DATA SCIENCE AND AI PROJECT

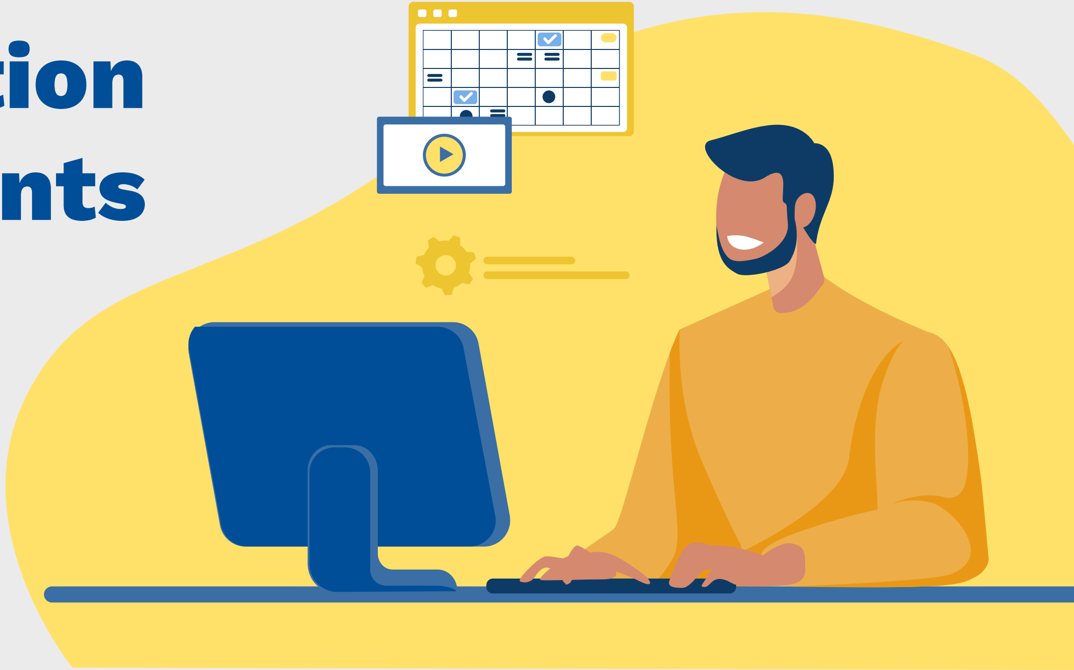
FSP6_Team03

Group Members: Ingale Omkar, Lau Chen Yi Wynne, Himari Ang Lixin

TABLE OF CONTENTS

01

Data Extraction Problem Statements



Problem Statements

01

To what extent is the price of Bitcoin dependent on the global financial system that is represented through stock indices?

02

Which model will provide a better accuracy in predicting bitcoin prices, VAR or LSTM?

Data Extraction – BTC

Alpha Vantage

date	1a. open (USD)	1b. open (USD)	2a. high (USD)	2b. high (USD)	3a. low (USD)	3b. low (USD)	4a. close (USD)	4b. close (USD)	5. volume	6. market cap (USD)
2021-03-24	54342.80	54342.80	57200.00	57200.00	51700.00	51700.00	52303.65	52303.65	83537.465021	83537.465021
2021-03-23	54083.25	54083.25	55830.90	55830.90	53000.00	53000.00	54340.89	54340.89	59789.365427	59789.365427
2021-03-22	57351.56	57351.56	58430.73	58430.73	53650.00	53650.00	54083.25	54083.25	62581.626169	62581.626169
2021-03-21	58100.02	58100.02	58589.10	58589.10	55450.11	55450.11	57351.56	57351.56	48564.470274	48564.470274
2021-03-20	58030.01	58030.01	59880.00	59880.00	57820.17	57820.17	58102.28	58102.28	44476.941776	44476.941776
...	...	...	...	...	...	...	...	...	...	...
2018-08-01	7735.67	7735.67	7750.00	7750.00	7430.00	7430.00	7604.58	7604.58	42582.312932	42582.312932
2018-07-31	8171.40	8171.40	8180.00	8180.00	7633.00	7633.00	7730.93	7730.93	48296.915587	48296.915587
2018-07-30	8210.99	8210.99	8273.00	8273.00	7866.00	7866.00	8173.92	8173.92	39692.416542	39692.416542
2018-07-29	8225.04	8225.04	8294.51	8294.51	8115.00	8115.00	8211.00	8211.00	25531.226185	25531.226185
2018-07-28	8188.57	8188.57	8246.54	8246.54	8067.00	8067.00	8225.04	8225.04	26215.173839	26215.173839

971 rows × 10 columns

Data Extraction - BTC

API

Date	Close
2018-07-28	8225.04
2018-07-29	8211.00
2018-07-30	8173.92
2018-07-31	7730.93
2018-08-01	7604.58
...	...
2021-03-20	58102.28
2021-03-21	57351.56
2021-03-22	54083.25
2021-03-23	54340.89
2021-03-24	52303.65

971 rows × 1 columns

+

Yahoo Finance

Date	Close
2014-09-17	457.334015
2014-09-18	424.440002
2014-09-19	394.795990
2014-09-20	408.903992
2014-09-21	398.821014
...	...
2018-07-23	7711.109863
2018-07-24	8424.269531
2018-07-25	8181.390137
2018-07-26	7951.580078
2018-07-27	8165.009766

1410 rows × 1 columns

→

Bitcoin Data

Date	Close
2014-09-17	457.334015
2014-09-18	424.440002
2014-09-19	394.795990
2014-09-20	408.903992
2014-09-21	398.821014
...	...
2021-03-20	58102.280000
2021-03-21	57351.560000
2021-03-22	54083.250000
2021-03-23	54340.890000
2021-03-24	52303.650000

2381 rows × 1 columns

Data Extraction – Other indices

- S&P/ASX 200 ('^AXJO')
- S&P BSE SENSEX ('^BSESN')
- Dow Jones Industrial Average ('^DJI')
- S&P500 index ('^GSPC')
- S&P/TSX Composite index ('^GSPTSE')
- NASDAQ Composite ('^IXIC')
- KOSPI Composite Index (^KS11)
- Nikkei ('^N225')
- NIFTY 50 ('^NSEI')
- NYSE COMPOSITE ('^NYA')

Date	^AXJO	^BSESN	^DJI	^GSPC	^GSPTSE	^IXIC	^KS11	^N225	^NSEI	^NYA
2014-09-17	5407.299805	26631.289062	17156.849609	2001.569946	15458.900391	4562.189941	2062.610107	15888.669922	7975.500000	10973.740234
2014-09-18	5415.799805	27112.210938	17265.990234	2011.359985	15465.500000	4593.430176	2047.739990	16067.570312	8114.750000	11024.059570
2014-09-19	5433.100098	27090.419922	17279.740234	2010.400024	15265.400391	4579.790039	2053.820068	16321.169922	8121.450195	10989.570312
2014-09-22	5363.000000	27206.740234	17172.679688	1994.290039	15129.000000	4527.689941	2039.270020	16205.900391	8146.299805	10892.639648
2014-09-23	5415.700195	26775.689453	17055.869141	1982.770020	15125.700195	4508.689941	2028.910034	NaN	8017.549805	10815.419922
...	...	...	...	...	...	...	...	...	...	...
2021-03-18	6745.899902	49216.519531	32862.300781	3915.459961	18836.500000	13116.169922	3066.010010	30216.750000	14557.849609	15589.089844
2021-03-19	6708.200195	49858.238281	32627.970703	3913.100098	18854.000000	13215.240234	3039.530029	29792.050781	14744.000000	15562.299805
2021-03-22	6752.500000	49771.289062	32731.199219	3940.590088	18815.099609	13377.540039	3035.459961	29174.150391	14736.400391	15551.559570
2021-03-23	6745.399902	50051.441406	32423.150391	3910.520020	18669.800781	13227.700195	3004.739990	28995.919922	14814.750000	15346.530273
2021-03-24	6778.799805	NaN	NaN	NaN	NaN	NaN	2996.350098	28405.519531	NaN	NaN

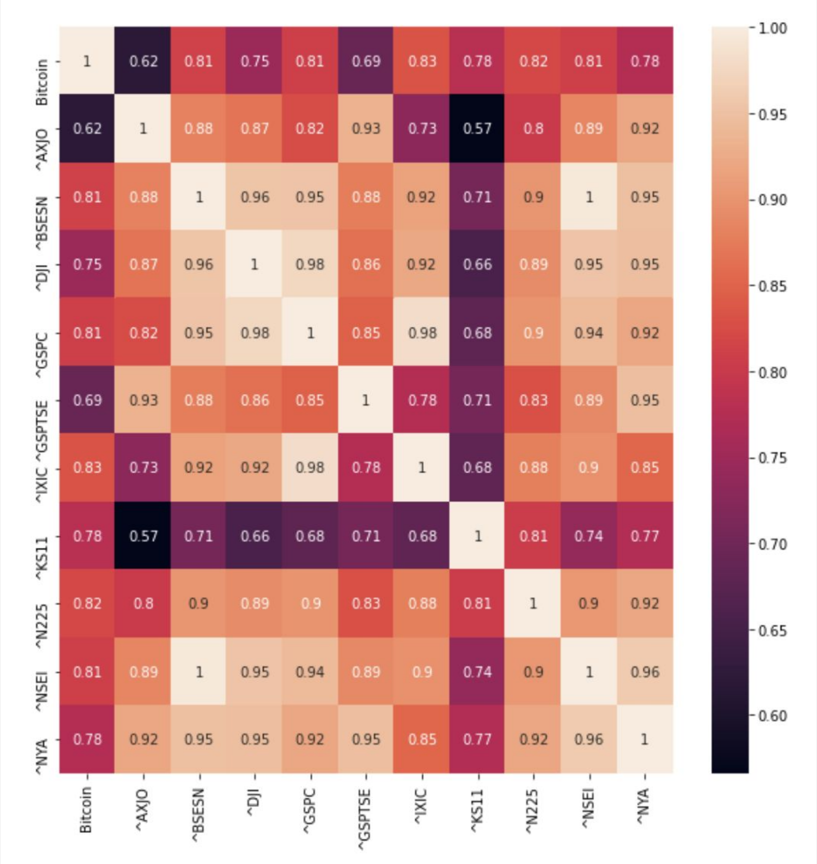
1697 rows × 10 columns

Data Cleaning - Combine the dataframes

	Bitcoin	^AXJO	^BSESN	^DJI	^GSPC	^GSPTSE	^IXIC	^KS11	^N225	^NSEI	
Date											
2014-09-17	457.334015	5407.299805	26631.289062	17156.849609	2001.569946	15458.900391	4562.189941	2062.610107	15888.669922	7975.500000	10973.7
2014-09-18	424.440002	5415.799805	27112.210938	17265.990234	2011.359985	15465.500000	4593.430176	2047.739990	16067.570312	8114.750000	11024.(
2014-09-19	394.795990	5433.100098	27090.419922	17279.740234	2010.400024	15265.400391	4579.790039	2053.820068	16321.169922	8121.450195	10989.5
2014-09-22	402.152008	5363.000000	27206.740234	17172.679688	1994.290039	15129.000000	4527.689941	2039.270020	16205.900391	8146.299805	10892.6
2014-09-23	435.790985	5415.700195	26775.689453	17055.869141	1982.770020	15125.700195	4508.689941	2028.910034	NaN	8017.549805	10815.4
...	...	...	...	...	...	...	...	...	...	...	
2021-03-18	57648.160000	6745.899902	49216.519531	32862.300781	3915.459961	18836.500000	13116.169922	3066.010010	30216.750000	14557.849609	15589.(
2021-03-19	58030.010000	6708.200195	49858.238281	32627.970703	3913.100098	18854.000000	13215.240234	3039.530029	29792.050781	14744.000000	15562.2
2021-03-22	54083.250000	6752.500000	49771.289062	32731.199219	3940.590088	18815.099609	13377.540039	3035.459961	29174.150391	14736.400391	15551.5
2021-03-23	54340.890000	6745.399902	50051.441406	32423.150391	3910.520020	18669.800781	13227.700195	3004.739990	28995.919922	14814.750000	15346.5
2021-03-24	52303.650000	6778.799805	NaN	NaN	NaN	NaN	NaN	2996.350098	28405.519531	NaN	

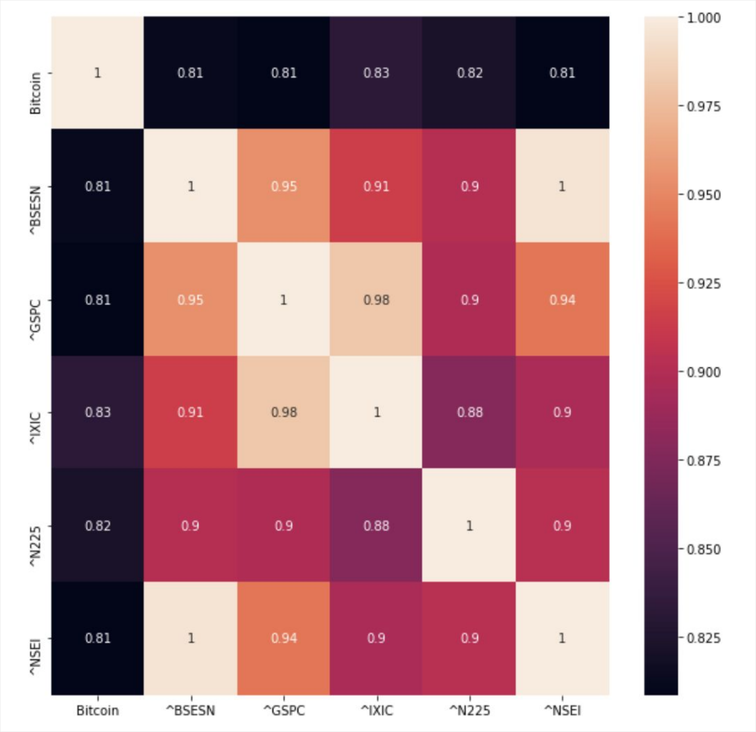
1697 rows × 11 columns

Data Cleaning - Combine the dataframes



Data Cleaning - Indices with >0.8 correlation with BTC

1) BSESN
2) GSPC
3) IXIC
4) N225
5) NSEI



Data Cleaning – Drop rows with “NaN”

Before dropping

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 1697 entries, 2014-09-17 to 2021-03-24
Data columns (total 6 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   Bitcoin  1697 non-null   float64
 1   ^BSESN   1595 non-null   float64
 2   ^GSPC    1641 non-null   float64
 3   ^IXIC    1641 non-null   float64
 4   ^N225    1591 non-null   float64
 5   ^NSEI    1595 non-null   float64
dtypes: float64(6)
memory usage: 92.8 KB
```

After dropping

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 1457 entries, 2014-09-17 to 2021-03-24
Data columns (total 6 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   Bitcoin  1457 non-null   float64
 1   ^BSESN   1457 non-null   float64
 2   ^GSPC    1457 non-null   float64
 3   ^IXIC    1457 non-null   float64
 4   ^N225    1457 non-null   float64
 5   ^NSEI    1457 non-null   float64
dtypes: float64(6)
memory usage: 79.7 KB
```

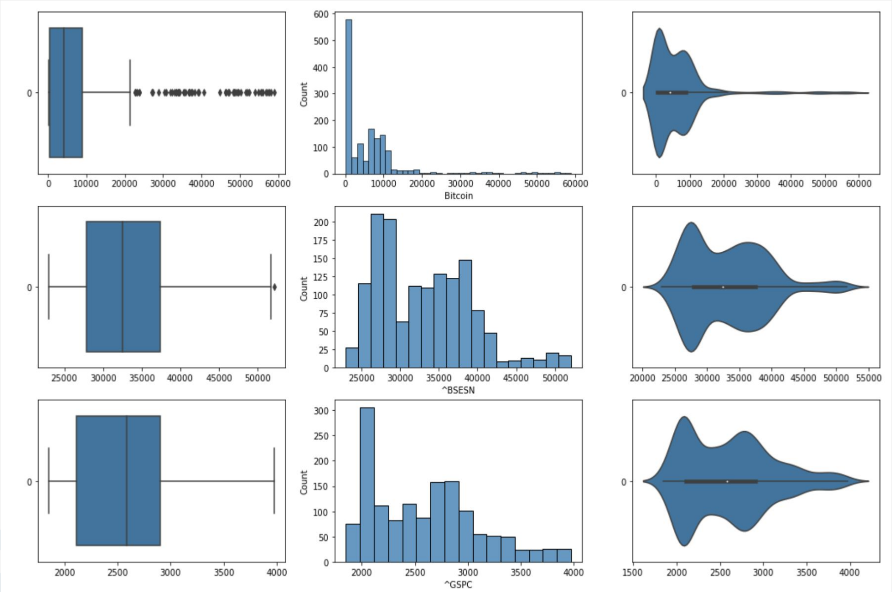
Data Cleaning - Final Dataframe

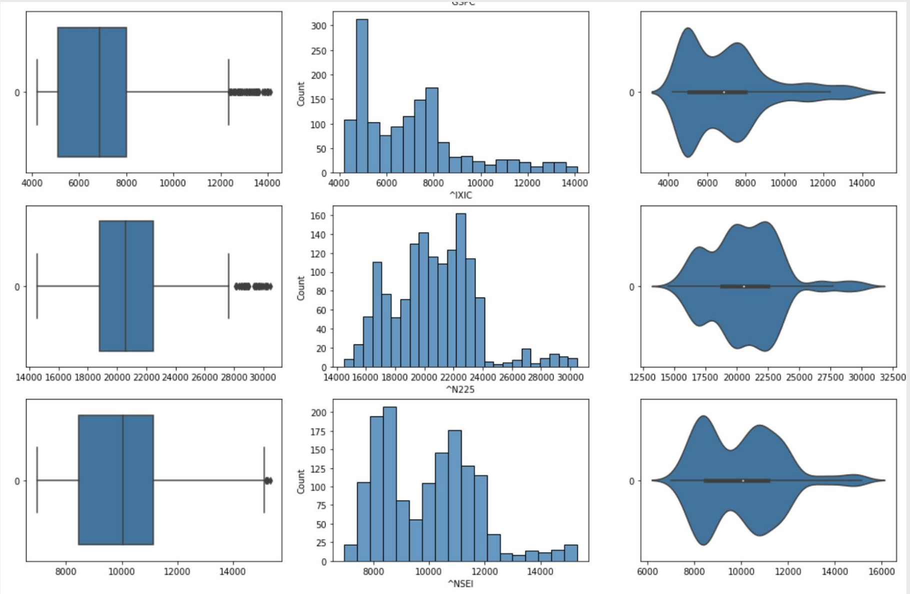
Date	Bitcoin	^BSESN	^GSPC	^IXIC	^N225	^NSEI
2014-09-17	457.334015	26631.289062	2001.569946	4562.189941	15888.669922	7975.500000
2014-09-18	424.440002	27112.210938	2011.359985	4593.430176	16067.570312	8114.750000
2014-09-19	394.795990	27090.419922	2010.400024	4579.790039	16321.169922	8121.450195
2014-09-22	402.152008	27206.740234	1994.290039	4527.689941	16205.900391	8146.299805
2014-09-24	423.204987	26744.689453	1998.300049	4555.220215	16167.450195	8002.399902
...	...	...	...	...	...	...
2021-03-18	57648.160000	49216.519531	3915.459961	13116.169922	30216.750000	14557.849609
2021-03-19	58030.010000	49858.238281	3913.100098	13215.240234	29792.050781	14744.000000
2021-03-22	54083.250000	49771.289062	3940.590088	13377.540039	29174.150391	14736.400391
2021-03-23	54340.890000	50051.441406	3910.520020	13227.700195	28995.919922	14814.750000
2021-03-24	52303.650000	49180.308594	3889.139893	12961.889648	28405.519531	14549.400391

1457 rows × 6 columns

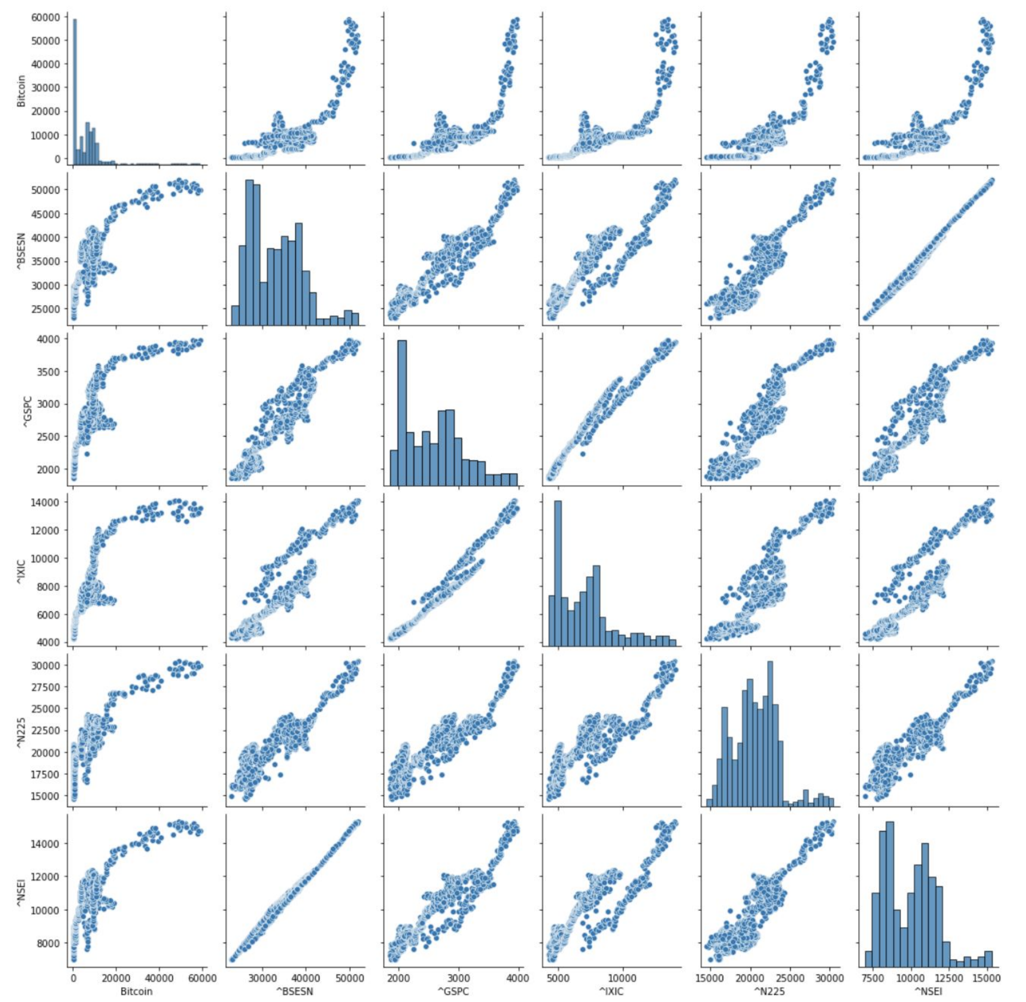
02
Exploratory Analysis

Boxplot, Histogram, ViolinPlot

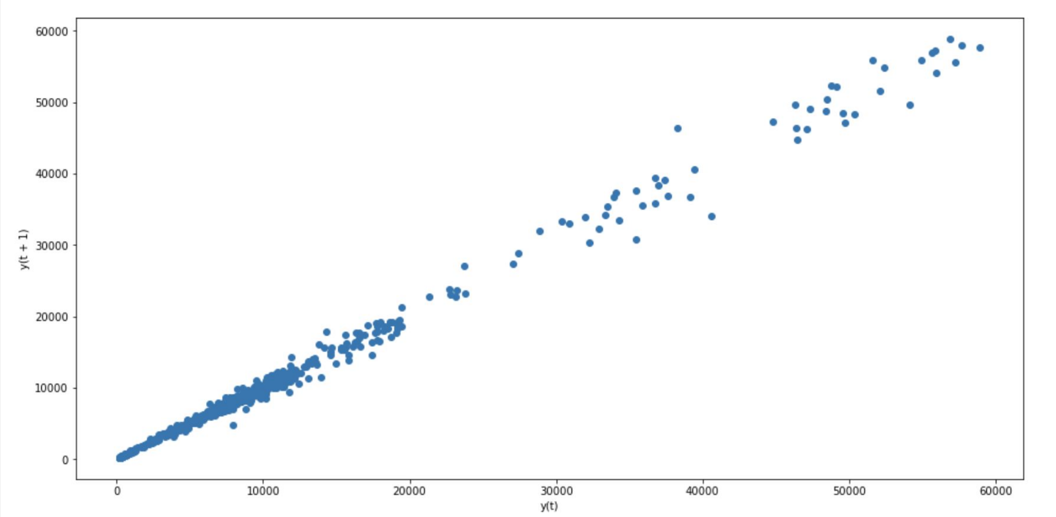




Pairplot



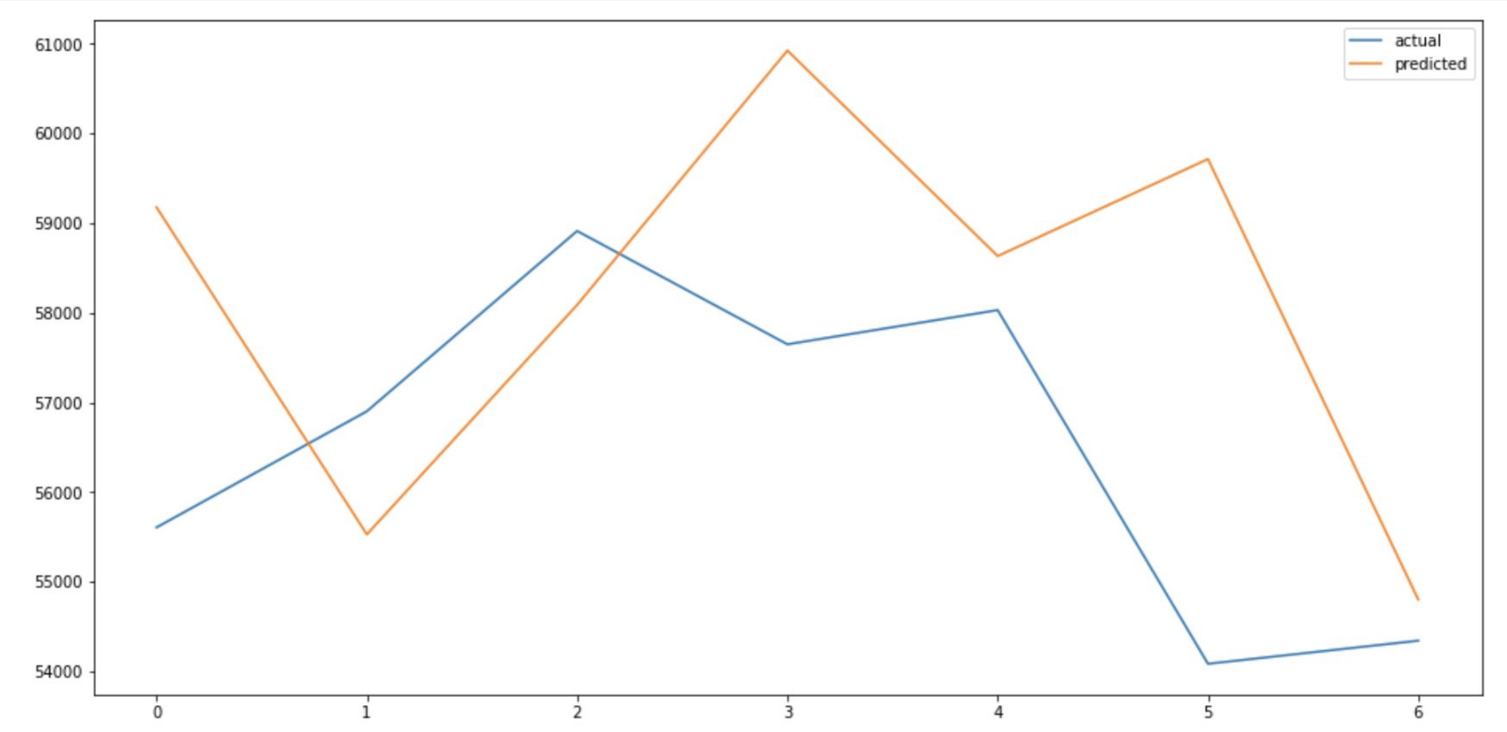
Autocorrelation – Lag plot



	t-1	t+1
t-1	1.000000	0.997519
t+1	0.997519	1.000000

Bitcoin is not affected randomly and has a correlation across time

Univariate Autoregression



RMSE:
2887.63

03
ML Models
Vector AutoRegression

ML Models

Multivariate Vector AutoRegression (VAR)

Univariate Long Short Term Memory (LSTM)

Multi-Variate Long Short Term Memory (LSTM)

Response Variable: Bitcoin
Predictor Feature: 5 indices

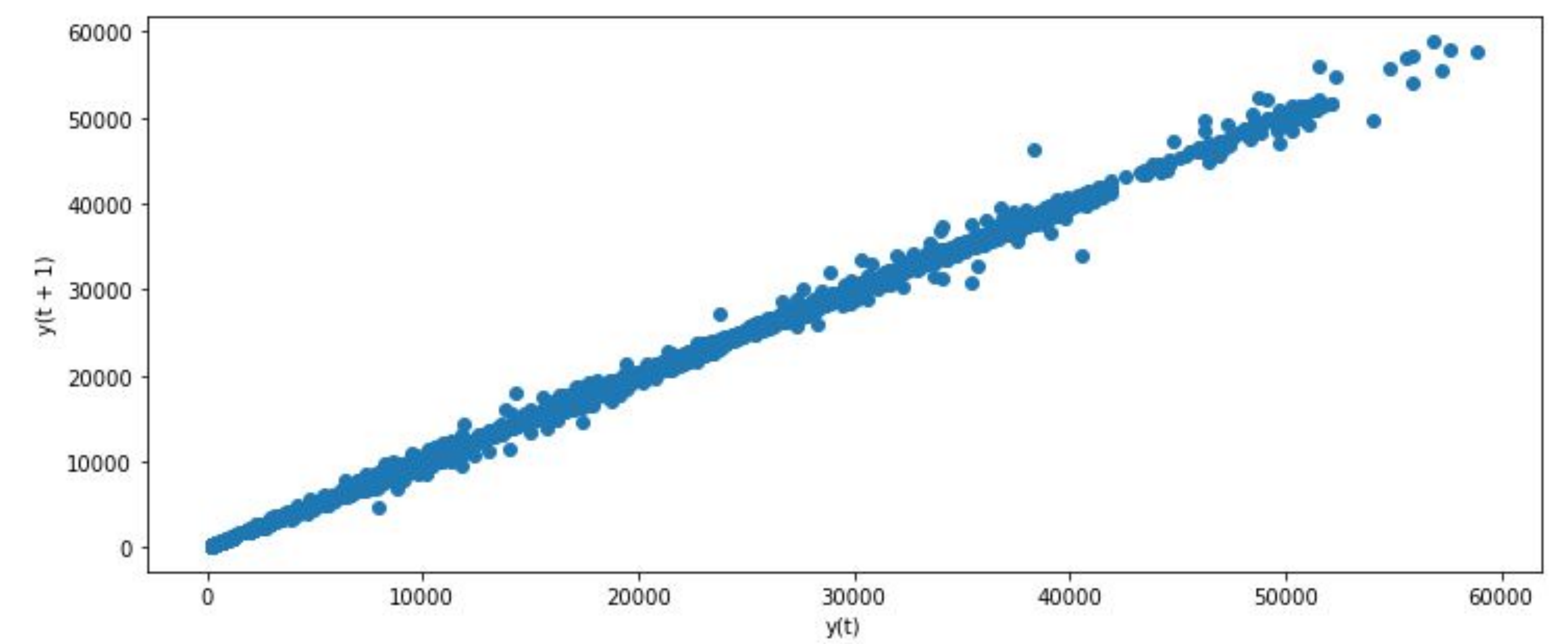
Linear Regression

- Tried to use linear regression
- Failed as the dimensions of input data were hard to control

Multivariate Autoregression

- Multivariate forecasting algorithm
- Takes lagged values of indices to predict future values of bitcoin

Autocorrelation – Lag plot



Multivariate Autoregression

```
# Create the train data set
X_train = combined.head(int(len(combined) - 7))
X_train
```

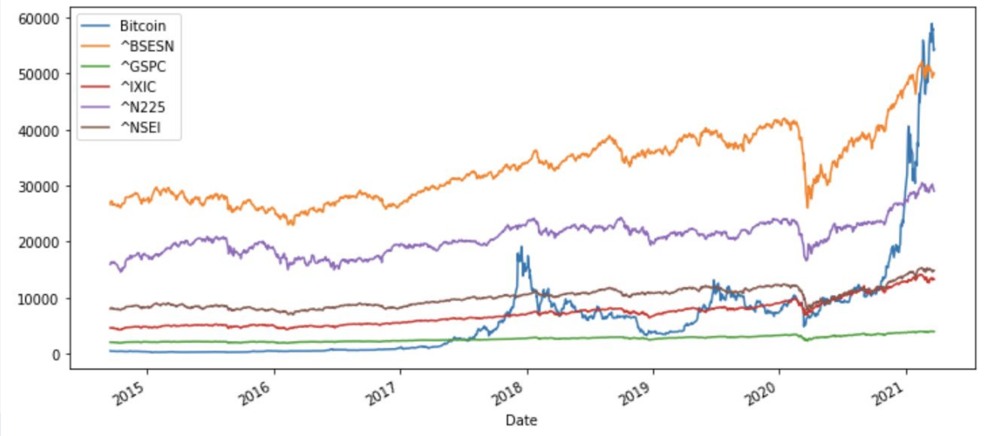
```
# Create the test data set
X_test = combined.tail(7)
X_test
```

```
# Create a copy of the train data set and find the differences
# between the current and its previous row for every row
X_train = X_train.copy()
X_train_diff =(X_train).diff().dropna()
X_train_diff.describe()
```

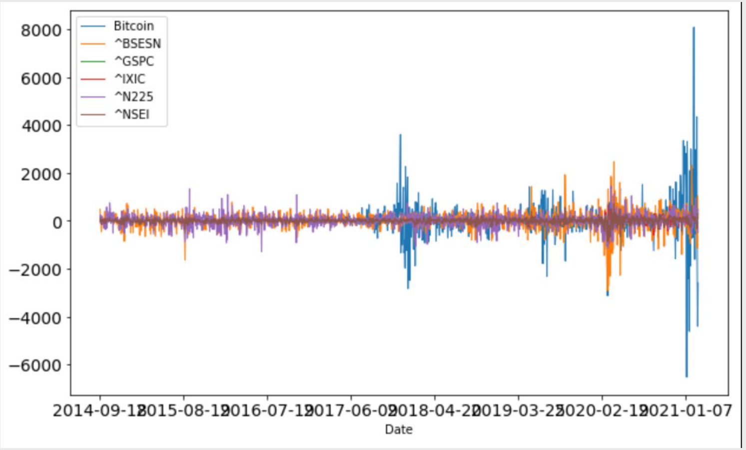
Multivariate Autoregression

Augmented Dickey Fuller Test:

Non-Stationary (p-value > 0.05)



Stationary



Multivariate Autoregression

Granger Causality Test:

Determine if the 5 indices are useful in forecasting price of Bitcoin

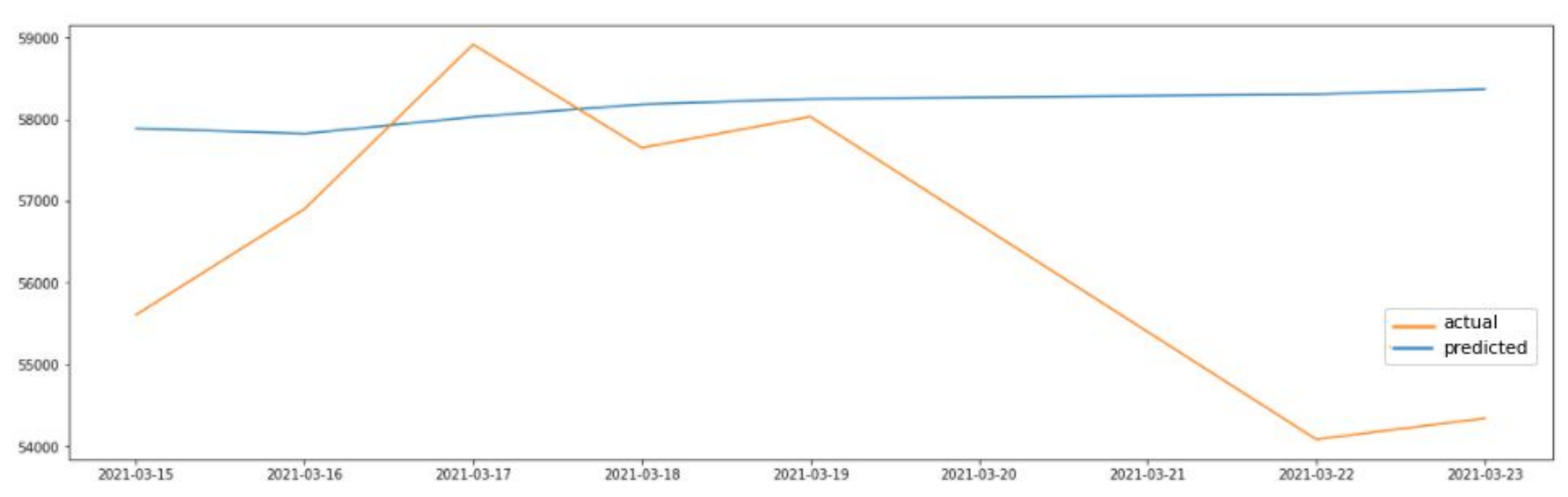
Although the correlation isn't high, the team decided to proceed with the model to see if the prediction is accurate

Correlation matrix of residuals

	Bitcoin	^BSESN	^GSPC	^IXIC	^N225	^NSEI
Bitcoin	1.000000	0.065217	0.142411	0.165703	0.067554	0.065787
^BSESN	0.065217	1.000000	0.407709	0.342432	0.344028	0.994781
^GSPC	0.142411	0.407709	1.000000	0.925500	0.347975	0.402952
^IXIC	0.165703	0.342432	0.925500	1.000000	0.299192	0.336072
^N225	0.067554	0.344028	0.347975	0.299192	1.000000	0.348288
^NSEI	0.065787	0.994781	0.402952	0.336072	0.348288	1.000000

Multivariate Autoregression

RMSE:
2425.69



04
ML Models

Long Short Term Memory ANN

LSTM Model

- Artificial Recurrent Neural Network
- Uses machine learning to predict the prices of bitcoin
- Uses past information to increase model performance
- Resistant to fluctuations of inputs that are random

Univariate and Multivariate LSTM

Scaling to make data set more manageable

```
# Scale the data for bitcoin
scaler = MinMaxScaler(feature_range=(0,1))
BTC = np.array(combined[['Bitcoin']])
scaled_data = scaler.fit_transform(BTC)
scaled_data
```

Splitting data set

```
# Creating the training dataset
training_data = scaled_data[:training_data_len, :]

# Split the data into x_train and y_train
x_train = []
y_train = []

for i in range(30, len(training_data)):
    x_train.append(training_data[i-30:i,0])
    y_train.append(training_data[i, 0])
```

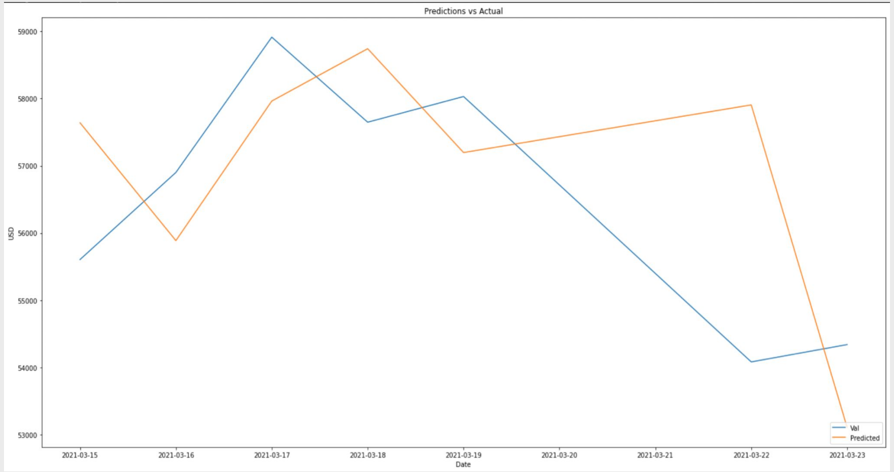
```
# Create the testing dataset
test_data = scaled_data[training_data_len - 30:, :]

# Create testing datasets: x_test, y_test
x_test = []
y_test = BTC[training_data_len:,:]

for i in range(30, len(test_data)):
    x_test.append(test_data[i-30:i,0])
```

Univariate LSTM

RMSE: 417.09

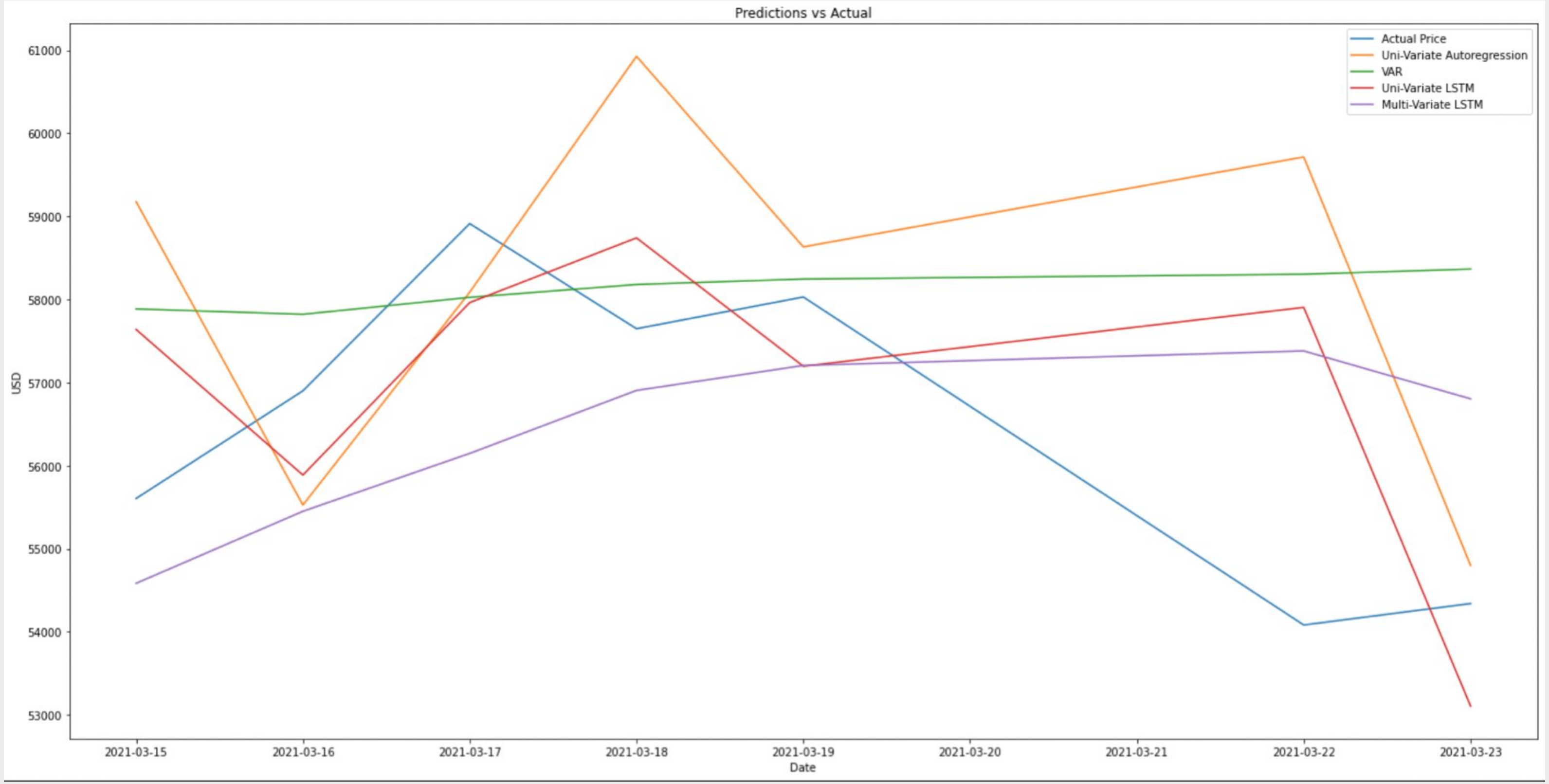


Multivariate LSTM

RMSE: 148.57

05 Conclusion

Conclusion



Predictions vs Actual
Actual Price
Uni-Variate Autoregression
VAR
Uni-Variate LSTM
Multi-Variate LSTM
61000
60000
59000
58000
57000
56000
55000
54000
53000
USD
2021-03-15
2021-03-16
2021-03-17
2021-03-18
2021-03-19
2021-03-20
2021-03-21
2021-03-22
2021-03-23
Date

Conclusion

Problem 1: To what extent is the price of Bitcoin dependent on the global financial system that is represented through stock indices

	Autoregression	VAR	Uni-LSTM	Multi-LSTM
Accuracy	102.94	97.22	100.79	99.84
RMS Error	$2887.63	$2425.69	$417.09	$148.57
Mean Forecast Error	-1620.73	1615.65	-417.09	148.57

Conclusion

Problem 1: Multivariate LSTM vs Univariate LSTM

Multivariate LSTM has a:

1) Better accuracy
2) Lower RMS Error
3) Lower Mean forecast Error

Multivariate models are using indices for prediction while the univariate ones are only using past prices of Bitcoin

Bitcoin is dependent on the global financial market based on prediction through indices

Conclusion

Problem 2: Multivariate LSTM vs VAR

Multivariate LSTM has a:

1) Better accuracy
2) Lower RMS error
3) Lower Mean Forecast Error

Multivariate LSTM is a better model than VAR to predict the price of Bitcoin using stock indices

Conclusion

Interesting fact

VAR: Only able to predict price of Bitcoin over a few days

LSTM: Able to predict accurately over a few months

LSTM capable of predicting long term dependencies because of its recurrent nature

Contributions

Omkar

Extracted and Cleaned data set

Basic Exploratory Analysis

Correlation to choose stock indices (heatmap)

Research on ML models

Successfully created Multivariate VAR model, Uni and Multivariate LSTM models

Wynne

Extracted and Cleaned data set

Basic Exploratory Analysis

Box plot, Histogram, Violin plot, Pair plot, and Heat map for chosen indices

Linear regression

Research on ML models

Himari

Extracted and Cleaned data set

Basic Exploratory Analysis

Linear regression

Univariate VAR model

Research on ML models

Successfully created Univariate VAR model

References:

VAR:

- https://otexts.com/fpp2/causality.html
- https://towardsdatascience.com/vector-autoregressive-for-forecasting-time-series-a60e6f168c70
- https://www.kaggle.com/sunithaak/guidance-on-vector-auto-regression-for-beginner-s
- https://www.machinelearningplus.com/time-series/vector-autoregression-examples-python/
- https://www.kaggle.com/lokeshkumarn/autoregression-model
- https://machinelearningmastery.com/autoregression-models-time-series-forecasting-python/
- https://towardsdatascience.com/time-series-forecasting-with-autoregressive-processes-ba629717401

LSTM:

- https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/
- https://towardsdatascience.com/predictive-analytics-time-series-forecasting-with-gru-and-bilstm-in-tensorflow-87588c852915
- https://laptrinhx.com/ann-classification-model-evaluation-and-parameter-tuning-3333931647/
- https://github.com/nilabja-bhattacharya/Cryptocurrency-Price-Prediction
- https://github.com/shreyas-muralidhara/Bitcoin-price-prediction
- https://medium.com/analytics-vidhya/rnn-vs-gru-vs-lstm-863b0b7b1573

Thank you!