Cryptocurrency Price Predictions

Project Goals

- Predict daily cryptocurrency prices using XGBoost and LSTM
- Analyze which factors are the most significant in predicting the target variable
- Use linear regression model from classical finance (Fama-French model) to test importance of a feature

Exploratory Data Analysis

Dataset

Goal: Predict the daily closing prices of **Bitcoin** and **Ethereum** (top 2 coins by market capitalization).

Historical cryptocurrency data was scraped from coinmarketcap using the cryptocmd library.

Dataset consisted of 6 features: **Daily open, high, low, close prices**, **volume** and **market capitalization**.

Historical price data was collected over all available period.

Bitcoin data: 2013/04/27 - 2023/07/19
Ethereum data: 2015/09/09 - 2023/07/19

Data Visualization

Data Wrangling

Problems and issues ----Missing Value:

Chosen Method ----Linear interpolation

Nature of Data: It's time series data ,they usually follow a trend. Linear interpolation respects this trend

Amount of Data: There are only 1183 rows of data per coin, eliminating any rows with missing values is not a viable option because it would significantly reduce our already small training set

Volatility of cryptocurrencies:Cryptocurrencies are highly volatile. Filling in missing data with the most frequent values could make the data unusually predictable for this kind of volatile data and potentially introduce artificial regularities.

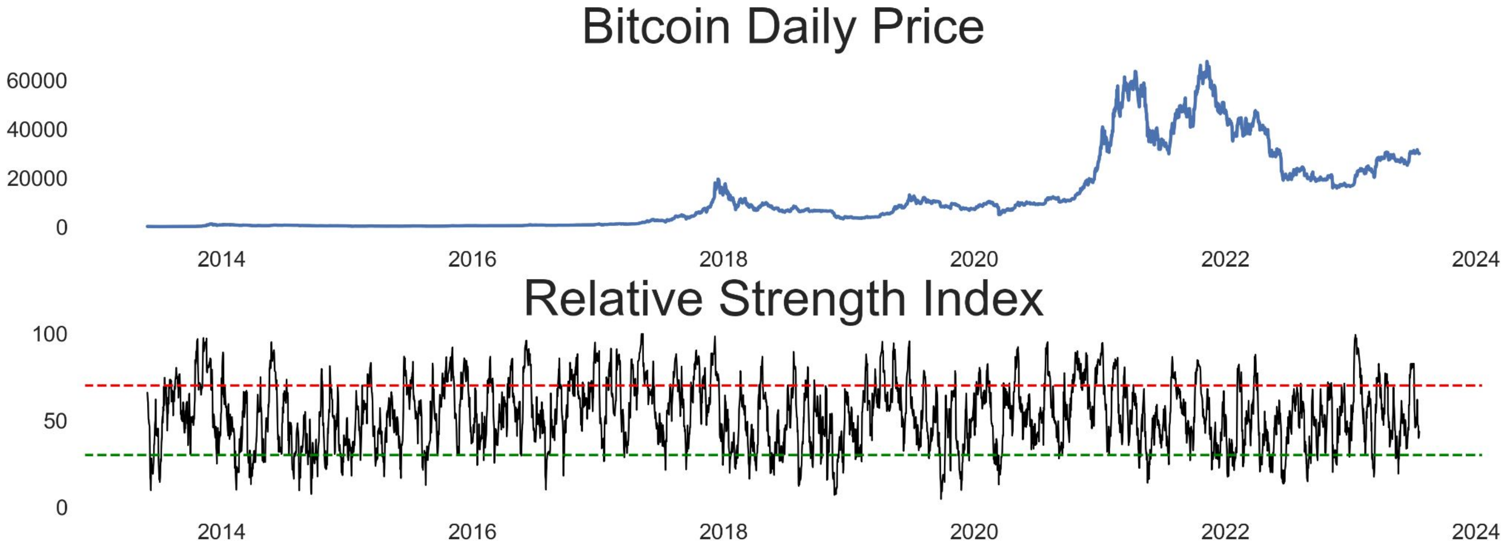
Before/After Data:

Additional Features

Day of the week: 1=Monday . . . 7=Sunday
Month: 1=January . . .12=December
Year: 2013 - 2023
Lags: 1, 7, and 14-day price lags
3-Day Simple Moving Average (SMA): Average price over the past 3 days
Moving Average Convergence/Divergence (MACD): Momentum oscillator used to trade trends, defined as 12 period EMA - 26 period EMA
Relative Strength Indicator (RSI): Technical indicator that signals upward or downward price momentum

Relative Strength Index (RSI)

Relative strength index provides signals about upward or downward price momentum.
An asset is usually considered overbought when RSI > 70 and oversold when RSI < 30.
RSI is computed using the expression: $100 - \left[\frac{100}{1+\frac{\text{Average gain}}{\text{Average loss}}}\right]$

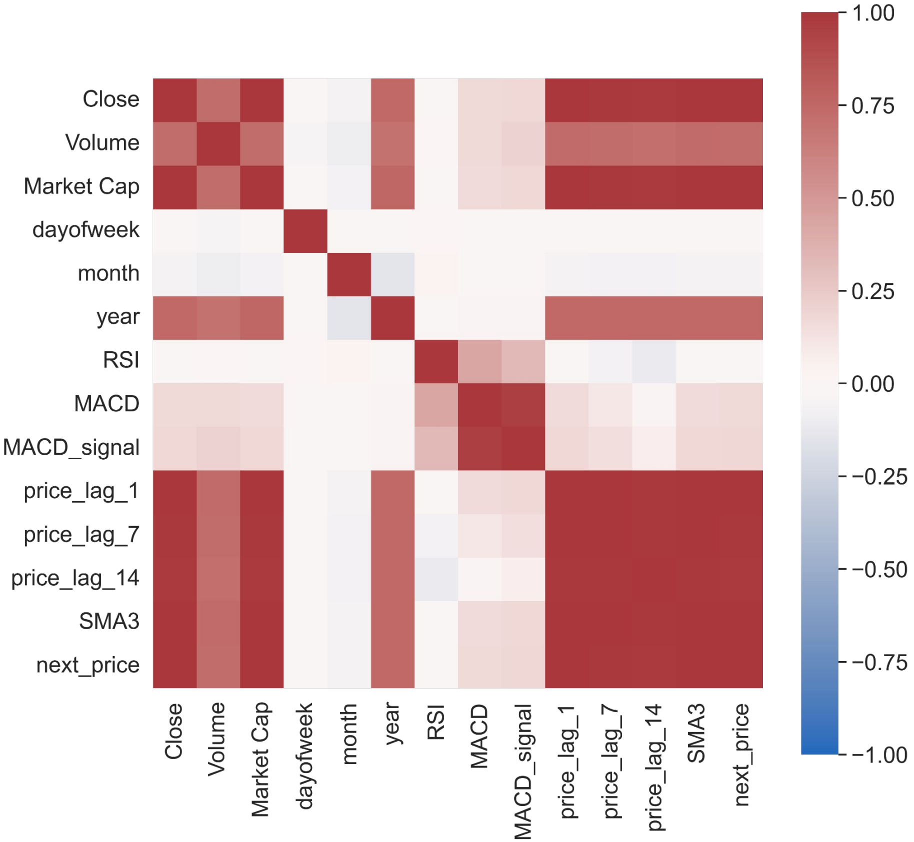


Target Variable

Our goal to predict the following day's price, one day at a time, given each day's closing price along with the other features.

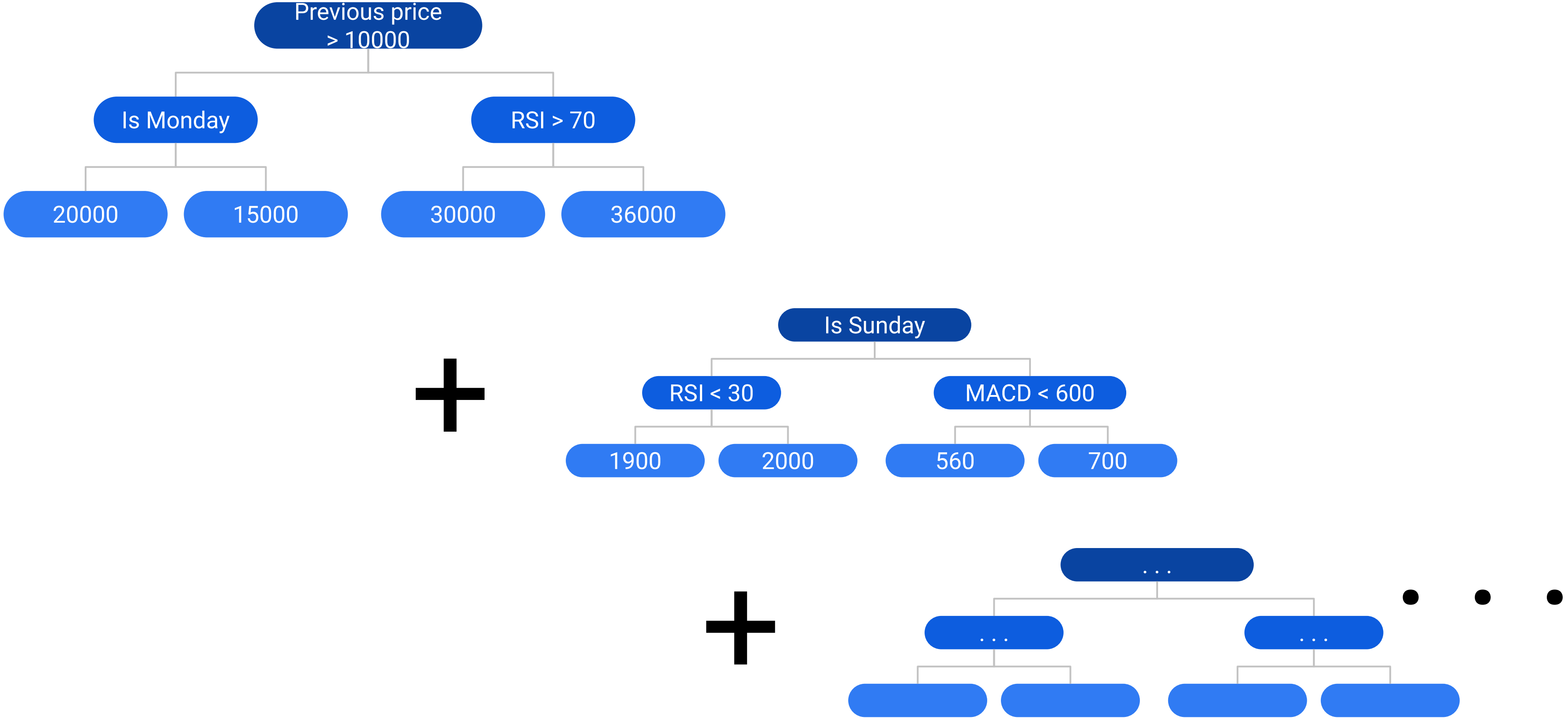
	Close	Volume	Market Cap	dayofweek	month	year	RSI	MACD	MACD_signal	price_lag_1	price_lag_7	price_lag_14	SMA3	next_price
Date														
2023-07-10	30414.472	14828209155.420	590834747693.020	1	7	2023	51.713	646.862	790.353	30171.234	31156.439	30271.131	30292.749	?

Correlation Heatmap



Making Predictions with Machine Learning Models

XGBoost


Previous price > 10000
Is Monday
RSI > 70
20000
15000
30000
36000
Is Sunday
RSI < 30
MACD < 600
1900
2000
560
700
. . .
. . .
. . .

Training XGBoost

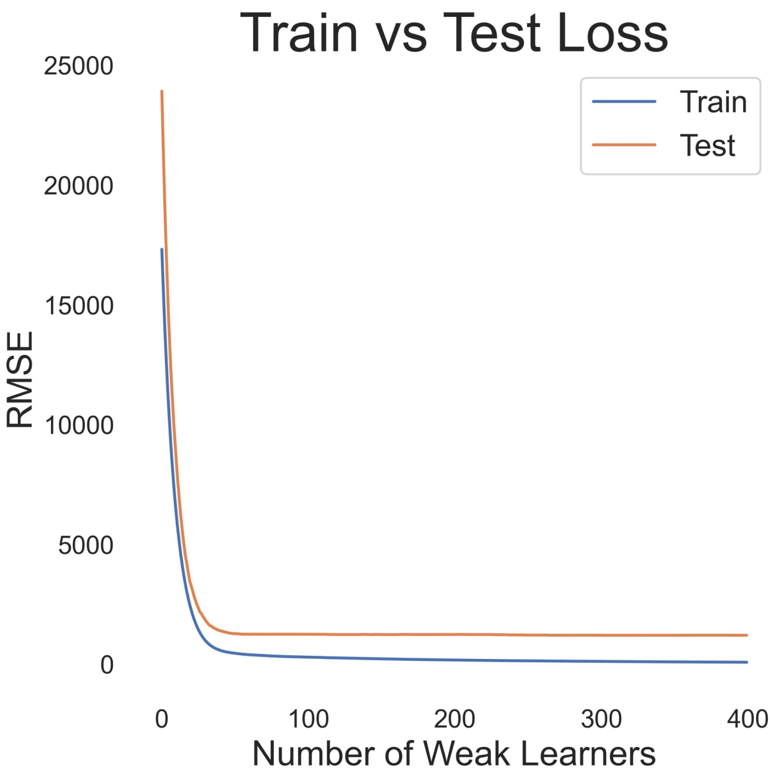
Train/test data split: 95%/5%

Hyperparameter tuning: randomized search

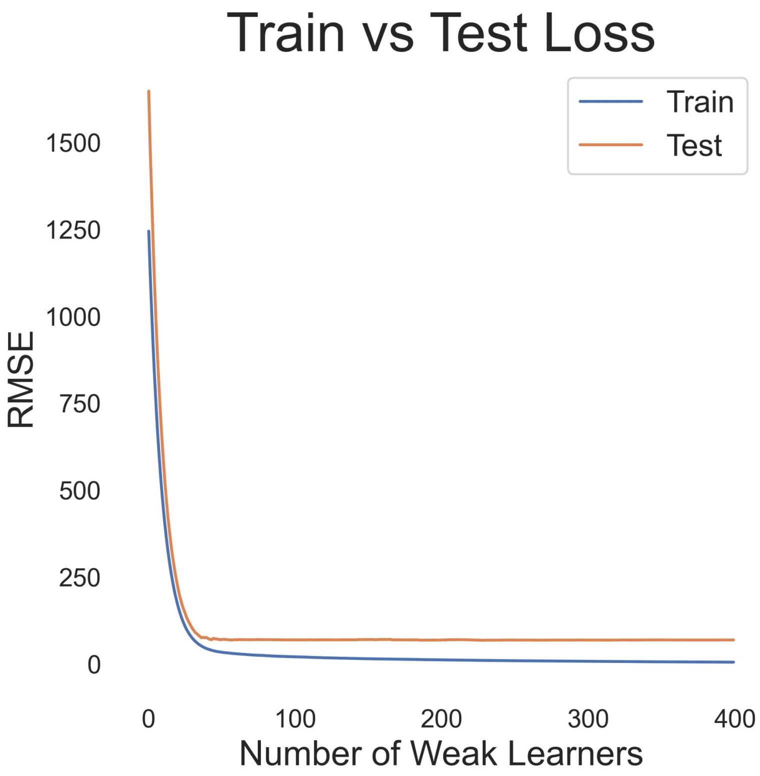
Best hyperparameters:

```
booster='gbtree'
objective='reg:squarederror'
learning_rate=0.1
n_estimators=100
```

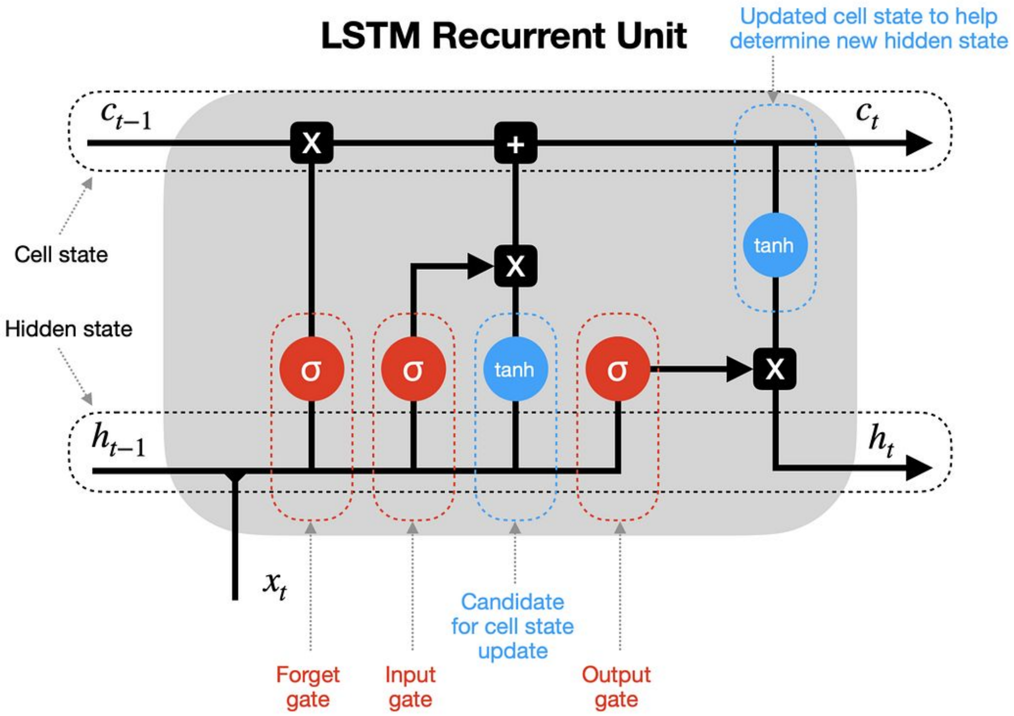
Bitcoin

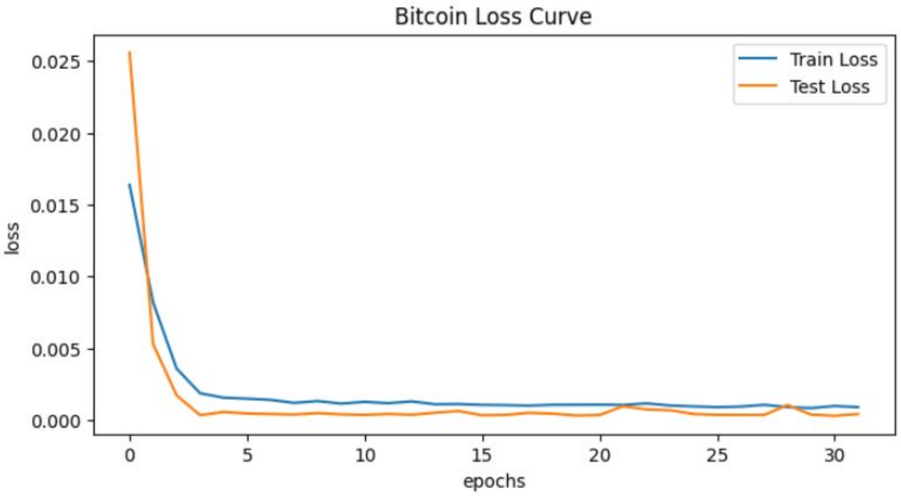


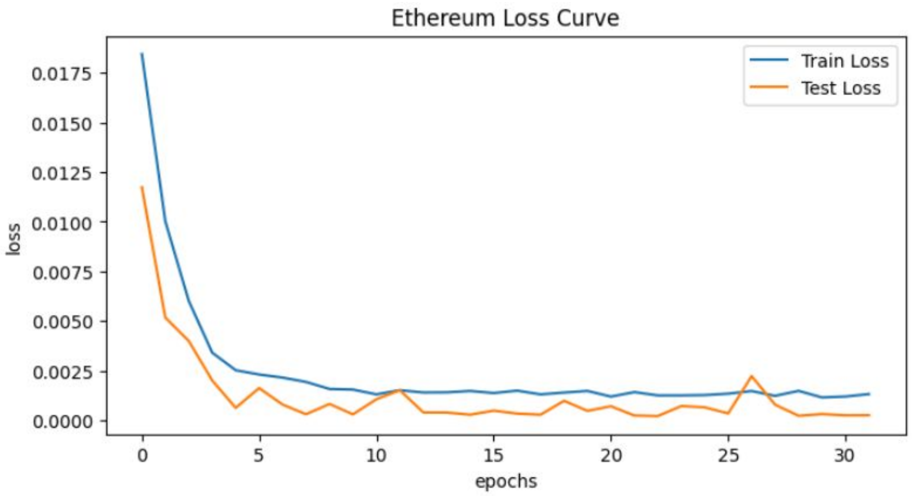
Ethereum



LSTM (Long Short-Term Memory)



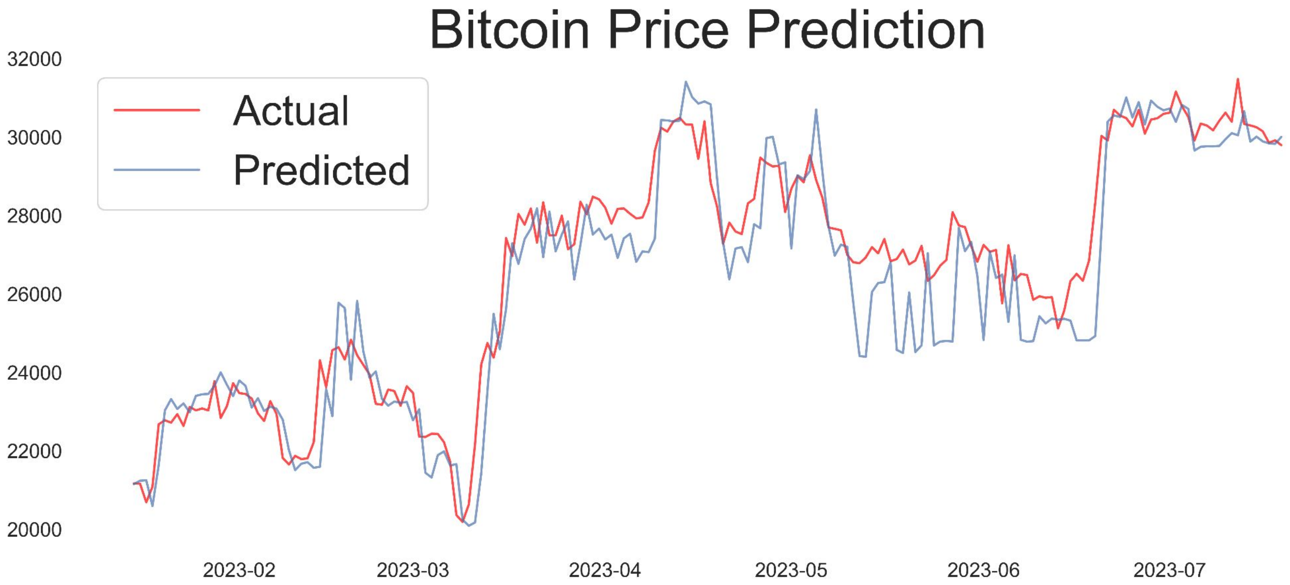

Bitcoin Loss Curve
Train Loss
Test Loss
0.025
0.020
0.015
0.010
0.005
0.000
loss
0
5
10
15
20
25
30
epochs


Ethereum Loss Curve
Train Loss
Test Loss
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
loss
0
5
10
15
20
25
30
epochs

Prediction Results

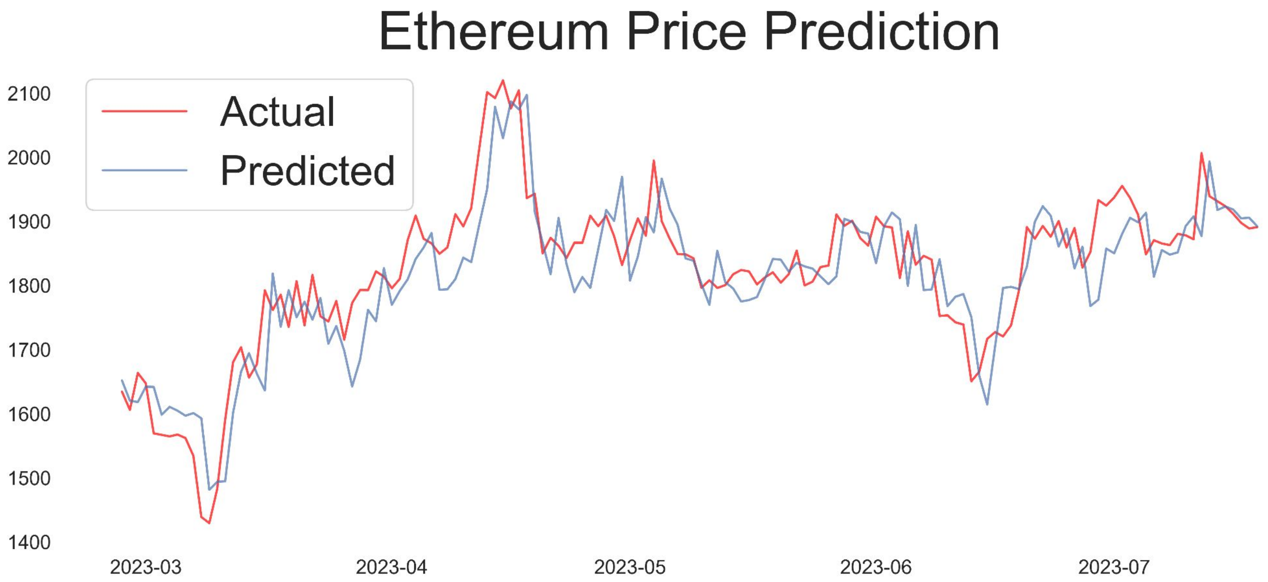
Bitcoin - XGBoost

Mean Absolute Percentage Error: **2.96%**
Root Mean Squared Error: 1067.33
Training data period: 2013/05/30 - 2023/01/14
Test data period: 2023/01/15 - 2023/07/19



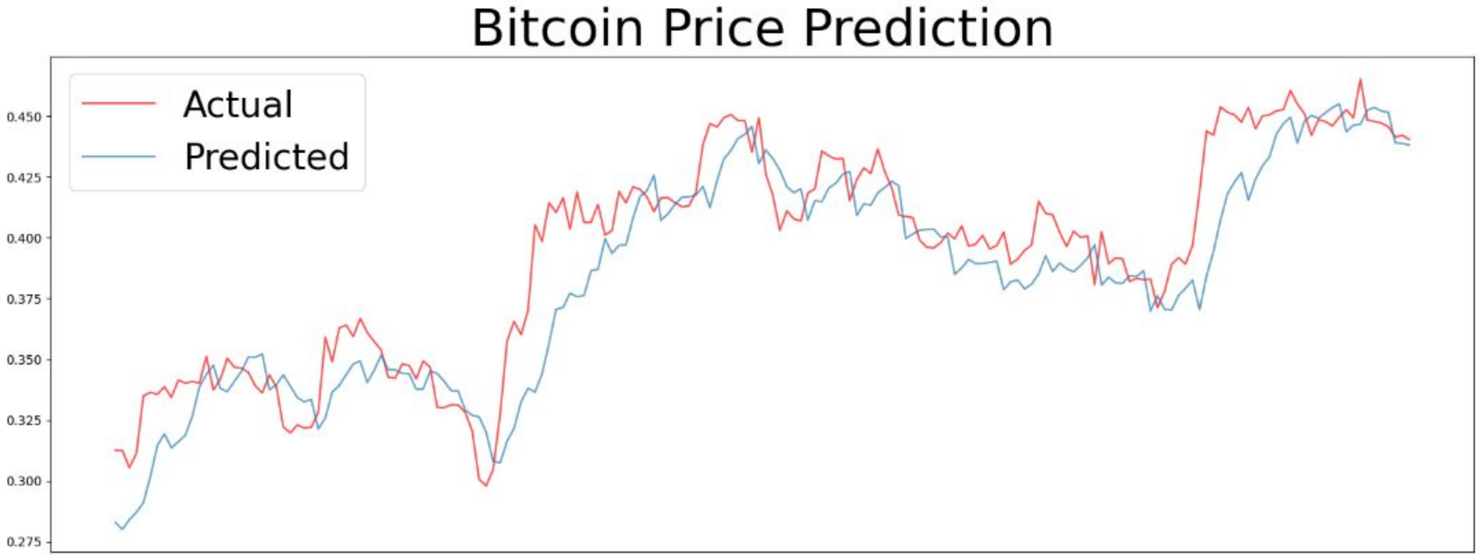
Ethereum - XGboost

Mean Absolute Percentage Error: **2.61%**
Root Mean Squared Error: 60.35
Training data period: 2015/09/09 - 2023/02/25
Test data period: 2023/02/26 - 2023/07/19



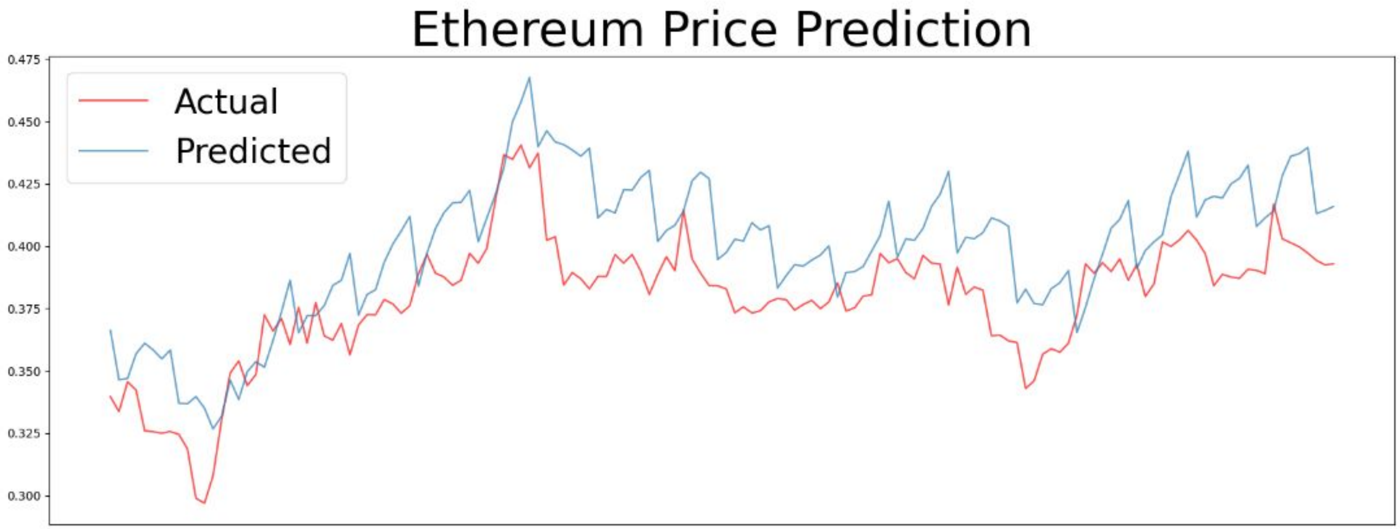
Bitcoin - LSTM

Mean Absolute Percentage Error: **3.87%**



Ethereum - LSTM

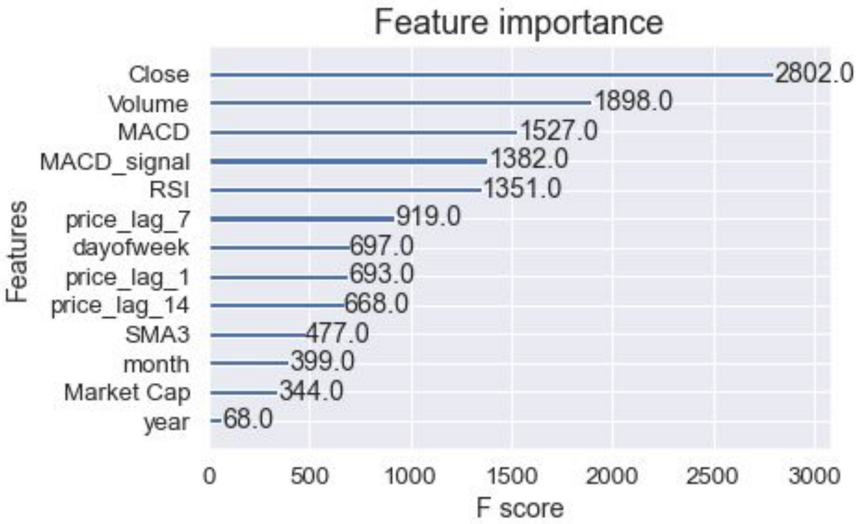
Mean Absolute Percentage Error: **5.82%**



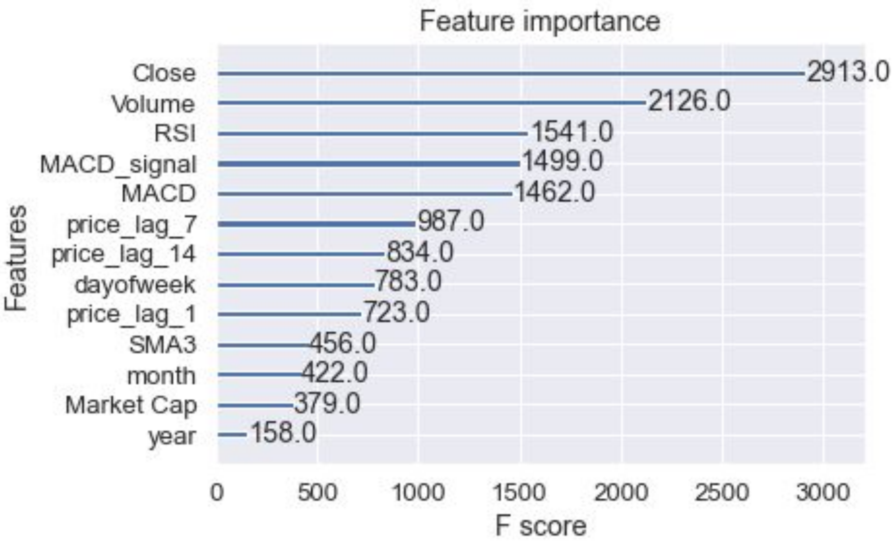
Feature Importance

These scores show how many times each feature was split on (used as a decision making criteria).

Bitcoin



Ethereum



Other Attempts

Fama-French 3 Factor Model - Data

Classic regression for finding factors that explain stock returns

$$R_i - R_f = \alpha_i + \beta_i(R_M - R_f) + s_i SMB + h_i BMB + \epsilon_i$$

Data Sources:

1. Coinmarketcap
   - Webscraped with Selenium
   - Token_id, token category, closing price, volume, market cap

2. Google BigQuery (API, SQL)
   - Ethereum blockchain data
   - large trader net purchase
     - Net purchase or sales of tokens by the largest traders as % of total traded

Corr (returns, large trader net purchase) = **-0.01** (P-val = 0.017)

Fama-French 3 Factor Model - Results

$$R_i - R_f = \alpha_i + \beta_i(R_M - R_f) + s_i SMB + h_i BMB + \epsilon_i$$

- Divide all tokens into 25 portfolios each week based on market cap and active trader net purchase. Calculate coefficients for each portfolio.

Factors:

$R_i - R_f$: return of portfolio i

SMB(Small minus Big): return of small tokens minus return of big tokens

BMB (Bullish minus Bearish): return difference between bullish and bearish tokens

- Bullish (Bearish): high (low) large trader net purchase

Estimation results of h_i for all 25 portfolios.

	Bearish	2	3	4	Bullish
Small	-0.90	-0.77	-0.32	-0.32	-0.27
2	-0.47	-0.25	-0.48	0.49	0.46
3	-0.36	-0.42	-0.42	-0.07	-0.19
4	-0.13	-0.18	-0.07	-0.09	-0.10
Big	-0.19	-0.04	0.10	0.07	0.02

Sentiment Analysis

- Utilized the **NewsCatcher News API** to fetch Google News about cryptocurrencies from 2017-2021 to match our dataset
- Utilized **Transformer** to get the sentiment scores
- Score ranges from -1 (very negative) to 1 (very positive)