




Crypto Market Forecasting

13기 김창현
14기 김종민
14기 임형우



목차

1. 중간 이전 스터디 설명
2. 대회소개
3. 평가지표
4. EDA
5. 모델링
6. 개선사항 및 추후 일정
7. 결론



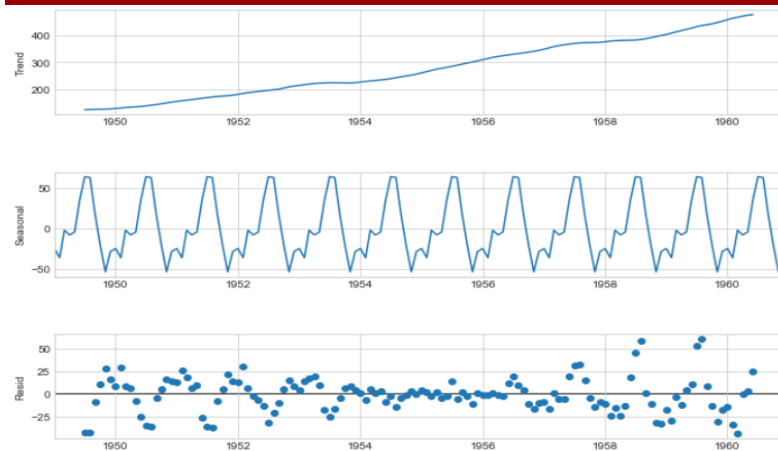
중간 이전 스터디 내용

- 1차 (21.09.26) : 시계열 데이터 분석 기본, 지수 평활법
- 2차 (21.10.03) : AR, MA, ARIMA, SARIMA
- 3차 (21.10.10) : 불균형 데이터 처리, SVM

김성범 교수님의 유튜브 자료 활용



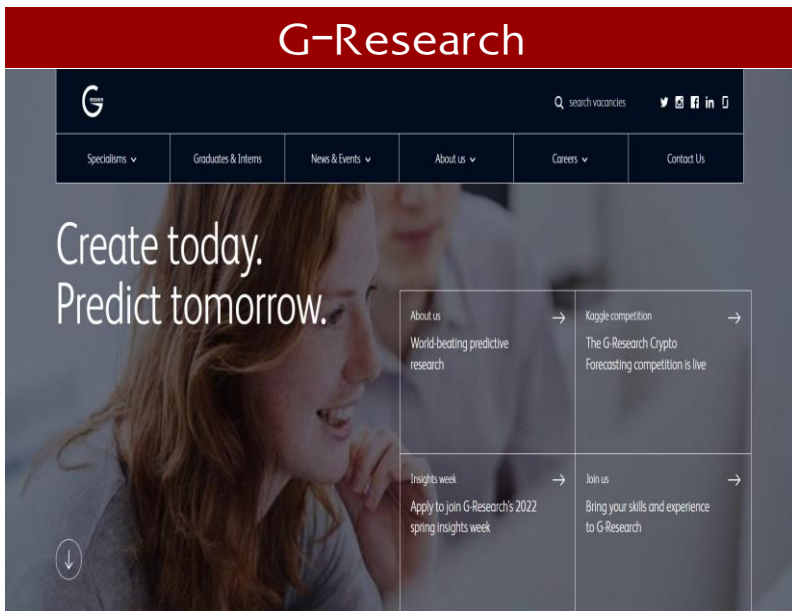
실습 진행



대회소개

- G-Research는 런던에 위치한 금융공학 연구소
- 14개의 유명한 암호화폐의 가격을 예측하는 대회

G-Research



The image shows the top section of the G-Research website. It features a red header with the company logo and name. Below the header is a navigation bar with links for Specialisms, Graduates & Interns, News & Events, About us, Careers, and Contact Us. The main content area has a large image of a woman looking thoughtful, with the text 'Create today. Predict tomorrow.' overlaid. To the right of the image is a grid of four links: 'About us', 'Kaggle competition', 'Insights week', and 'Join us', each with a right-pointing arrow.

Kaggle Competition



The image shows the Kaggle Competition page for G-Research. The banner at the top features the competition title 'G-Research Crypto Forecasting' and the prize money '\$125,000'. Below the banner is a navigation bar with links for Overview, Data, Code, Discussion, Leaderboard, Rules, Team, My Submissions, and Submit Predictions. The main content area is divided into two columns. The left column contains a table with four rows: Description, Evaluation, Timeline, and Code Requirements. The right column contains the competition details, including a description of the task, the prize money, and the deadline.

Overview	
Description	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. Could some of these price movements have been predicted in advance?
Evaluation	In this competition, you'll use your machine learning expertise to forecast short term returns in 14 popular cryptocurrencies. We have amassed a dataset of millions of rows of high-frequency market data dating back to 2018 which you can use to build your model. Once the submission deadline has passed, your final score will be calculated over the following 3 months using live crypto data as it is collected.
Timeline	The simultaneous activity of thousands of traders ensures that most signals will be transitory, persistent alpha will be exceptionally difficult to find, and the danger of overfitting will be considerable. In addition, since 2018, interest in the cryptomarket has exploded, so the volatility and correlation structure in our data are likely to be highly non-stationary. The successful contestant will pay careful attention to these considerations, and in the process gain valuable insight into the art and science of financial forecasting.
Prizes	G-Research is Europe's leading quantitative finance research firm. We have long explored the extent of market prediction possibilities, making use of machine learning, big data, and some of the most advanced technology available. Specializing in data science and AI education for workforces, Cambridge Spark is partnering with G-Research for this competition. Watch our introduction to the competition below:
Code Requirements	



평가지표

- 대회 목표는 15분 후 가격을 예측하는 것 (log returns over 15 minutes)
- 암호화폐는 서로의 상관관계가 매우 높음, 따라서 개별 자산의 가격을 예측하기 위해 시장 수익률(weighted average market returns)을 제외하고 예측

Evaluation Metric

Prediction targets and evaluation

This forecasting competition aims to predict returns in the near future for prices P^a , for each asset a . For each row in the dataset, we include the target for prediction, Target. Target is derived from log returns (R^a) over 15 minutes.

$$R^a(t) = \log(P^a(t+16) / P^a(t+1))$$

Target

Crypto asset returns are highly correlated, following to a large extent the overall crypto market. As we want to test your ability to predict returns for individual assets, we perform a linear residualization, removing the market signal from individual asset returns when creating the target. In more detail, if $M(t)$ is the weighted average market returns, the target is:

$$M(t) = \frac{\sum_a w^a R^a(t)}{\sum_a w^a}$$
$$\beta^a = \frac{\langle M \cdot R^a \rangle}{\langle M^2 \rangle}$$

$$\text{Target}^a(t) = R^a(t) - \beta^a M(t)$$

where the bracket $\langle . \rangle$ represent the rolling average over time (3750 minute windows), and same asset weights w^a used for the evaluation metric.



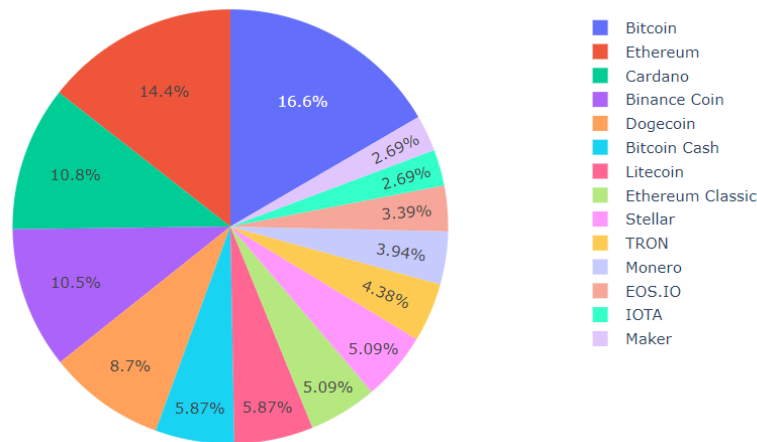
EDA

- 비트코인, 이더리움 등 14개의 유명한 암호화폐의 가격을 예측하는 대회
- 시가총액 등의 요소를 고려하여 자산 별로 가중치를 부과함

G-Research

Asset_ID	Weight	Asset_Name	weight_percentage
1	6.779922	Bitcoin	16.584998
6	5.894403	Ethereum	14.418848
3	4.406719	Cardano	10.779686
0	4.304065	Binance Coin	10.528574
4	3.555348	Dogecoin	8.697068
2	2.397895	Bitcoin Cash	5.865715
9	2.397895	Litecoin	5.865715
7	2.079442	Ethereum Classic	5.086716
12	2.079442	Stellar	5.086716
13	1.791759	TRON	4.382990
11	1.609438	Monero	3.936996
5	1.386294	EOS.IO	3.391144
8	1.098612	IOTA	2.687418
10	1.098612	Maker	2.687418

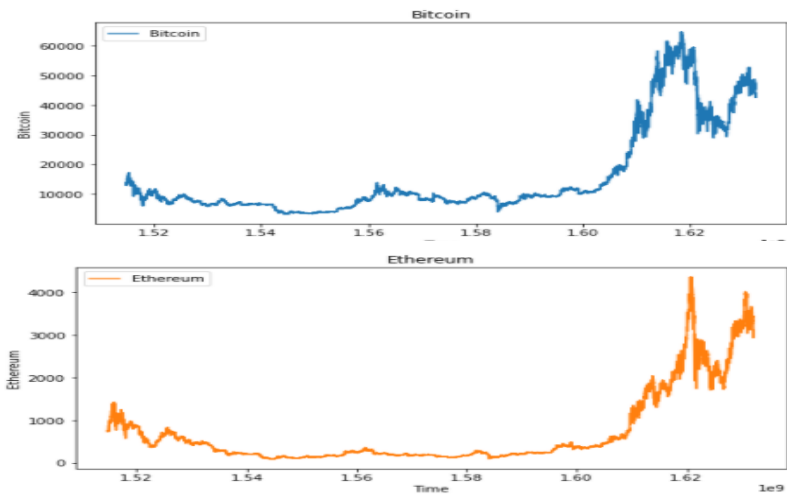
전체 중 개별 자산이 차지하는 비율



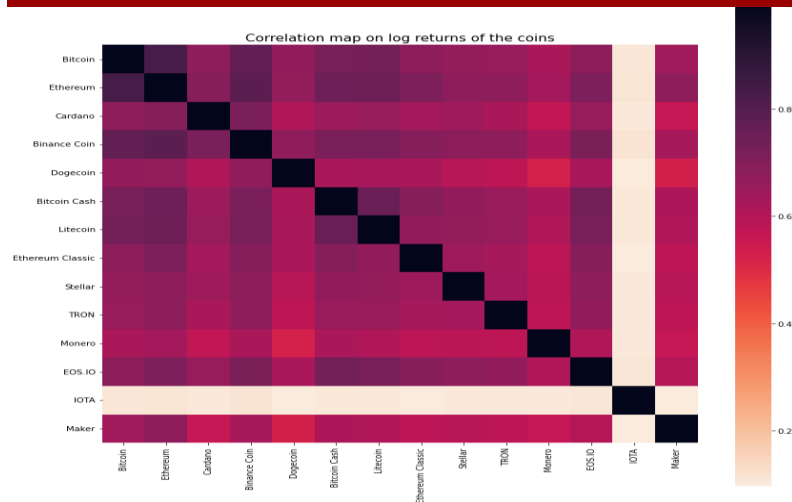
EDA

- 대표적인 암호화폐인 비트코인과 이더리움의 가격 변화
- 가장화폐는 전반적으로 높은 상관관계를 보이고 있음

Visualization



Correlation Plot



모델1: ARIMA

- 자기회귀 누적 이동평균 모델
- 아래와 같이 증가 예측에 좋은 성능을 나타냄

SARIMAX Modeling

```
parameters      aic
3      (1, 0) -446.108741
1      (0, 1) -445.320464
6      (2, 0) -444.145513
4      (1, 1) -444.129700
2      (0, 2) -442.735646
```

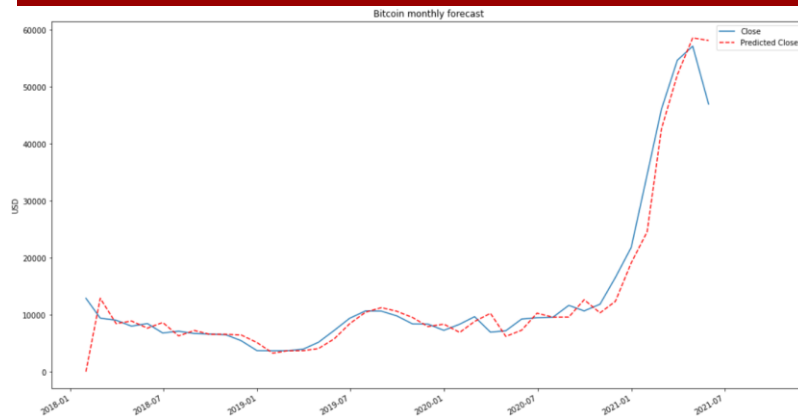
SARIMAX Results

Dep. Variable:	close_box	No. Observations:	45
Model:	SARIMAX(1, 1, 0)	Log Likelihood	225.054
Date:	Sun, 26 Dec 2021	AIC	-446.109
Time:	15:16:22	BIC	-442.540
Sample:	01-31-2018	HQIC	-444.785
	- 09-30-2021		
Covariance Type:	opg		

	coef	std err	z	P> z	[0.025	0.975]
ar.L1	0.3821	0.133	2.874	0.004	0.121	0.643
sigma2	2.099e-06	4.32e-07	4.856	0.000	1.25e-06	2.95e-06

Ljung-Box (L1) (Q): 0.02 Jarque-Bera (JB): 3.57
Prob(Q): 0.90 Prob(JB): 0.17
Heteroskedasticity (H): 0.32 Skew: -0.65
Prob(H) (two-sided): 0.03 Kurtosis: 3.51

Predicted Close



모델2: Boosting

- XGBRegressor
- MSE

RandomSearch

```
params = {  
    'n_estimators': [50, 100],  
    'min_child_weight': [4, 5],  
    'subsample': [i/10.0 for i in range(6, 11)],  
    'max_depth': [2, 3, 4, 6, 7],  
    'booster': ['gbtree', 'gblinear'],  
    'eval_metric': ['rmse'],  
    'eta': [i/10.0 for i in range(3, 6)],  
}  
  
reg = XGBRegressor(nthread=-1)
```

Mean Square Error

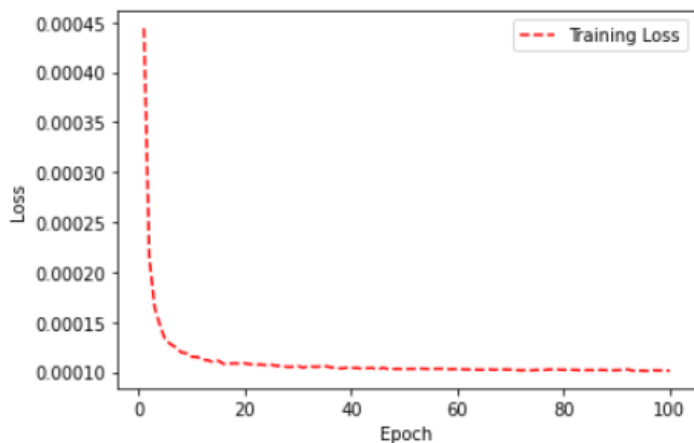
```
0.004238179249426806 0.0049390172739860305 0.0025165882548923205  
0.005079723405628494 0.006558070666150133 0.0029010544120279245  
0.005590263647039461 0.005107588600470915 0.004878439869055079  
0.005203426527239624 0.004893873597870384 0.009508390497024509  
0.005799805105563594 0.009212997343334134
```



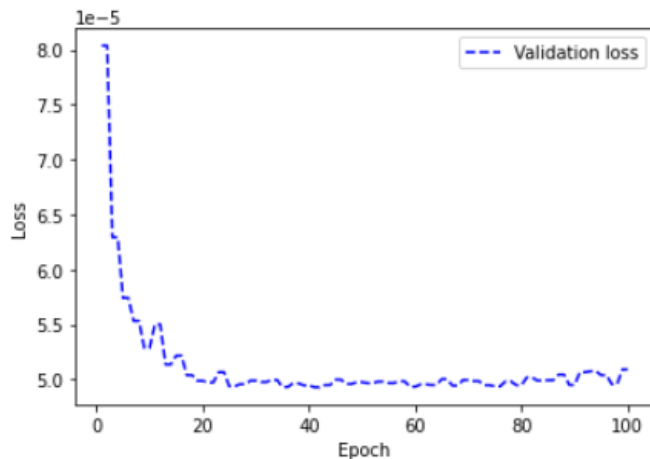
모델:3 LSTM

- 비트코인, 이더리움 등 14개의 유명한 암호화폐의 가격을 예측하는 대회
- 시가총액 등의 요소를 고려하여 자산 별로 가중치를 부과함

Training Loss Plot



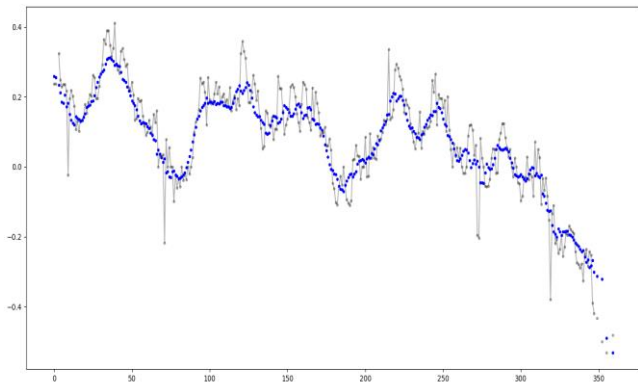
Validation Loss Plot



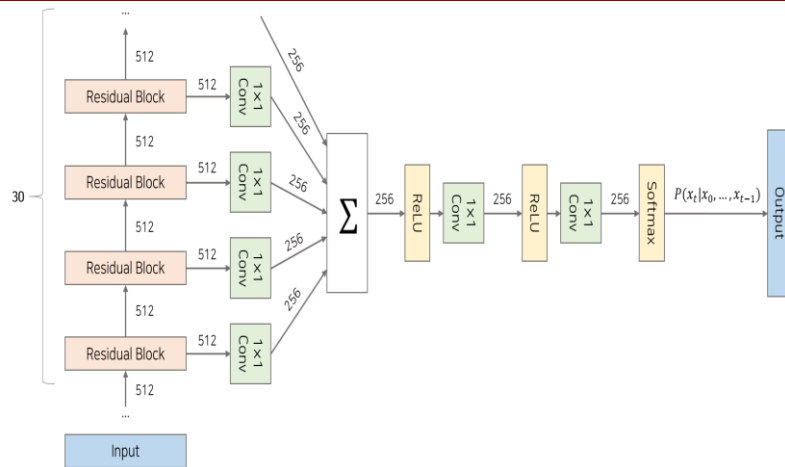
개선 사항 및 추후 계획

- Feature Engineering을 통해 더 많은 금융 Feature 생성 및 활용
- WaveNet, GRU 등 고성능의 다른 모델링 기법 사용해서 성능 비교

Moving Average



WaveNet



Thank You

