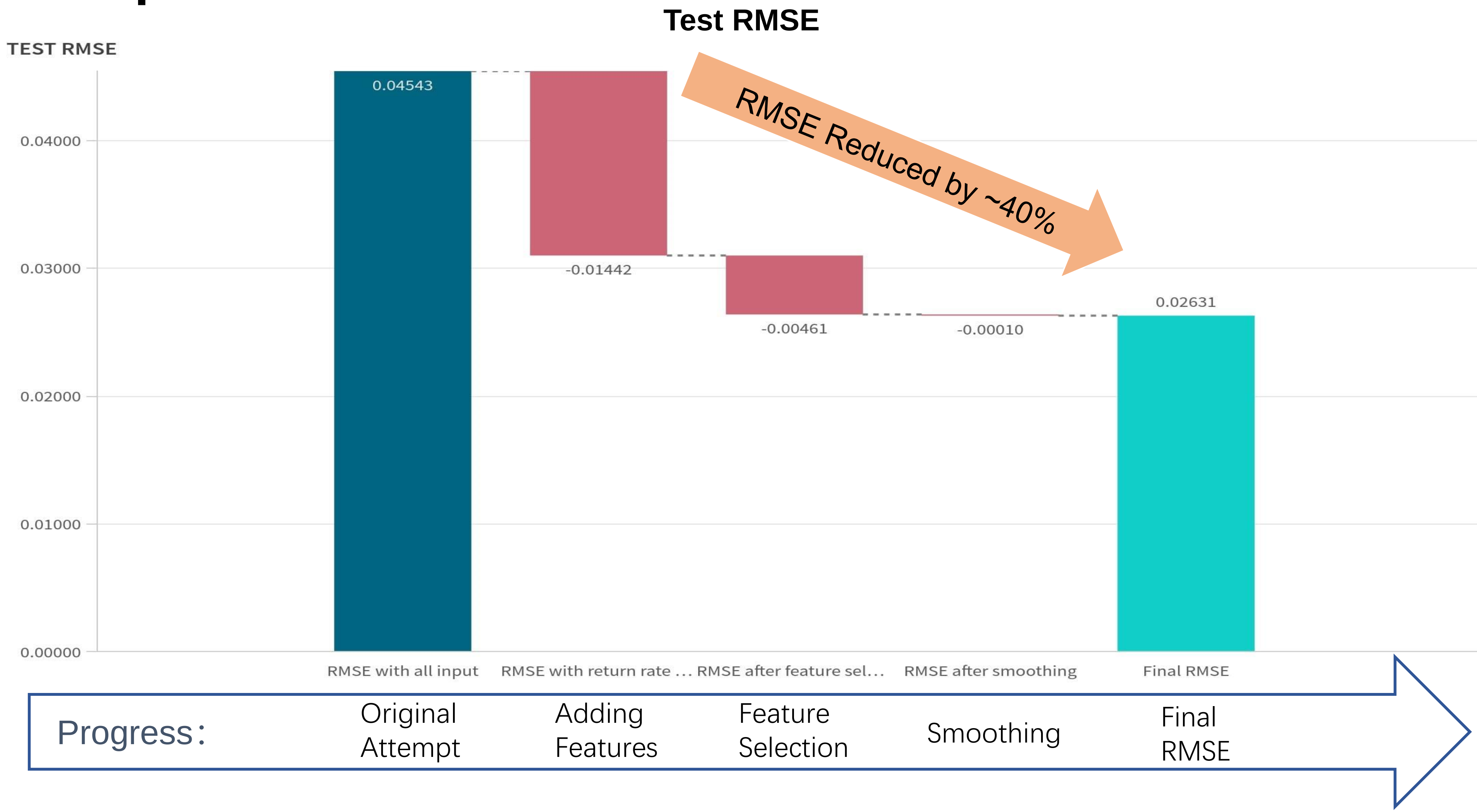
- **2、 Results:**

last 10 days data
to predict
BTC price of the next day that is not in the data set

Price for 2021.9.15
Predicted: 45858.684
True price: 48150.90.



Model Improvement



Model Limitation

1

average error is low while the error of each day is sometimes significant

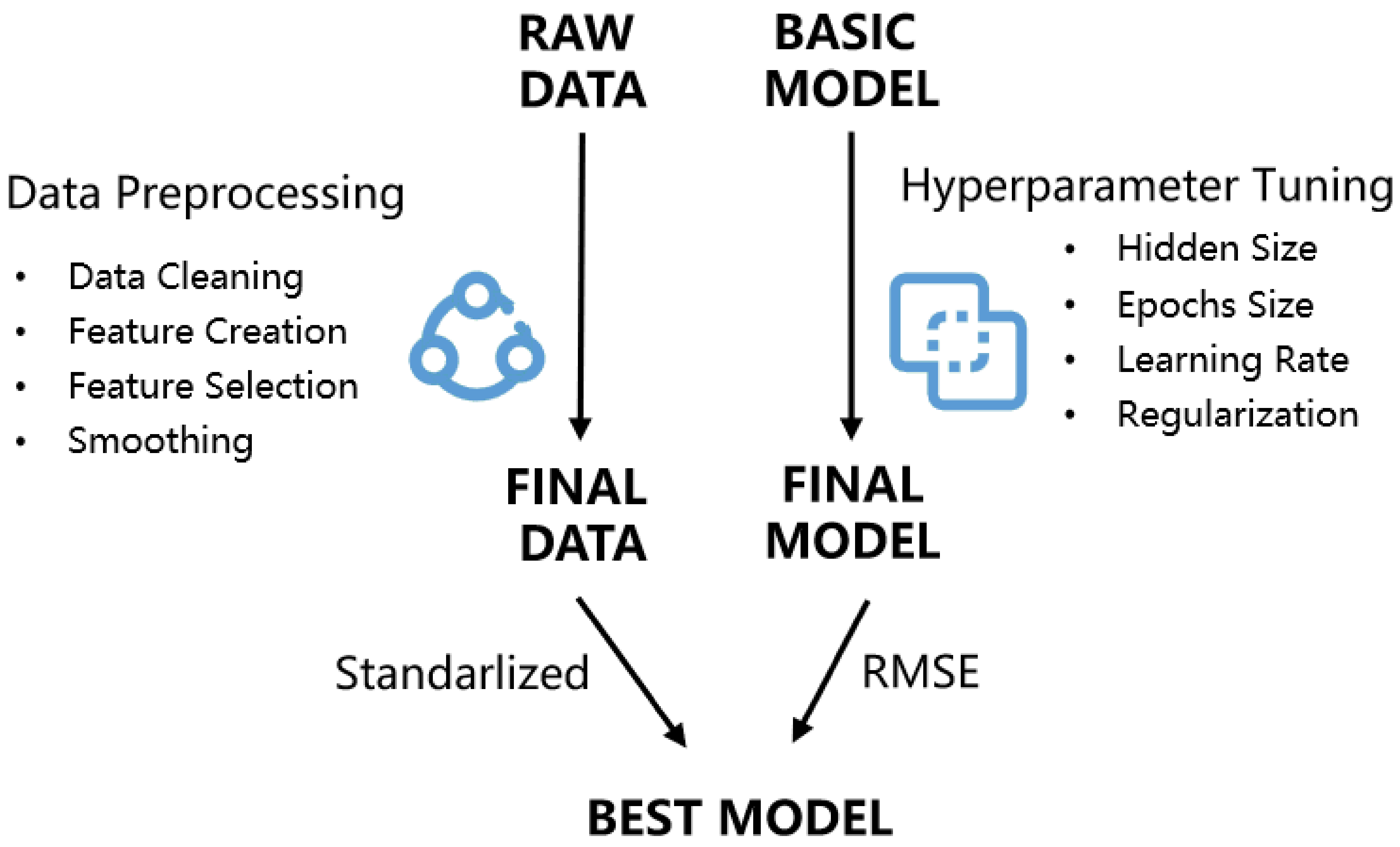
2

LSTM can only be used for predicting very near future

3

small dataset
no validation seperation
hence have risk for overfit

/ Conclusion


RAW DATA
BASIC MODEL
Data Preprocessing
• Data Cleaning
• Feature Creation
• Feature Selection
• Smoothing
Hyperparameter Tuning
• Hidden Size
• Epochs Size
• Learning Rate
• Regularization
FINAL DATA
FINAL MODEL
Standarlized
RMSE
BEST MODEL

REFERENCE

1. Chen, Zheshi et al. “Bitcoin price prediction using machine learning: An approach to sample dimension engineering.” J. Comput. Appl. Math. 365 (2020): n. pag.

2. Raju, Shobha Manival and Ali Mohammad Tarif. “Real-Time Prediction of BITCOIN Price using Machine Learning Techniques and Public Sentiment Analysis.” ArXiv abs/2006.14473 (2020): n. pag.

3. Bollen, Johan et al. “Twitter mood predicts the stock market.” ArXiv abs/1010.3003 (2011): n. pag.

4. Dutta, Aniruddha et al. “A Gated Recurrent Unit Approach to Bitcoin Price Prediction.” arXiv: Pricing of Securities (2019): n. pag.

Group member

Ke Ma	A0212524U
Xiao liang	A0232007X
Mingzhe Xu	A0232022A
Shaobin Qiao	A0232196E
Yishun Liu	A0231849X
Yunyi Gao	A0231927A

Thank you!