Capstone Project 2: Bitcoin Price Prediction

1. Problem statement

The goal of this project is to build a model to predict a bitcoin prices using prices pf other crypto currencies and the numbers of social media posts from previous day. All the information was imported through coingecko.com's API, and the time span of the study was from 06/16/2021 to 09/21/2021.

Deliverables: code, a written report, and a slide deck.

2. Background

Cryptocurrency has been one of the hottest topics in the last few years, especially in the younger generations. A cryptocurrency is basically a digital and encrypted asset that uses blockchain technology along with computers to stand alone on the web without any central party's control. Governments, financial institutions, and banks have not yet fully accepted this type of new currency and many people still believe it's just nothing but a scam.

Nevertheless, many people believe the potentials of the cryptocurrency due to its convenience of fund transfer without third party making a cut as the middleman. The transactions between cryptocurrency holders are beyond boundaries of countries, which makes it possible for money laundry and other crimes but still doesn't lower too much of the greatness of the currency.

Due to its popularity, many parties in the world have been trying their best to figure out how to predict the prize of cryptocurrency to make profits. Recently, some big companies like Tesla and MicroStrategy have invested in Bitcoin, the most well-known cryptocurrency, either to make profit or to hedge. Famous investor Cathie Wood and her Ark Invest is also a big believer of Bitcoin and have invested great amount of money in Bitcoin, and she even predicted the Bitcoin will hit one million, which is about 25 times more from the current price!

In this study, I would use the other cryptocurrency prices and the number of social media posts to predict the price of Bitcoin.

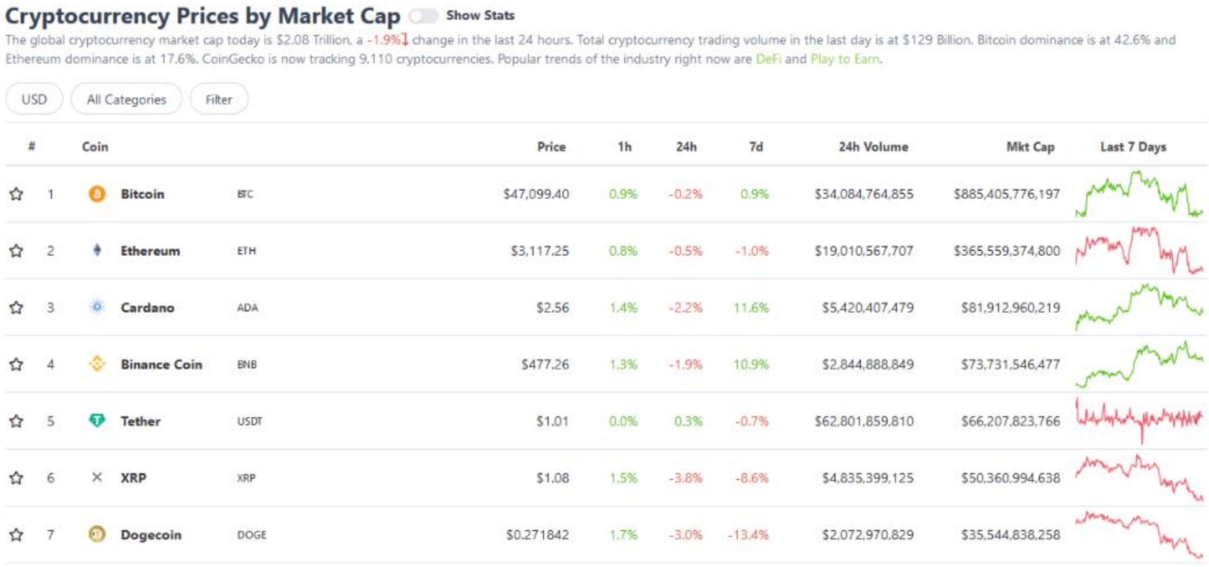
3. Data inspection and cleaning

I retrieve two parts of data using coingecko’s api, cryptocurrency prices and social media data.

Cryptocurrency Prices includes the top 7 cryptocurrencies as fig 1. Bitcoin, Ethereum, Cardano, Binance Coin, Tether, XRP, and Dogecoin. Here I used the last 100 days of data

The url of the price data source is the following:

https://api.coingecko.com/api/v3/coins/btc/market_chart?vs_currency=usd&days=100



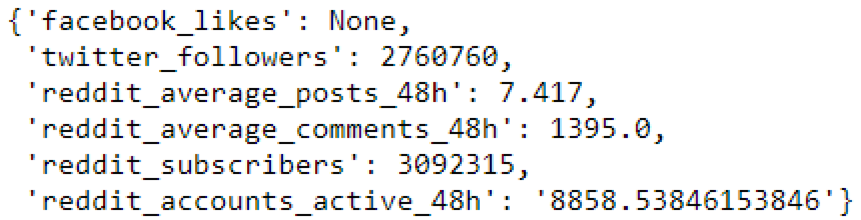
Cryptocurrency Prices by Market Cap Show Stats

The global cryptocurrency market cap today is $2.08 Trillion, a -1.9% change in the last 24 hours. Total cryptocurrency trading volume in the last day is at $129 Billion. Bitcoin dominance is at 42.6% and Ethereum dominance is at 17.6%. CoinGecko is now tracking 9,110 cryptocurrencies. Popular trends of the industry right now are DeFi and Play to Earn.

	#	Coin		Price	1h	24h	7d	24h Volume	Mkt Cap	Last 7 Days
☆	1	Bitcoin	BTC	$47,099.40	0.9%	-0.2%	0.9%	$34,084,764,855	$885,405,776,197	
☆	2	Ethereum	ETH	$3,117.25	0.8%	-0.5%	-1.0%	$19,010,567,707	$365,559,374,800	
☆	3	Cardano	ADA	$2.56	1.4%	-2.2%	11.6%	$5,420,407,479	$81,912,960,219	
☆	4	Binance Coin	BNB	$477.26	1.3%	-1.9%	10.9%	$2,844,888,849	$73,731,546,477	
☆	5	Tether	USDT	$1.01	0.0%	0.3%	-0.7%	$62,801,859,810	$66,207,823,766	
☆	6	XRP	XRP	$1.08	1.5%	-3.8%	-8.6%	$4,835,399,125	$50,360,994,638	
☆	7	Dogecoin	DOGE	$0.271842	1.7%	-3.0%	-13.4%	$2,072,970,829	$35,544,838,258	

Fig 1. Top 7 most popular cryptocurrency (derived from https://www.coingecko.com/).

Reddit Data includes the following bitcoin related Facebook likes, twitter followers, reddit average posts in the last 48 hours, reddit average comment in the last 48 hours, reddit subscribers, and reddit active accounts discussing in the last 48 hours.



```
{'facebook_likes': None,
 'twitter_followers': 2760760,
 'reddit_average_posts_48h': 7.417,
 'reddit_average_comments_48h': 1395.0,
 'reddit_subscribers': 3092315,
 'reddit_accounts_active_48h': '8858.53846153846'}
```

Fig 2. A sample of social media's data related to cryptocurrency

The url of the Reddit data source is the following:

https://api.coingecko.com/api/v3/coins/btc/market_chart?vs_currency=usd&days=100

Steps to clean up cryptocurrency price data:

1. First, I saved each cryptocurrency price respectively.
2. I joined all the data to a data frame on time_stamp.

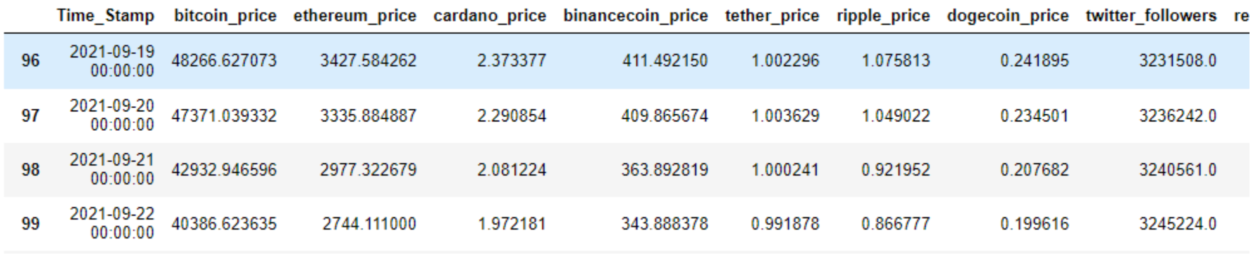
Steps to clean up social media's data:

1. I used the time_stamp from previous data frame to pull the data through coingecko's api.
2. The api would stop working sometimes while pulling the data, so I add a loop to request info several times to make sure I got all the information.
3. The Fabebook likes data didn't have any information so I dropped the column.
4. I then combined price and social media data in a data frame
5. I removed the last row since it's pulling today's data and it didn't have completed data.
6. There was one row that didn't have information on some of the cryptocurrency, so I used rolling average to fill.

Adding features:

I added two feature twitter_followers_diff and reddit_subscribers_diff to show the differences of the follower and subscribers for each day.

Looking at the last few columns as an example:



	Time_Stamp	bitcoin_price	ethereum_price	cardano_price	binancecoin_price	tether_price	ripple_price	dogecoin_price	twitter_followers	re
96	2021-09-19 00:00:00	48266.627073	3427.584262	2.373377	411.492150	1.002296	1.075813	0.241895	3231508.0	
97	2021-09-20 00:00:00	47371.039332	3335.884887	2.290854	409.865674	1.003629	1.049022	0.234501	3236242.0	
98	2021-09-21 00:00:00	42932.946596	2977.322679	2.081224	363.892819	1.000241	0.921952	0.207682	3240561.0	
99	2021-09-22 00:00:00	40386.623635	2744.111000	1.972181	343.888378	0.991878	0.866777	0.199616	3245224.0	

Fig 3. Some examples of the data frame

Except for Bitcoin prices, all the other columns are numerical. The following figure is the visualization of crypto prices.

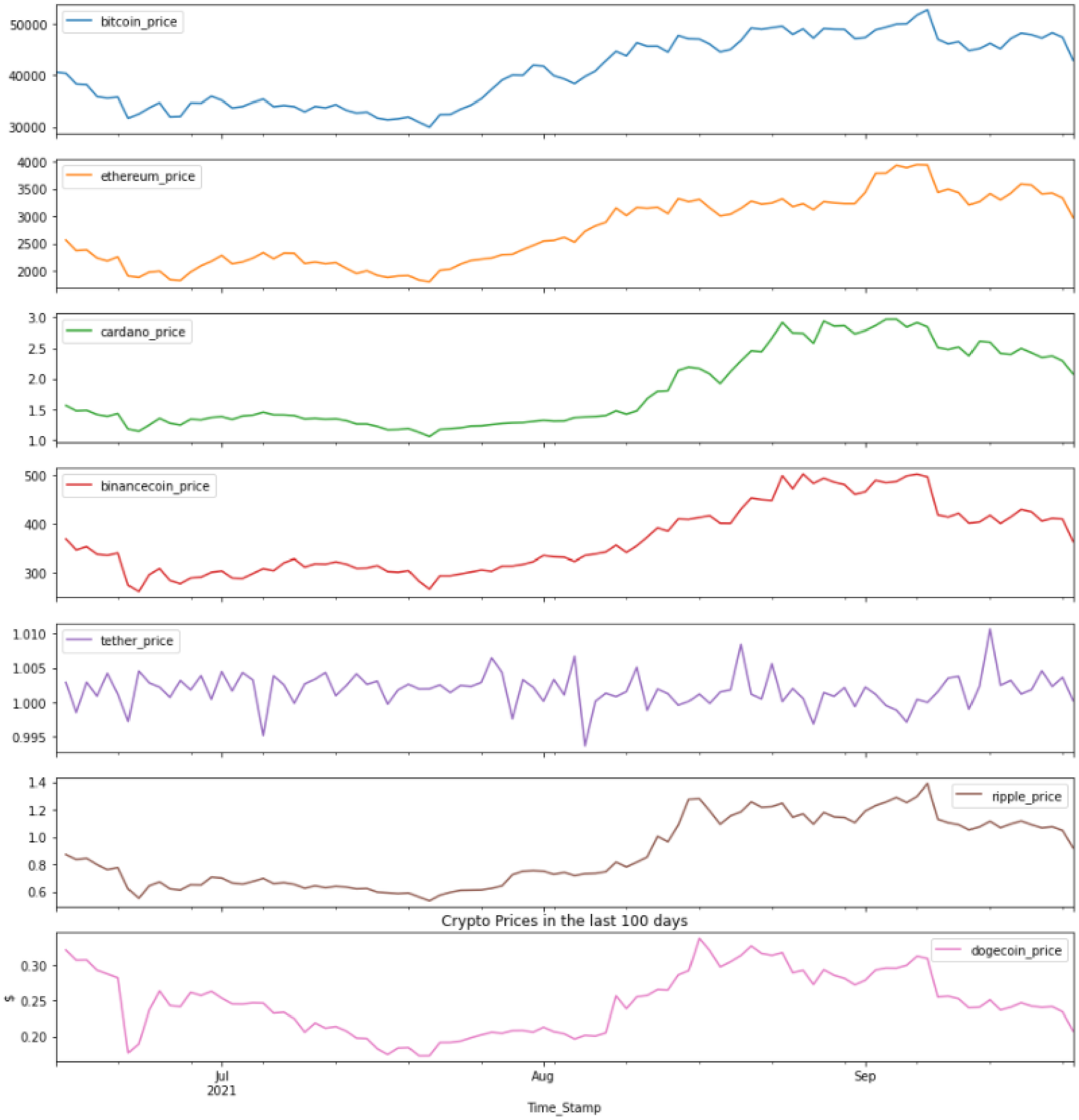


Fig 4. Visualization of 7 cryptocurrenies we picked

Scale data:

Cryptocurrency prices and socail media are all numerical but are very different in numbers, so I used MaxMinScaler() to scale all the data

The following figure showed high correlation between cryptocurrency prices except for tether prices. And Reddit subscribers seemed highly correlated to cryptocurrency prices.

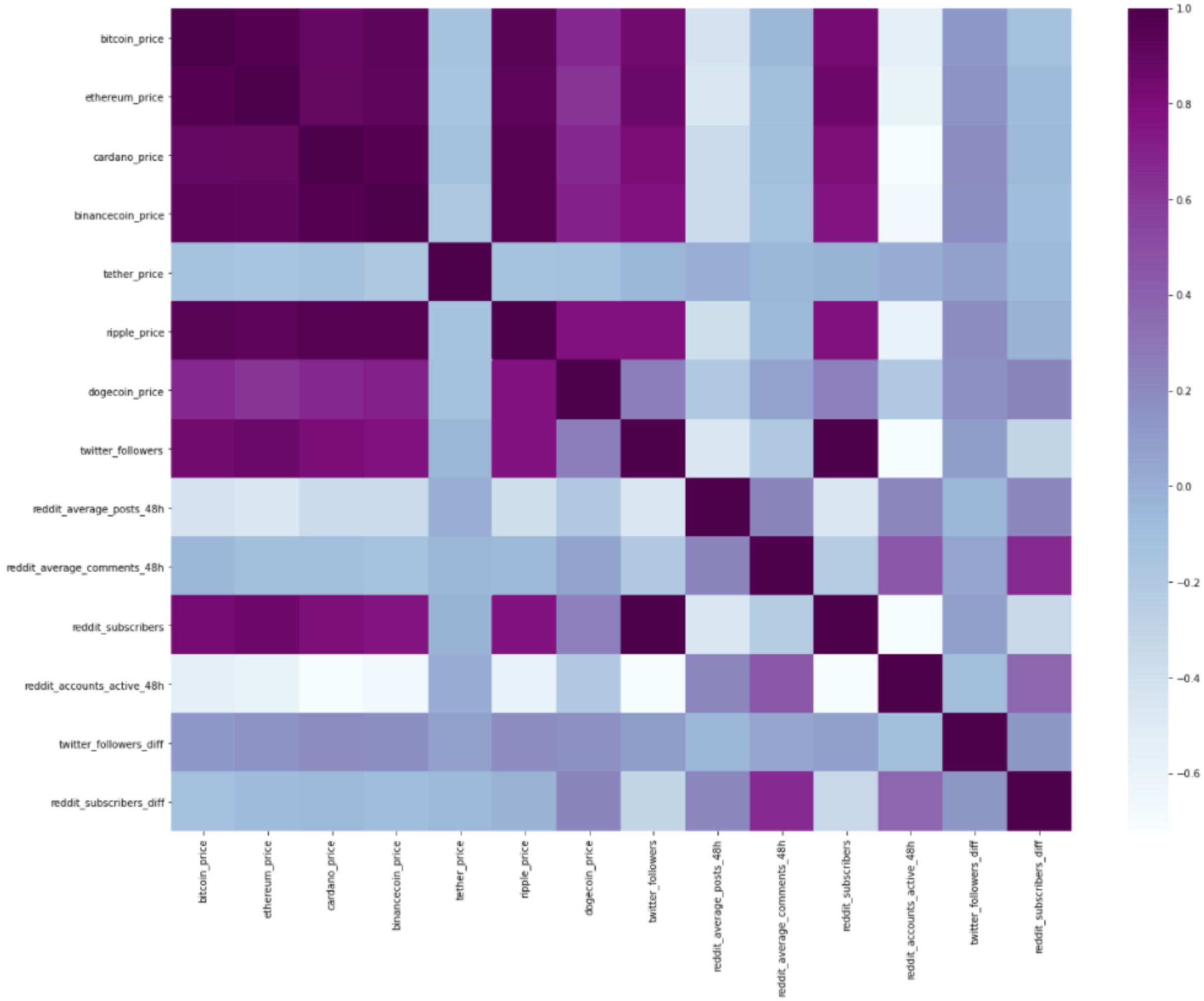


Fig 5. Correlation between prices

Last step before model was shifting bitcoin price to a day before in the data frame because I would need to use previous data to predict the bitcoin price of the next day.

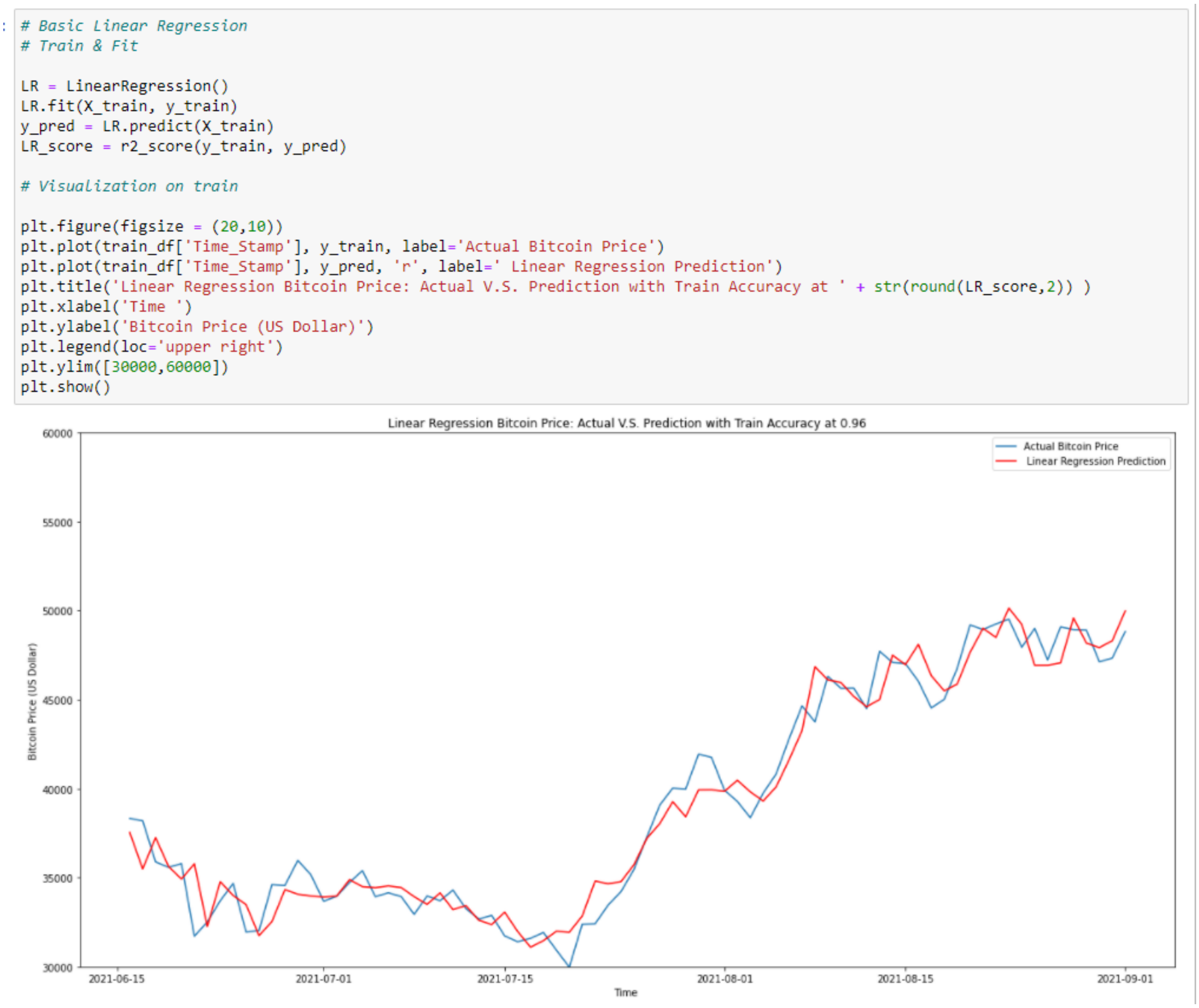
4. Model Building

I used the first 80% of the data to train and the last 20% of the data to test. I selected 4 models to predict bitcoin prices:

Linear Regression, Lasso Regression, Ridge Regression, and Elastic Net Regression

Linear Regression:

Using default linear regression model, as it shows in figure 6, training data has a high R^2 score at 0.96.

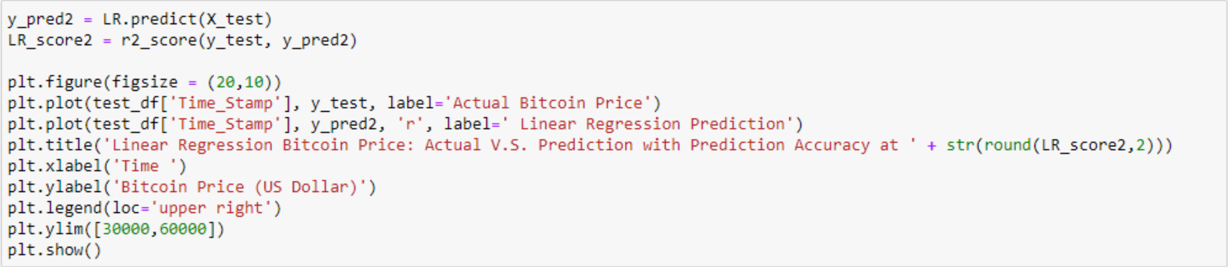


```
# Basic Linear Regression
# Train & Fit

LR = LinearRegression()
LR.fit(X_train, y_train)
y_pred = LR.predict(X_train)
LR_score = r2_score(y_train, y_pred)

# Visualization on train

plt.figure(figsize = (20,10))
plt.plot(train_df['Time_Stamp'], y_train, label='Actual Bitcoin Price')
plt.plot(train_df['Time_Stamp'], y_pred, 'r', label=' Linear Regression Prediction')
plt.title('Linear Regression Bitcoin Price: Actual V.S. Prediction with Train Accuracy at ' + str(round(LR_score,2)) )
plt.xlabel('Time ')
plt.ylabel('Bitcoin Price (US Dollar)')
plt.legend(loc='upper right')
plt.ylim([30000,60000])
plt.show()
```

Linear Regression Bitcoin Price: Actual V.S. Prediction with Train Accuracy at 0.96
Actual Bitcoin Price
Linear Regression Prediction
Bitcoin Price (US Dollar)
60000
55000
50000
45000
40000
35000
30000
2021-06-15
2021-07-01
2021-07-15
2021-08-01
2021-08-15
2021-09-01
Time

Fig 6. Linear Regression in training data

However, in the testing data R^2 is negative as it shows in figure 7.



```
y_pred2 = LR.predict(X_test)
LR_score2 = r2_score(y_test, y_pred2)

plt.figure(figsize = (20,10))
plt.plot(test_df['Time_Stamp'], y_test, label='Actual Bitcoin Price')
plt.plot(test_df['Time_Stamp'], y_pred2, 'r', label=' Linear Regression Prediction')
plt.title('Linear Regression Bitcoin Price: Actual V.S. Prediction with Prediction Accuracy at ' + str(round(LR_score2,2)))
plt.xlabel('Time ')
plt.ylabel('Bitcoin Price (US Dollar)')
plt.legend(loc='upper right')
plt.ylim([30000,60000])
plt.show()
```

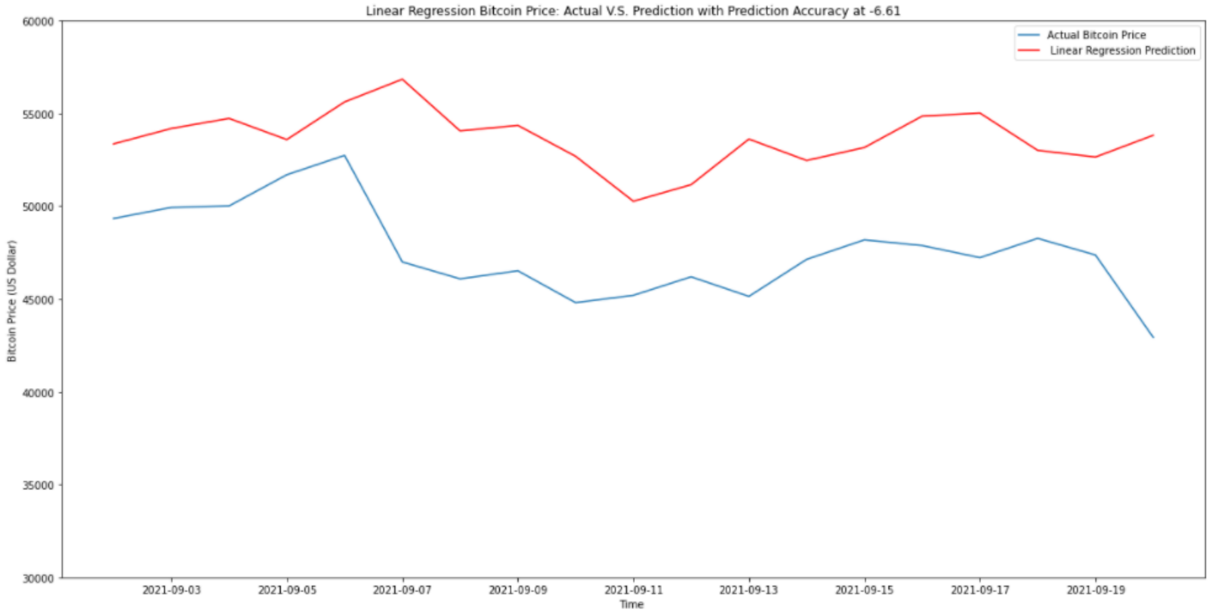
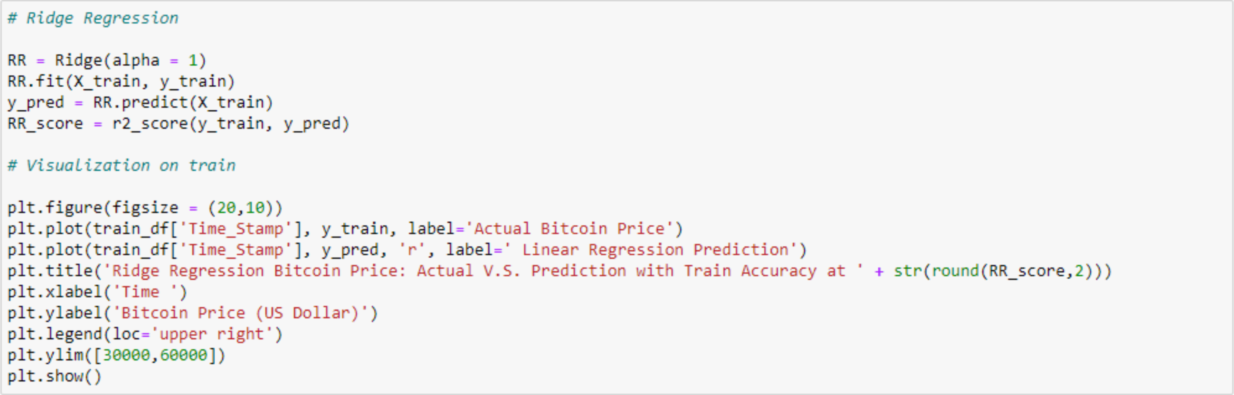


Fig 7. Linear Regression in testing data

Ridge Regression:

As it shows in figure 6, like linear regression, default model gave training a high R^2 score at 0.93.



```
# Ridge Regression

RR = Ridge(alpha = 1)
RR.fit(X_train, y_train)
y_pred = RR.predict(X_train)
RR_score = r2_score(y_train, y_pred)

# Visualization on train

plt.figure(figsize = (20,10))
plt.plot(train_df['Time_Stamp'], y_train, label='Actual Bitcoin Price')
plt.plot(train_df['Time_Stamp'], y_pred, 'r', label=' Linear Regression Prediction')
plt.title('Ridge Regression Bitcoin Price: Actual V.S. Prediction with Train Accuracy at ' + str(round(RR_score,2)))
plt.xlabel('Time ')
plt.ylabel('Bitcoin Price (US Dollar)')
plt.legend(loc='upper right')
plt.ylim([30000,60000])
plt.show()
```

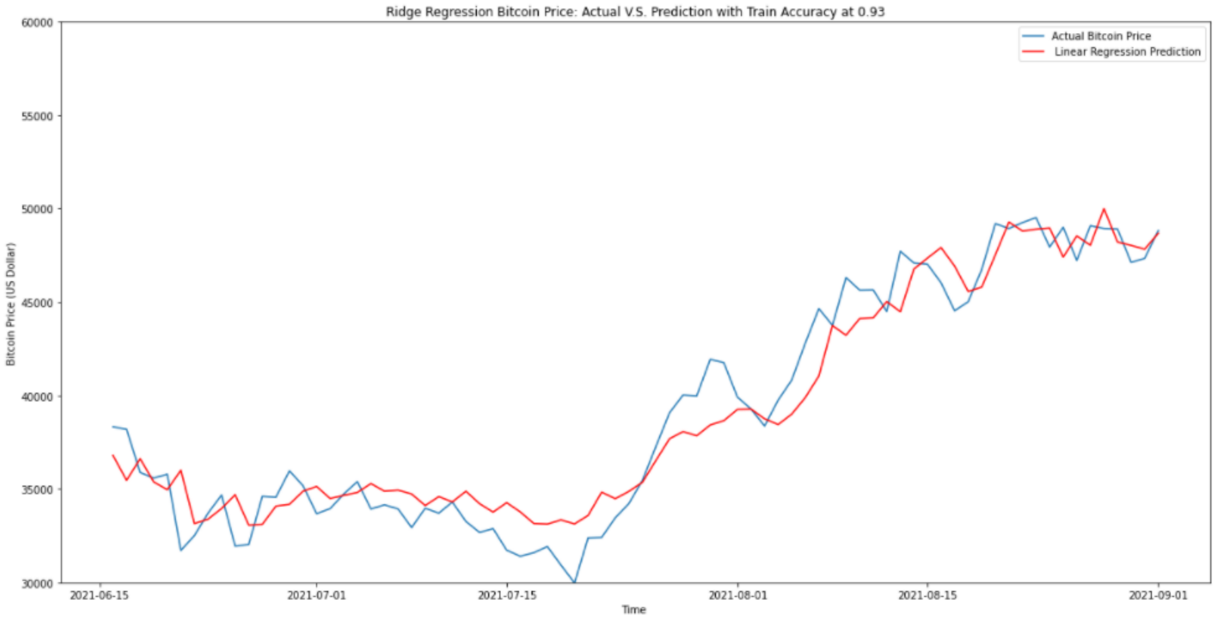


Fig 8. Ridge Regression in training data

However, testing data only had R^2 score at -1.2 in figure 9.

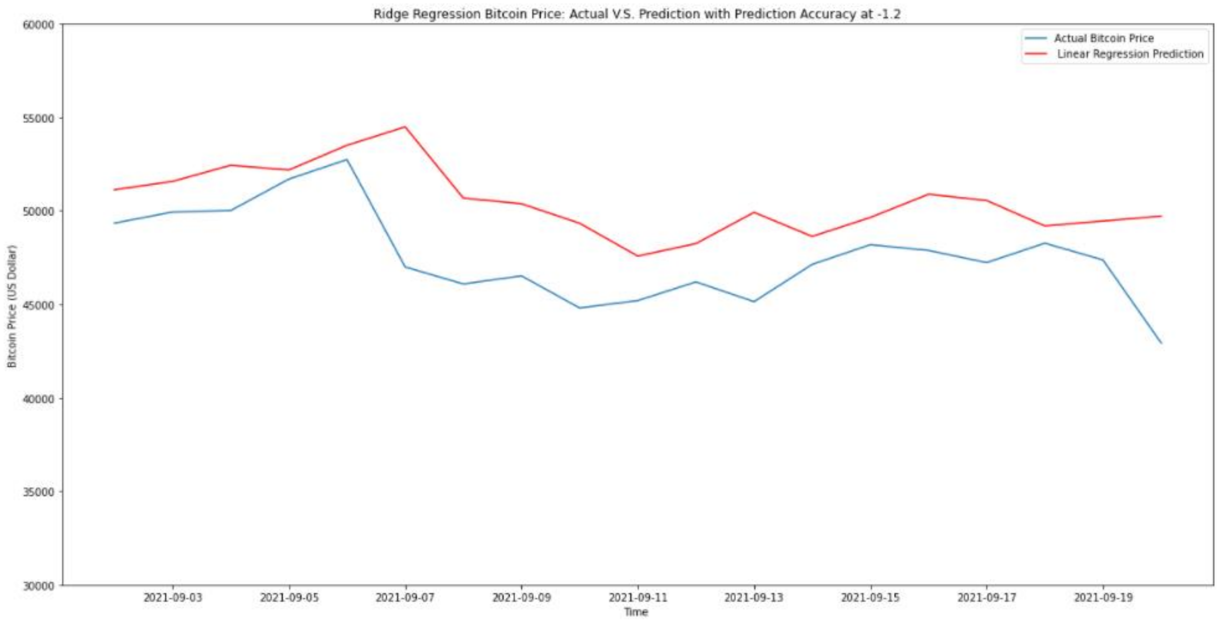


Fig 9. Ridge Regression in testing data

I then used hyperparameter tuning on alpha and used the best alpha at 0.1 to fit and test.

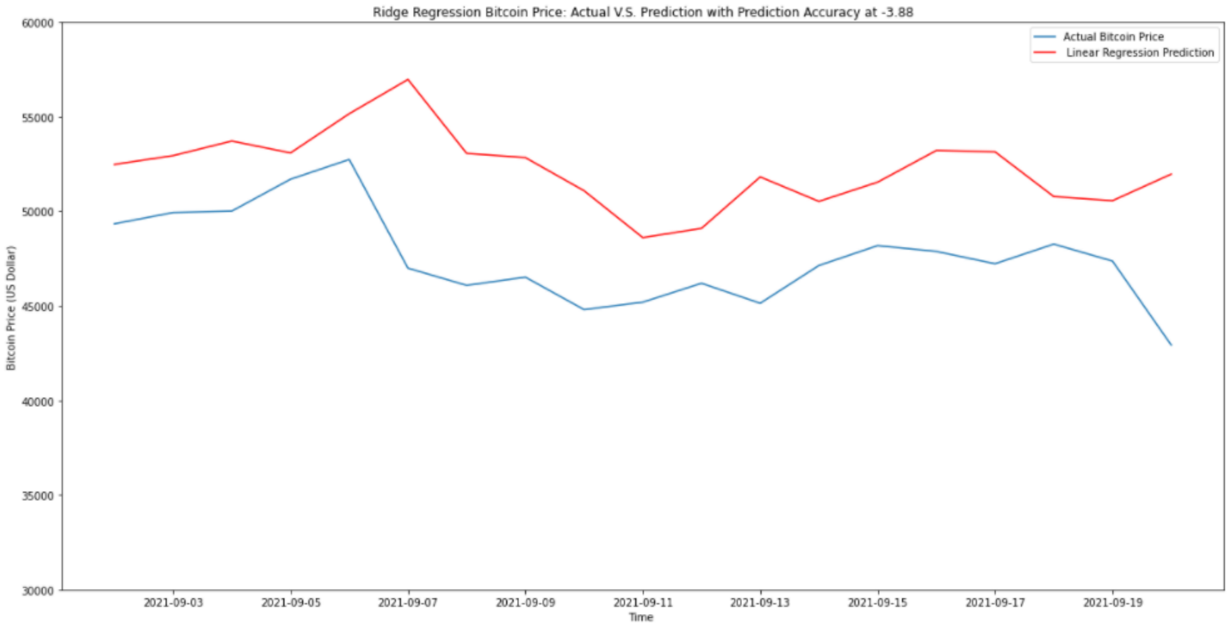


Fig 10. Ridge Regression in testing data after hyperpramarter tuning

In figure 10, you can see R^2 score didn't improve and was still negative at -3.88.

Lasso Regression:

Like Linear Regression and Ridge Regression, Lasso Regression had great R^2 score but poor score on testing data as you can see in figure 11. R^2 score was only at -5.67.

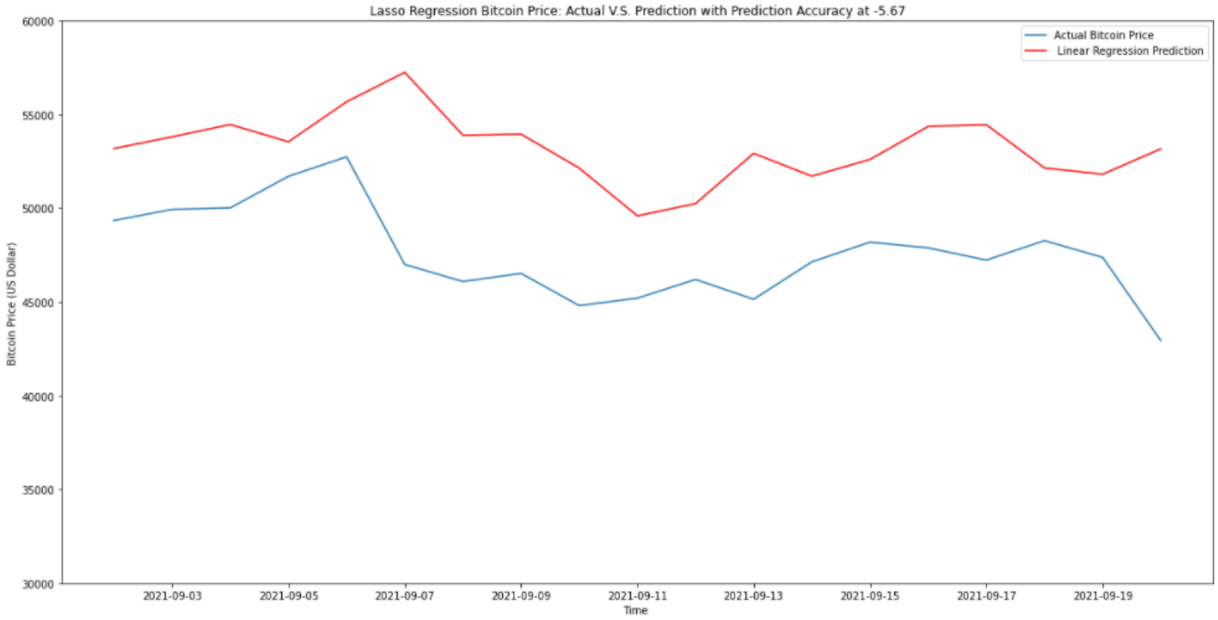


Fig 11. Lasso Regression in testing data

Hyperparameter tuning on alpha gave the best alpha at 10, but it still didn't help R^2 score. It was negative at -4.13

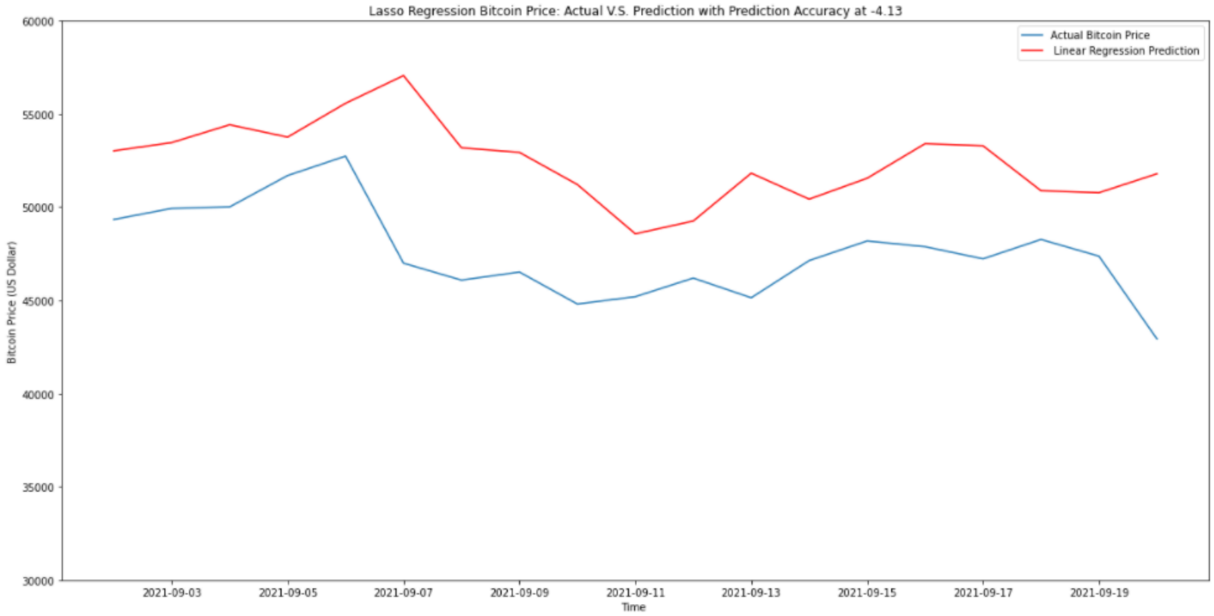


Fig 11. Lasso Regression in testing data after hyperparmarter tuning

Elastic Net Regression:

The default model of Elastic Net Regression, different from all the other previously built models, had an awful R2 score at 0.58 as you can see inf figure 12.

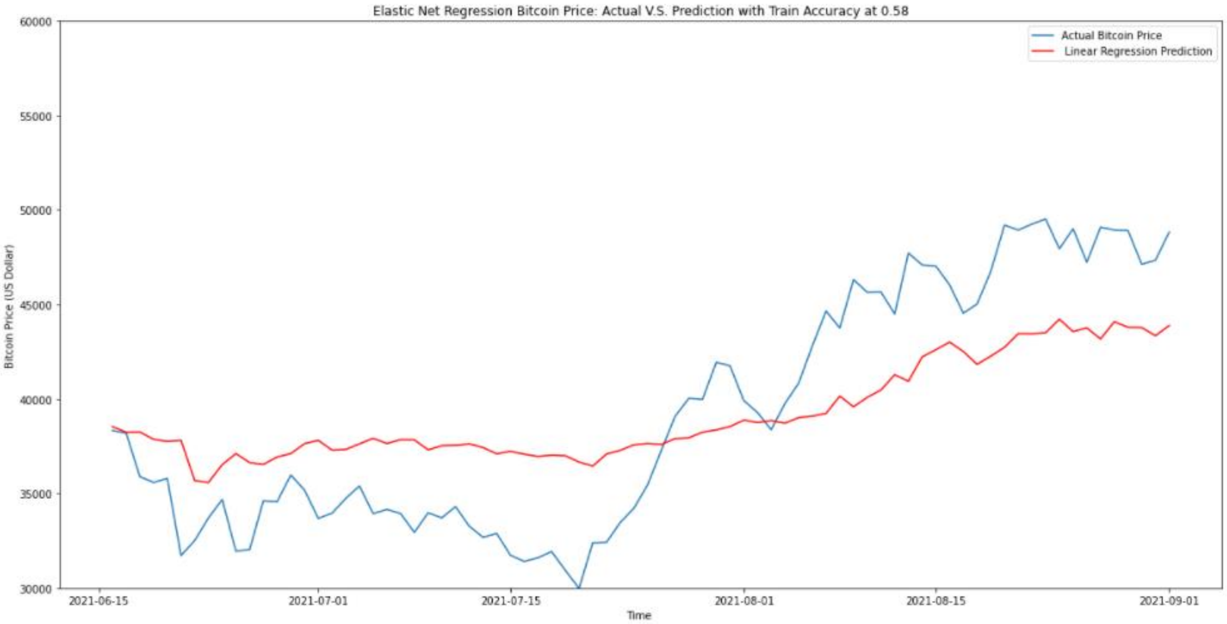


Fig 12. Elastic Net Regression in training data

The R^2 score was still negative in testing data except for the last data point.

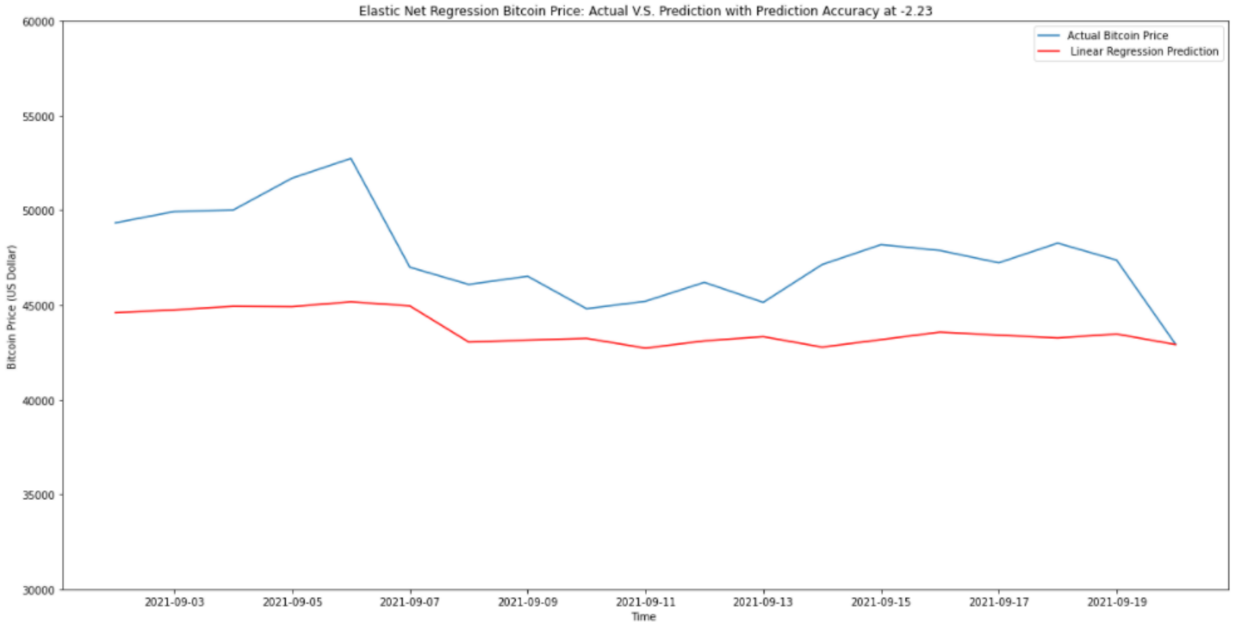


Fig 13. Elastic Net Regression in testing data

Hyperparameter tuning on parameters alpha and l1_ratio got me the exact same model as lasso model we built earlier. The best parameters were alpha at 10 and l1_ratio at 1.

Feature importance:

The following figures are the feature importance looking at the coef_ value for each model:

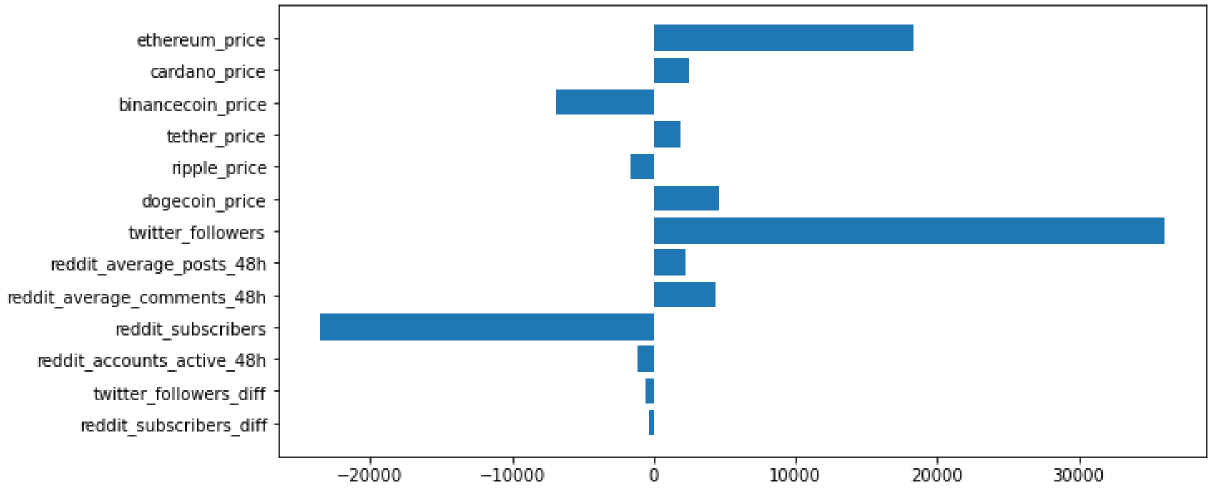


Fig 14. Logistic Regression feature importance

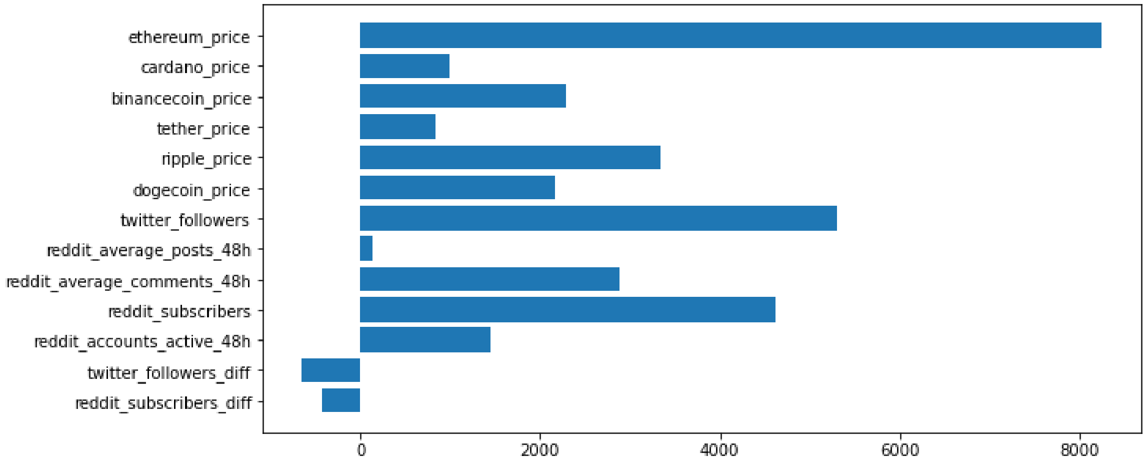


Fig 15. Ridge Regression feature importance

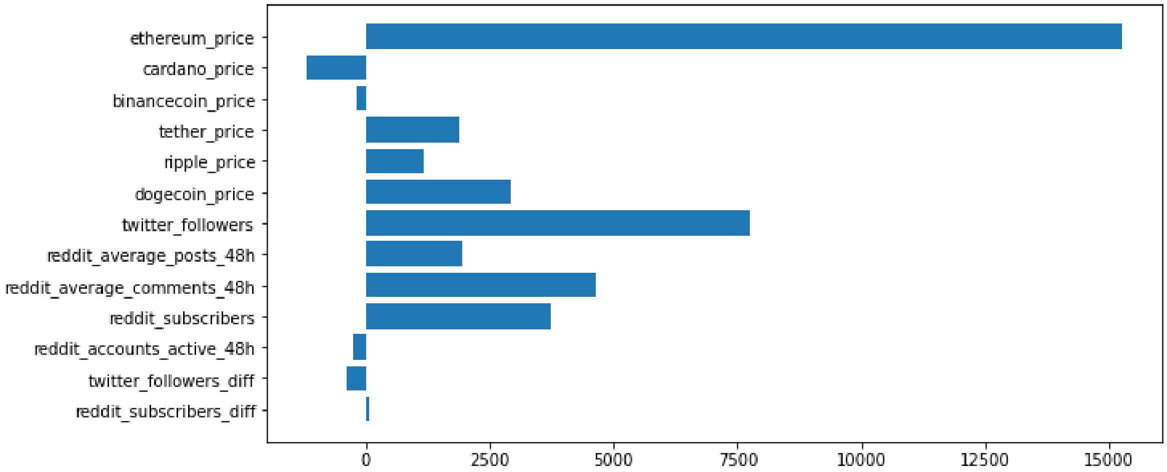


Fig 16. Ridge Regression feature importance after hyperparamter tuning

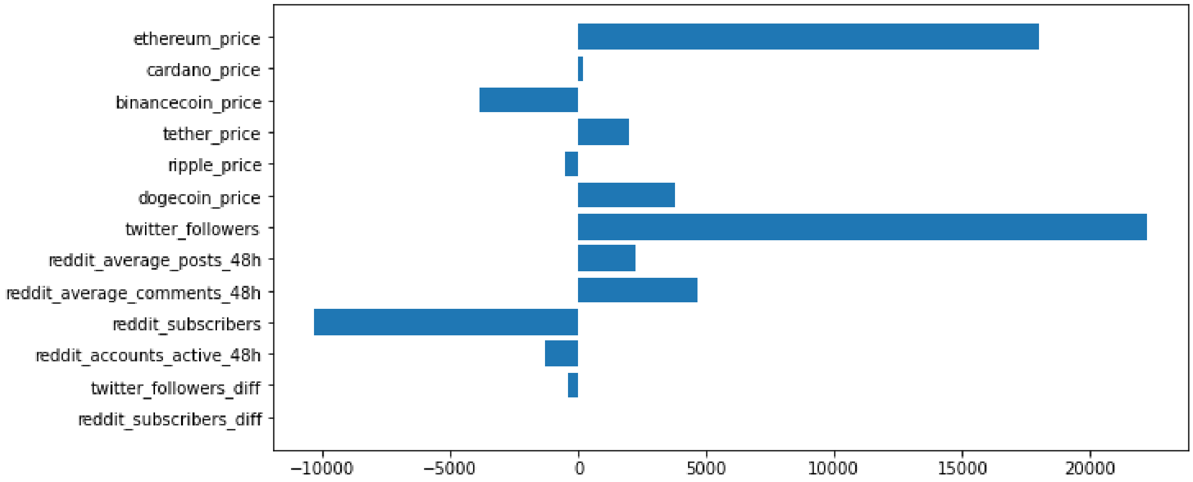


Fig 17. Lasso Regression feature importance

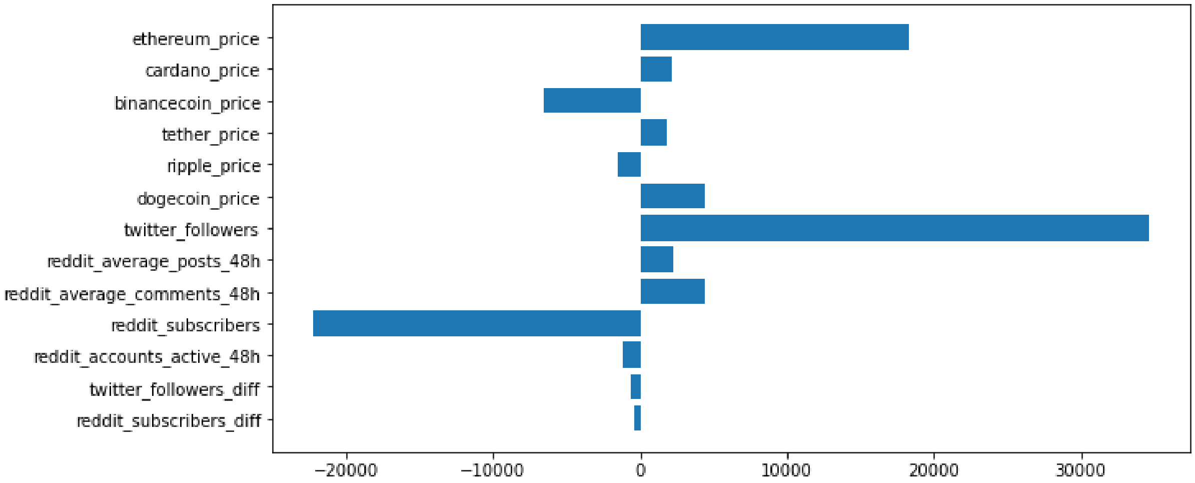


Fig 18. Lasso Regression feature importance after hyperparameter tuning

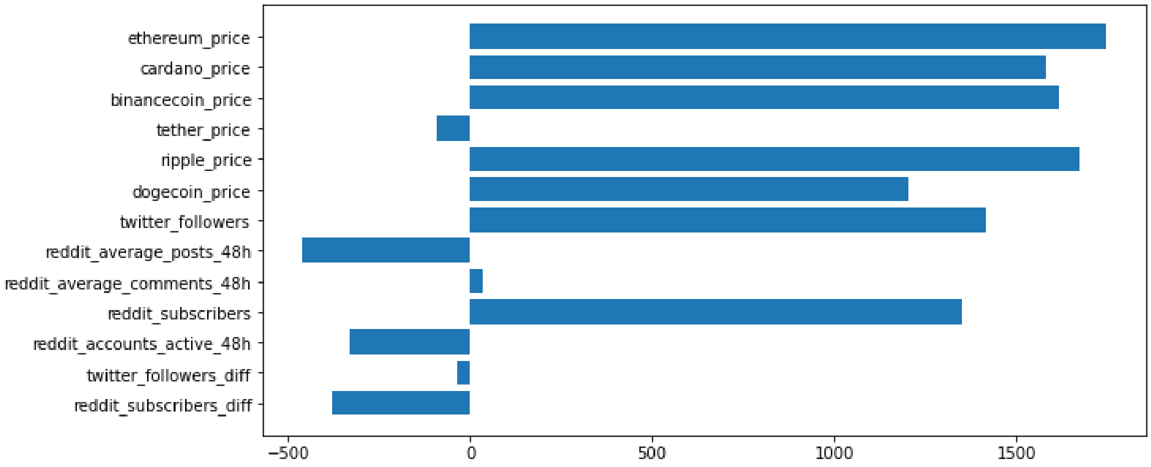


Fig 19. Elastic Net Regression feature importance

From figure 14-19, we can the Ethereum prices and twitter follower seem to be the most important features in all the models except for Elastic Net Regression model. In Elastic Net Regression model, 7 features are approximately tied on importance.

5. Summary:

4 models, Linear Regression, Lasso Regression, Ridge Regression, and Elastic Net Regression have been tested to try to build a good model that can predict future bitcoin prices using previous date' cryptocurrency prices and social media posts and followers. Hyperparameter tuning have been performed using grid search method but still couldn't get decent or even positive R^2 score. Possible cause was probably due to the recent huge fluctuation caused by China's housing sector mass collapse. Also, FED recently signaled bond-buying taper coming soon. Due to both events, stock and bond markets have been heavily affected, and we could assume Bitcoin price was also negatively affected that caused the prediction of bitcoin prices in each models was higher than the actual price.

6. Recommendations for clients to use the findings:

1. Using merely 6 other cryptocurrency prices and social media post and followers didn’t seem to provide good prediction on Bitcoin price. Investor on Bitcoin would need to discover more data.

2. Among all the findings, Ethereum price and twitter followers seem to be something we can keep in the future model because they are the most important features in the models built.

3. Bitcoin as the most popular cryptocurrency, it will be the first thing people liquidate which makes it highly unpredictable comparing with other equity that are hard to liquidate such as houses or gold. Therefore, recent negative or positive news may impact Bitcoin price heavily. Bitcoin can be traded 24-7 which doesn’t give investors time to digest good or bad news so the prices will fluctuate a lot in a short period of time after big news announcement. Investors should set up sell limit on Bitcoin to hedge themselves from huge loss.

7. Ideas for further studies:

Stock and bond market seem to have high correlation with Bitcoin market. We could add that into future model to see if it makes improvement on accuracy. Also, news regarding Bitcoin can be a good feature to consider.