SENTIMENT ANALYSIS OF CRYPTOCURRENCY TWEETS: PREDICTING BITCOIN MARKET SENTIMENT AND PRICE MOVEMENTS USING MACHINE LEARNING

Amarpreet Kaur
St. No: 501213603

Toronto Metropolitan University

SENTIMENT ANALYSIS OF CRYPTOCURRENCY TWEETS: PREDICTING BITCOIN MARKET SENTIMENT AND PRICE MOVEMENTS USING MACHINE LEARNING

Amarpreet Kaur

Student Id. 501213603

Toronto Metropolitan University, ON

Supervisor: Tamer Abdou

Date of Submission: Jul 24, 2023

Toronto Metropolitan University

INTRODUCTION

Sentiment analysis of cryptocurrency-related tweets, with a focus on Bitcoin, allows us to uncover market sentiment and public perception, providing valuable insights for making informed decisions and investments in the dynamic world of digital currencies.

Toronto Metropolitan University

DATA SOURCES

- **From Kaggle** Tweets - 1,69,761 rows and 13 columns and

  Historical data - 12,037 columns and 10 rows
- **From CryptoCompare API** for live data- Bitcoin price data

To predict sentiment, price and change in price based one sentiments (tweets) and to predict price based on historical data.

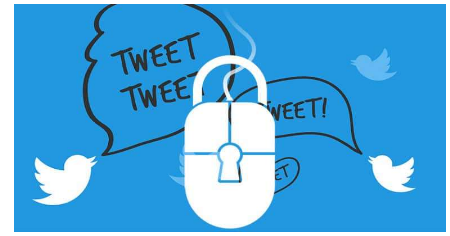
Toronto Metropolitan University

EXECUTIVE SUMMARY
PROBLEM STATEMENTS

1. How does sentiment analysis of tweets relating to Bitcoin correlate with the price dynamics of Bitcoin, specifically **examining the influence of sentiment on Bitcoin's price change**?
2. How do different machine learning algorithms, such as Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest, Logistic Regression, and Gradient Boosting, **perform in sentiment prediction** for cryptocurriencies-related tweets on Twitter, and which algorithm(s) yield the highest prediction accuracy?
3. What is the comparative effectiveness of various machine learning regression models, including Multiple Linear Regression, Decision Tree Regressor, Random Forest Regressor, Support Vector Regressor, Gradient Boosting Regressor, and Neural Network Regressor, in **predicting cryptocurriencies prices and changes in price (delta) based on sentiments from tweets**?
4. How does the utilization of Time Series Models improve the accuracy of **cryptocurriencies price prediction** for Bitcoin (BTC), Ethereum (ETH), Ripple (XRP), Litecoin (LTC), and USD Coin (USDC) based on their historical data? What are the potential limitations and challenges associated with using Time Series Models for cryptocurriencies price prediction?

Toronto Metropolitan University

EXECUTIVE SUMMARY
SOLUTION



1. **Sentiment Calculation** – **VADER and Text Blob**
2. **Sentiment prediction** – **Five Classification Models**
3. **Price and Price change (Delta) Predictions based on Sentiment** (Sentiment as independent attribute)- **Six Regression Models**
4. **Prediction of five cryptocurrencies based on Historical Data** – **Four Time Series Models**

EXECUTIVE SUMMARY
TOOLS USED

The chosen programming language for this project is **Python**, which offers a wide range of libraries for data analysis and model building.

EXECUTIVE SUMMARY
CONCLUSION

1. The analysis of the tweet scores, including the **lag, magnitude, and slope**, in relation to the changes in Bitcoin prices provides valuable insights into the dynamics of Twitter sentiment and its impact on the cryptocurrency market.
2. **The Support Vector Machines (SVM)** model consistently performed well across various metrics and demonstrated strong performance with high accuracy based on cross-validation results.
3. **Random Forest and Gradient Boosting regressors** exhibit better performance in prediction price as well as price change (delta).
4. Some currencies may benefit from using SMA, whereas ARIMA, LSTM, and Prophet **each possess unique strengths and weaknesses.**

APPROACH

1. **Data Collection**
   - Tweets Dataset
   - Crypto Dataset

2. **Preprocessing**
   - Tweets Cleaning
   - Sentiment Analysis with VADER AND Text Blob
   - Data Preparation & EDA

3. **Sentiment Analysis**
   - Sentiment Score Vs Price
   - Sentiment Score Vs Price Change

4. **Prediction**
   - Models, evaluation, and Cross Validation
   - Sentiment
   - Price
   - Price Change

5. **Results Interpretation**
   - Conclusion
   - Recommendations

Toronto Metropolitan University

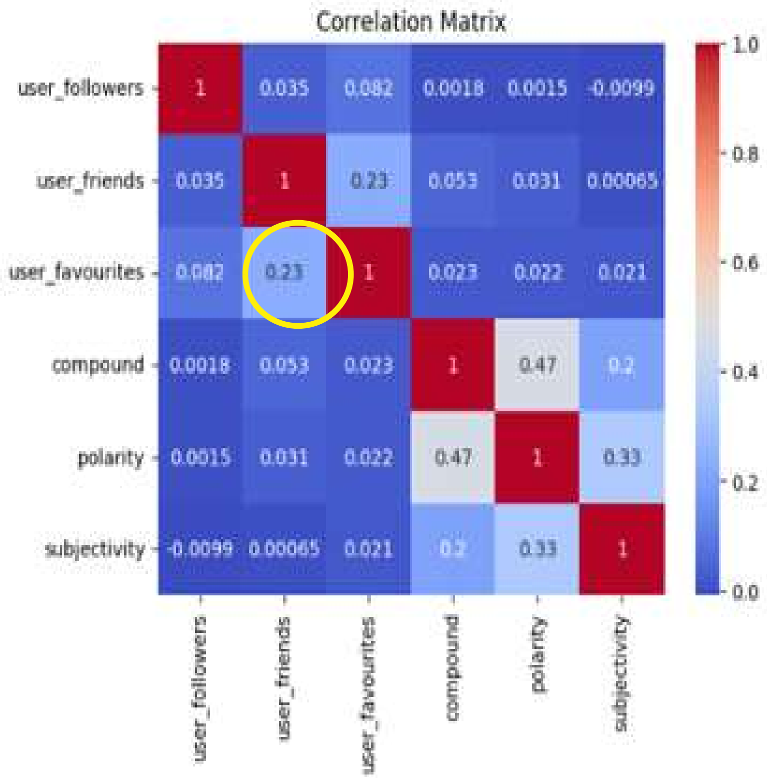
EXPLORATORY DATA ANALYSIS – (EDA) TWEETS

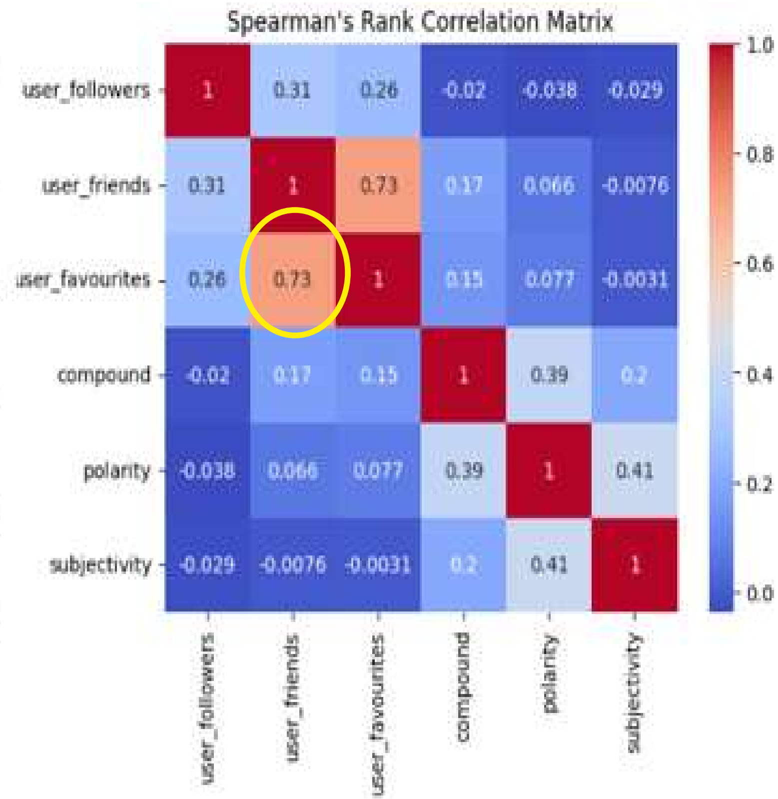
Data Dictionary

No	Columns	Descriptions	Data Type
1	user_name	The name of the user, as they've defined it.	object (string)
2	user_location	The user-defined location for this account's profile.	object (string)
3	user_description	The user-defined UTF-8 string describing their account.	object (string)
4	user_created	Time and date, when the account was created.	object (string)
5	user_followers	The number of followers an account currently has.	float64 (number)
6	user_friends	The number of friends an account currently has.	float64 (number)
7	user_favourites	The number of favorites an account currently has	float64 (number)
8	user_verified	When true, indicates that the user has a verified account	bool (boolean)
9	date	UTC time and date when the Tweet was created	datetime64[ns]
10	text	The actual UTF-8 text of the Tweet	object (string)
11	hashtags	All the other hashtags posted in the tweet along with #Bitcoin & #btc	object (string)
12	source	Utility used to post the Tweet; Tweets from the Twitter website have a source value - web	object (string)
13	is_retweet	Indicates whether this Tweet has been Retweeted by the authenticating user.	float64 (number)

Toronto Metropolitan University

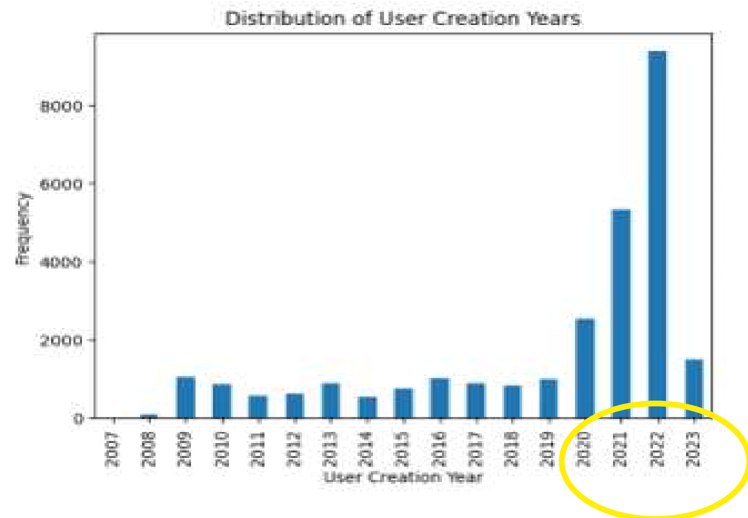
CORRELATION AND STATISTICAL SUMMARY

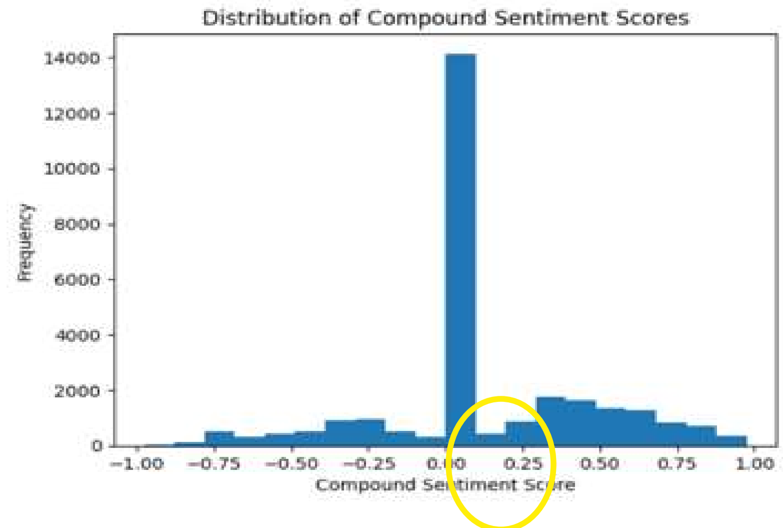


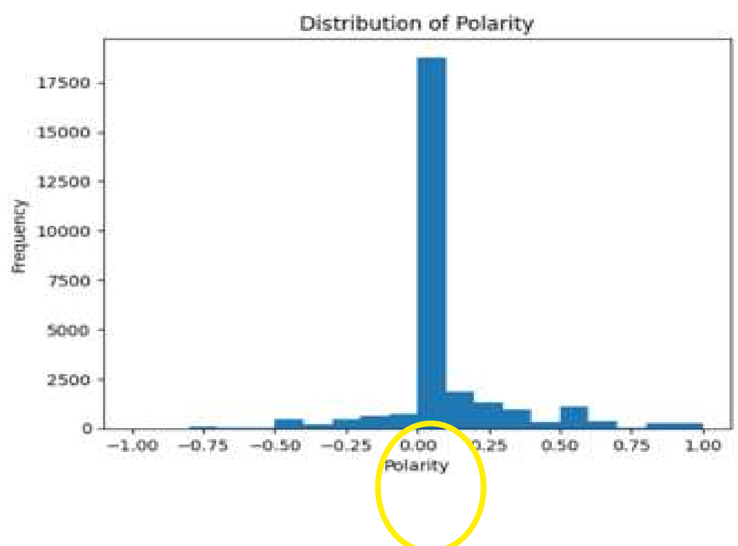


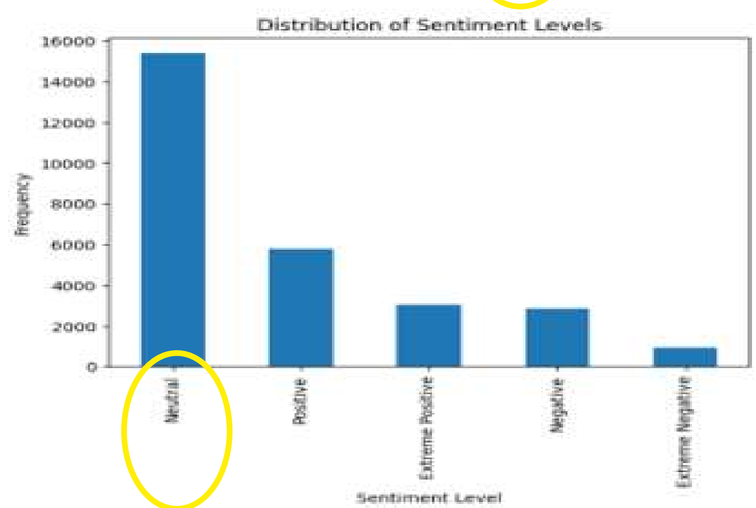
User friends and user favourites are highly correlated.

UNIVARIATE ANALYSIS





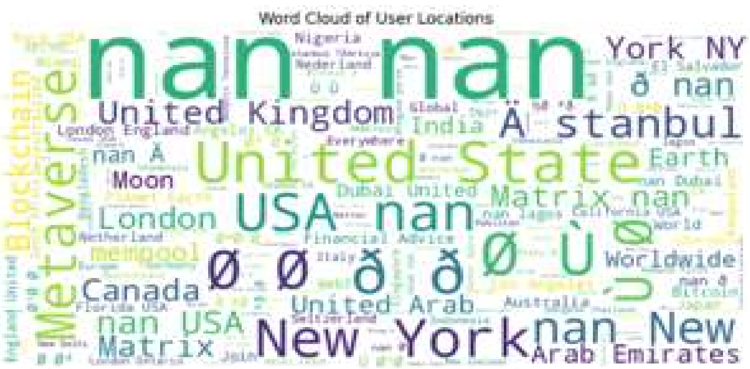


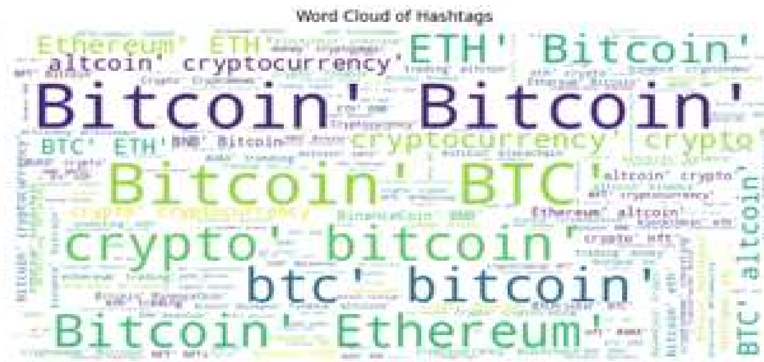


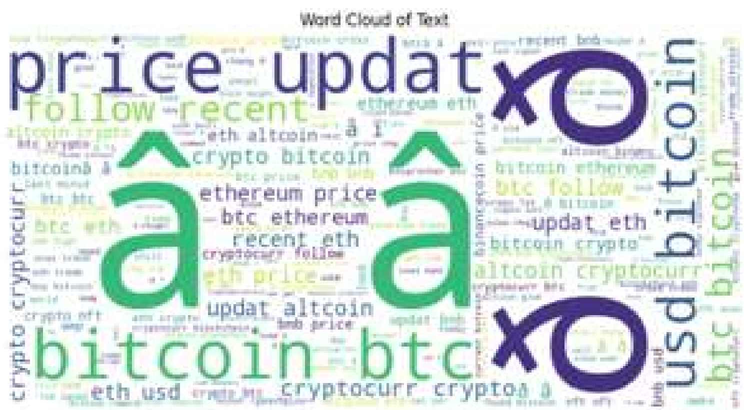
Toronto
Metropolitan
University

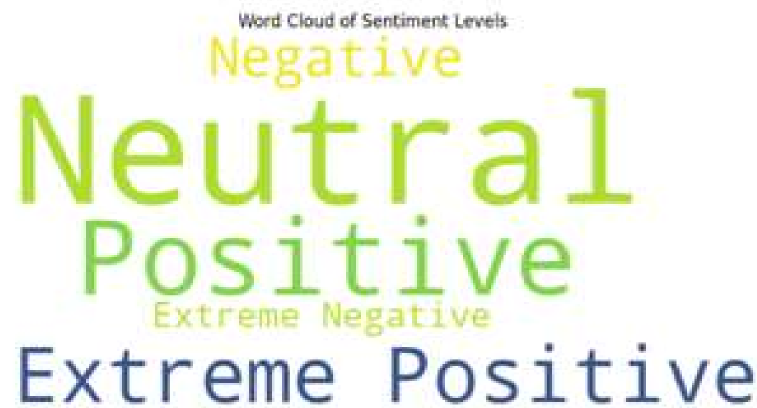
UNIVARIATE ANALYSIS

WORD CLOUD

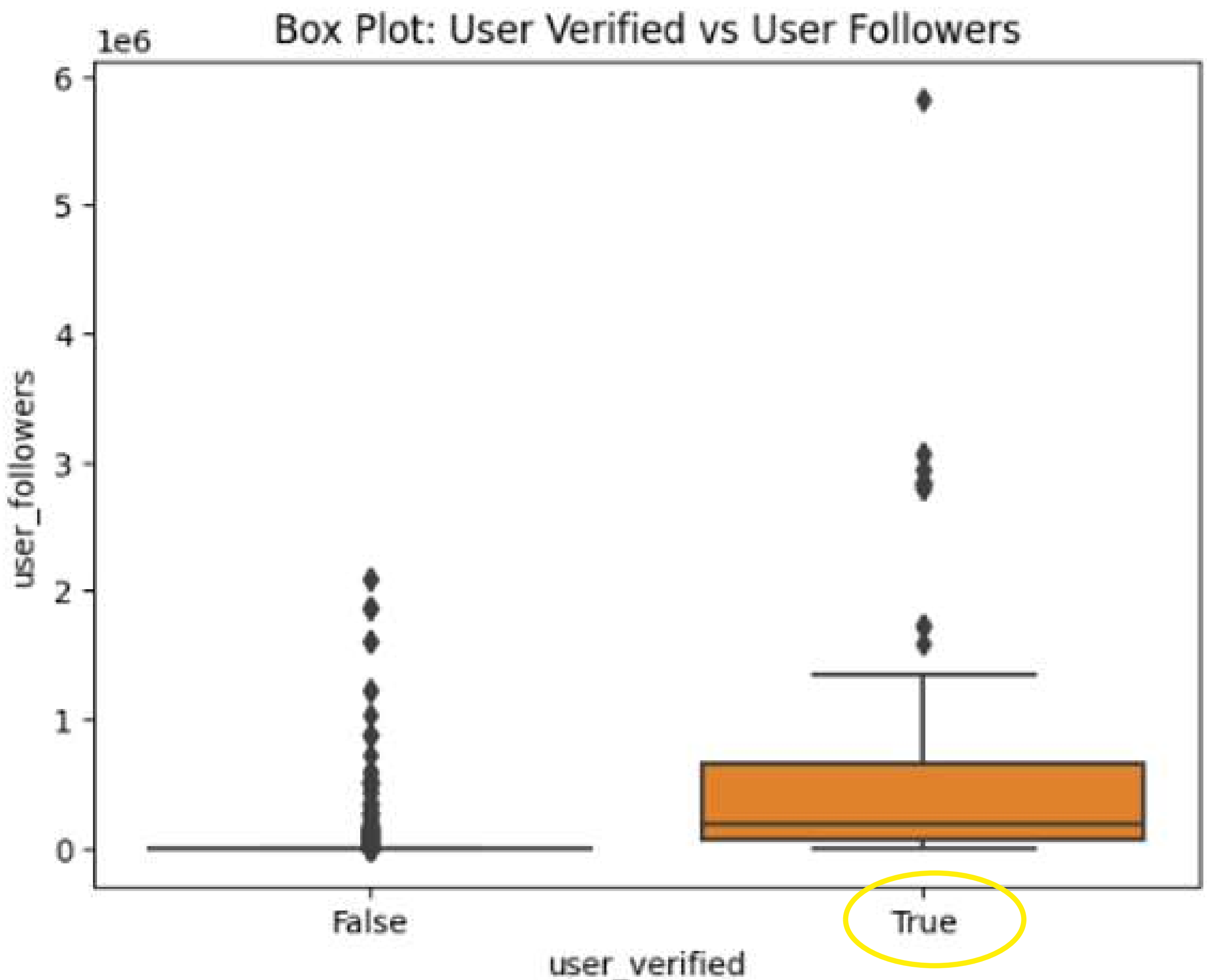




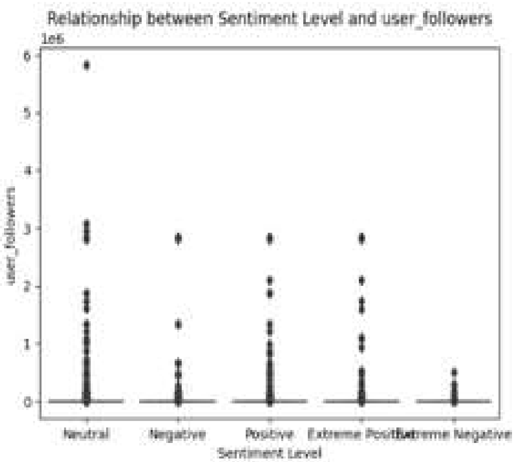


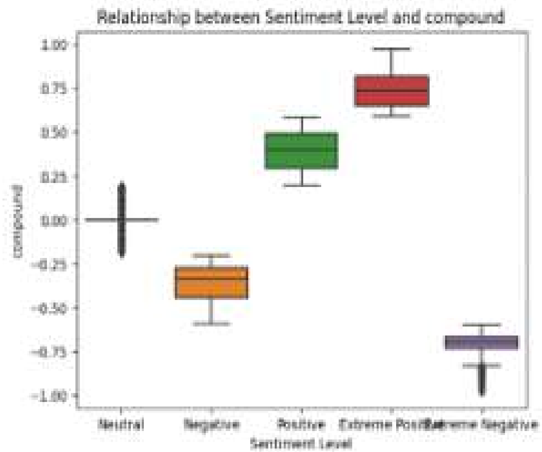


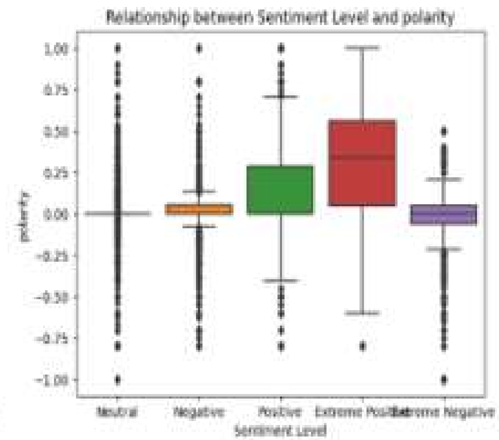
BIVARIATE ANALYSIS

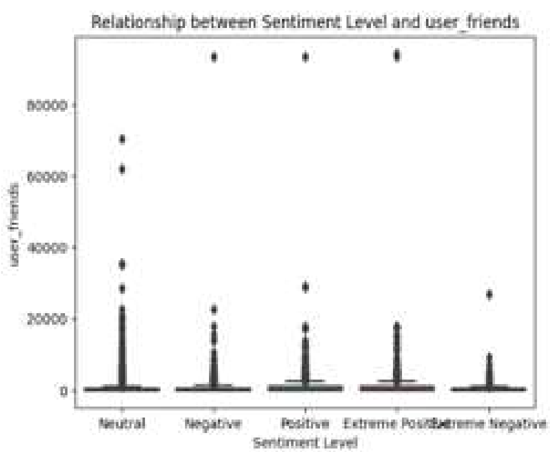


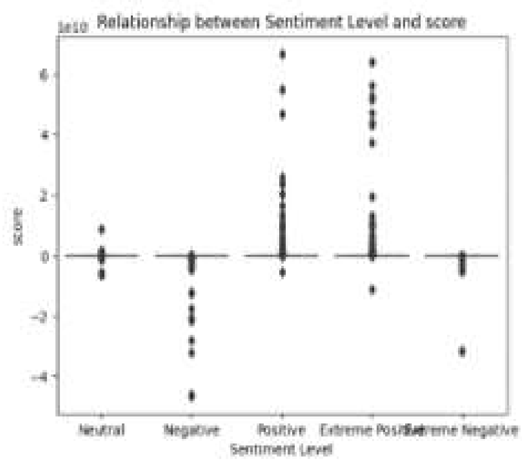
BIVARIATE ANALYSIS

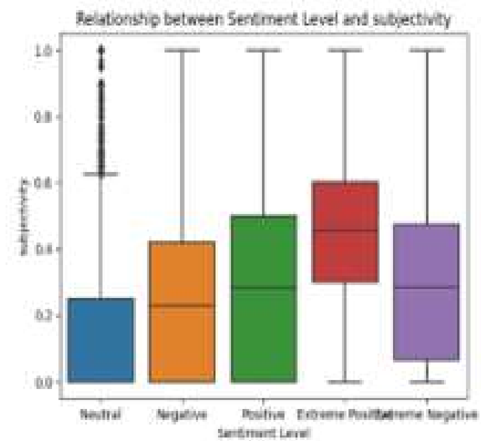








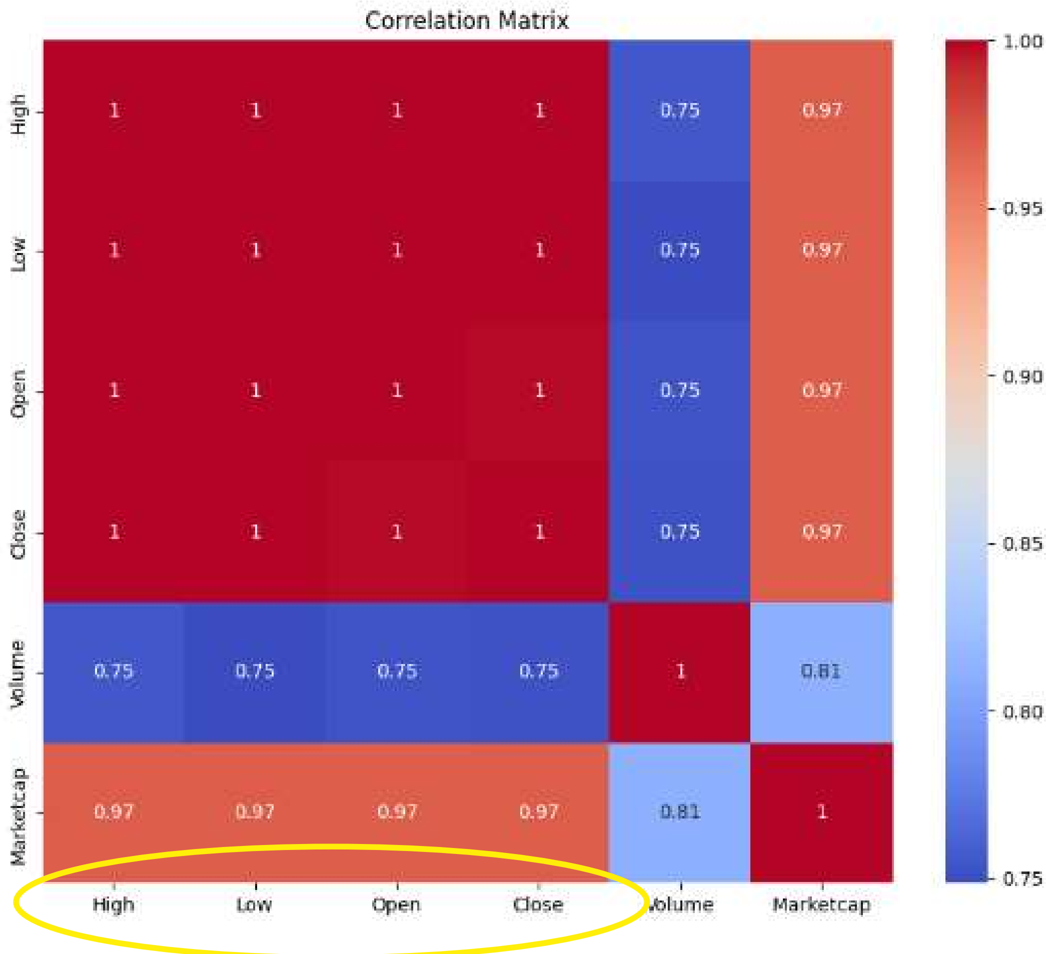




CRYPTOCURRENCIES
DATA DESCRIPTION

Column Name	Description	Variable Type	Data Type
SNo	Serial number	Numerical (Nominal)	int64
Name	Name of the crypto currency	Categorical (Nominal)	object
Symbol	The crypto currency symbol	Categorical (Nominal)	object
Date	The date of the recorded data point	Date (Dataframe)	object
High	The highest price reached by the crypto currency on the given date	Numerical (Continuous)	float64
Low	The lowest price reached by the crypto currency on the given date	Numerical (Continuous)	float64
Open	The opening price of the crypto currency on the given date	Numerical (Continuous)	float64
Close	The closing price of the crypto currency on the given date	Numerical (Continuous)	float64
Volume	The trading volume of the crypto currency on the given date	Numerical (Continuous)	float64
Marketcap	The market capitalization of the crypto currency on the given date	Numerical (Continuous)	float64

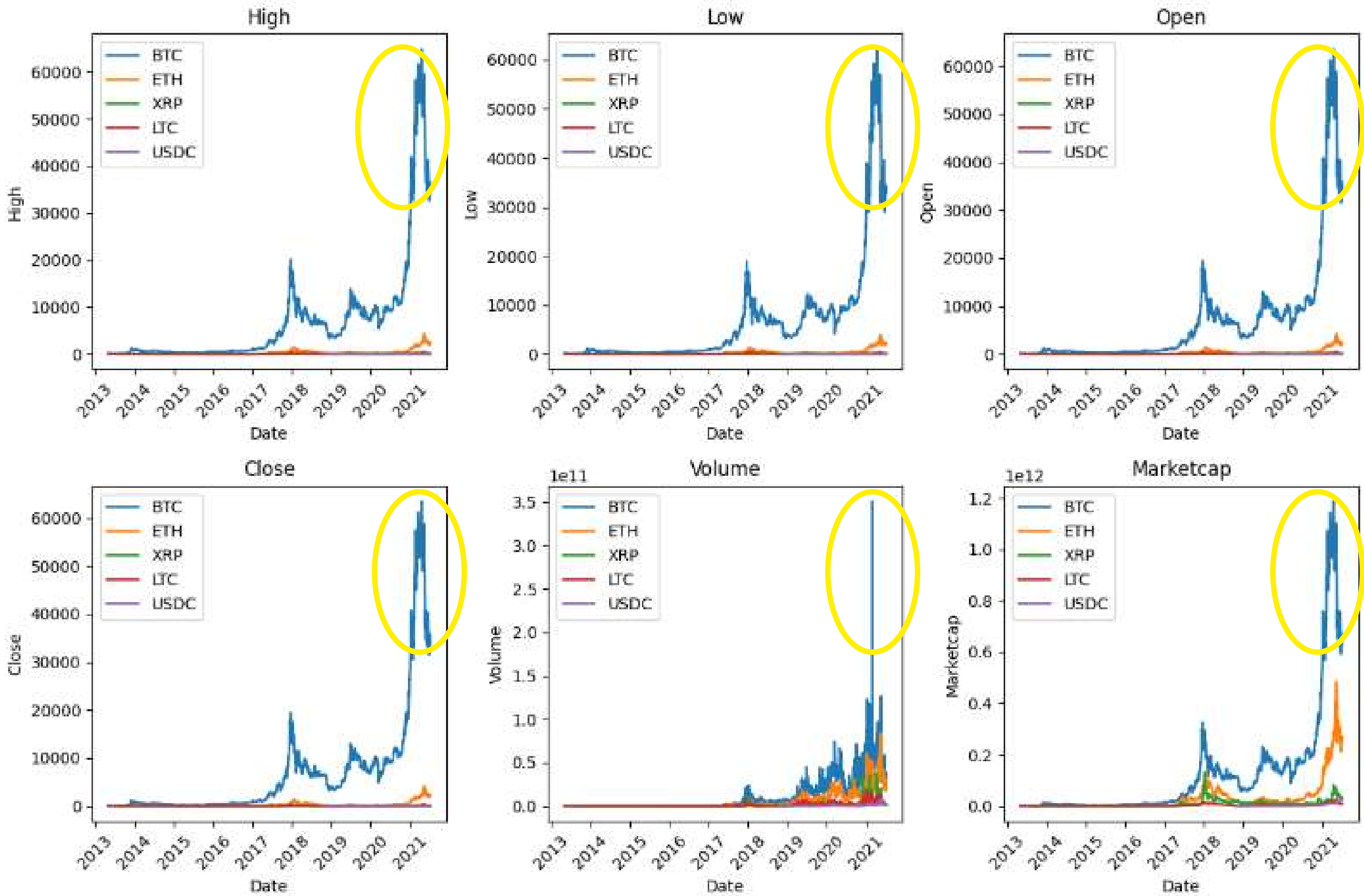
CORRELATION AND STATISTICAL SUMMARY



Toronto Metropolitan University

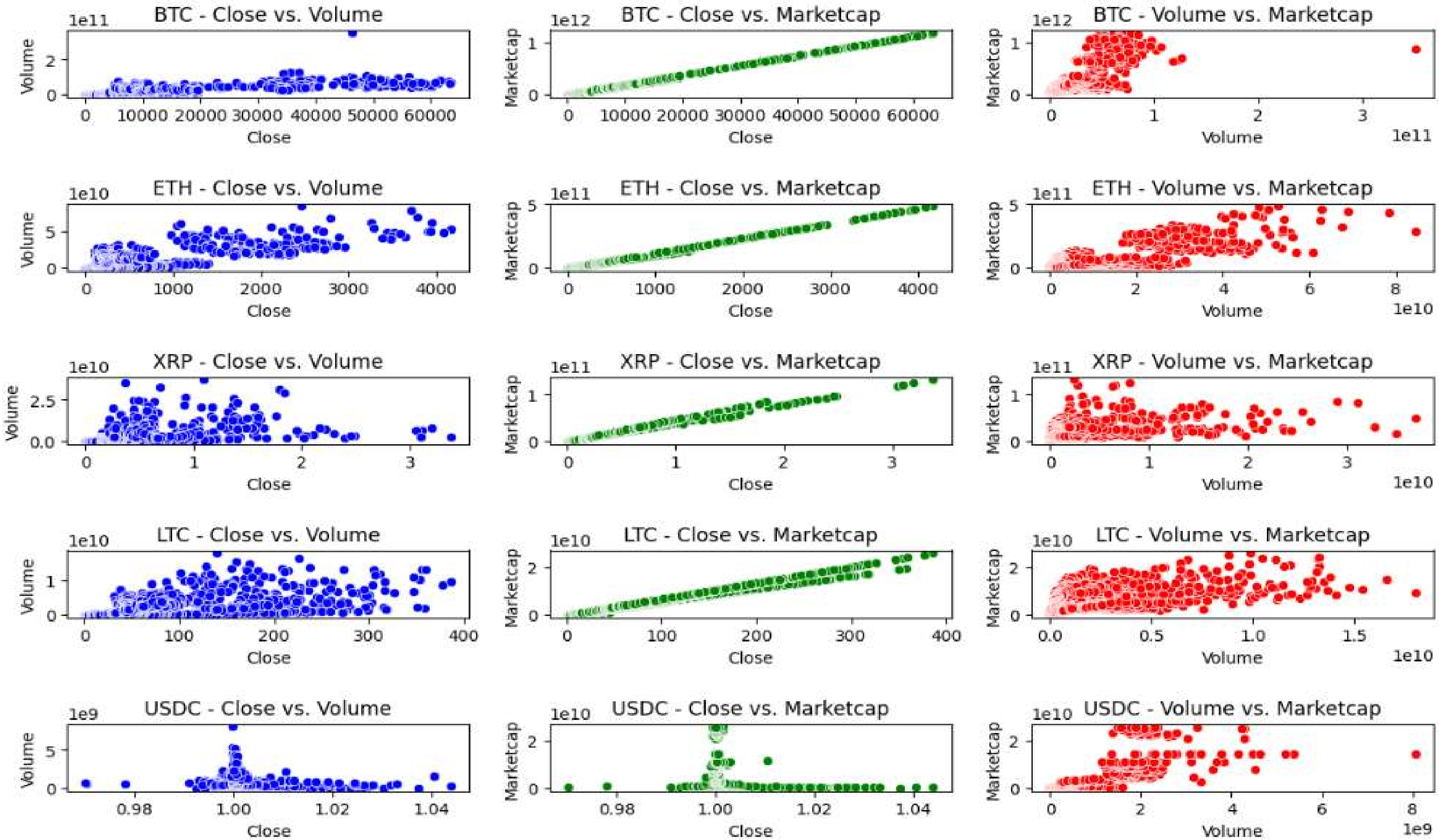
UNIVARIATE ANALYSIS

Time-series plot for each coin with each numeric column



Toronto Metropolitan University

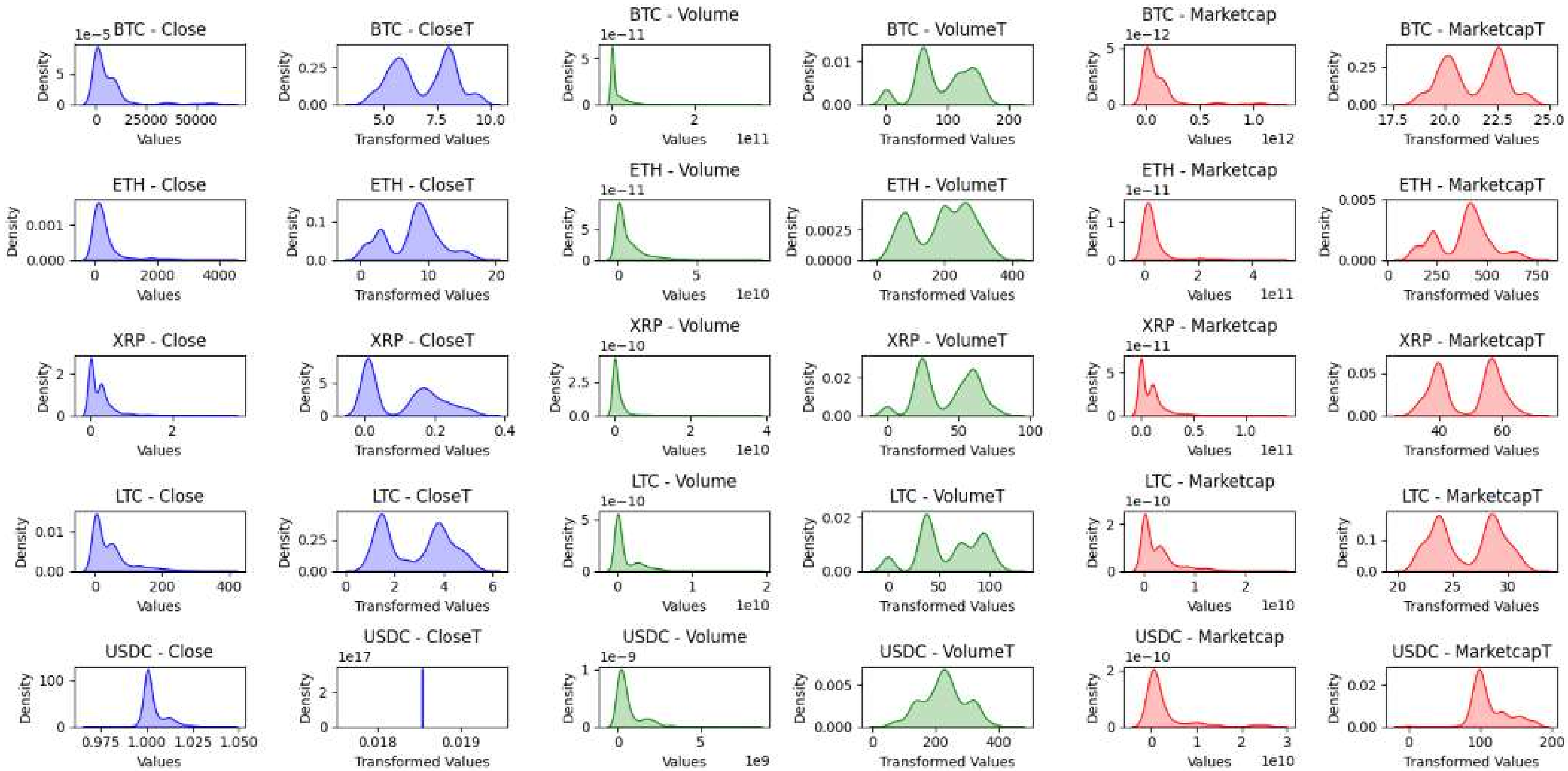
BIVARIATE ANALYSIS



Toronto Metropolitan University

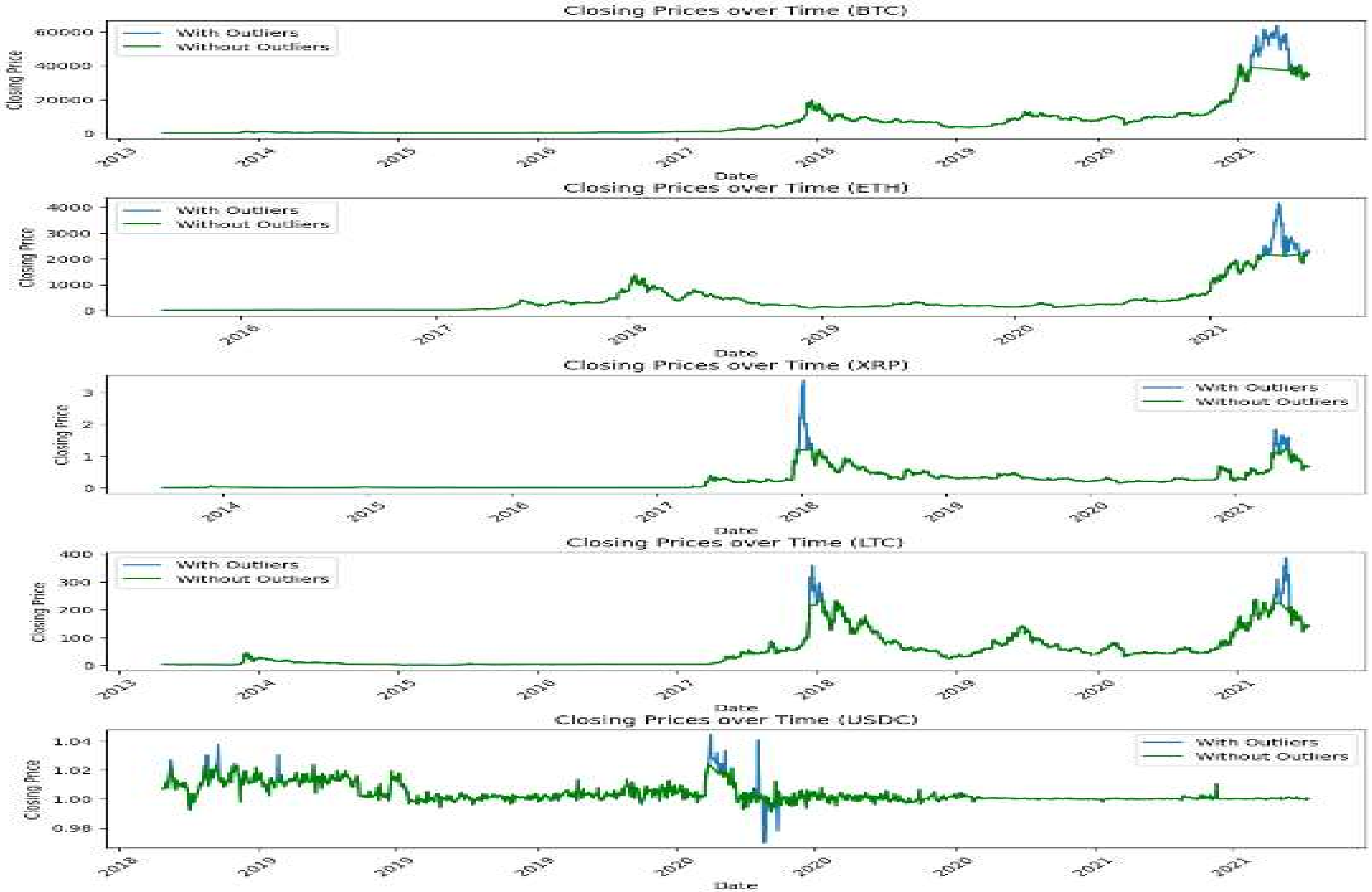
SKEWNESS

Plotting the original and transformed sidewise to compare.



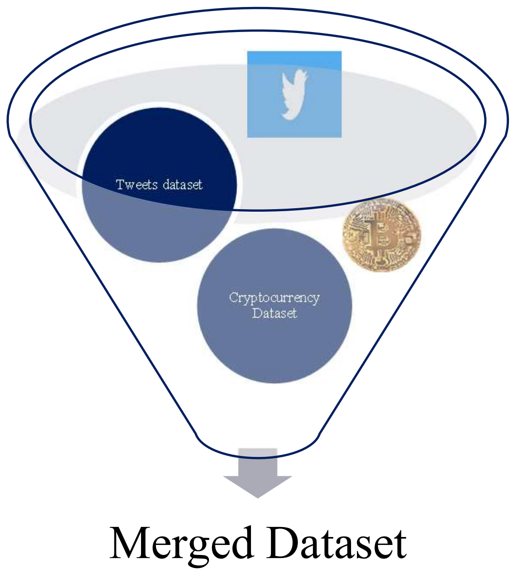
OUTLIERS ANALYSIS

Plotting with and without outliers

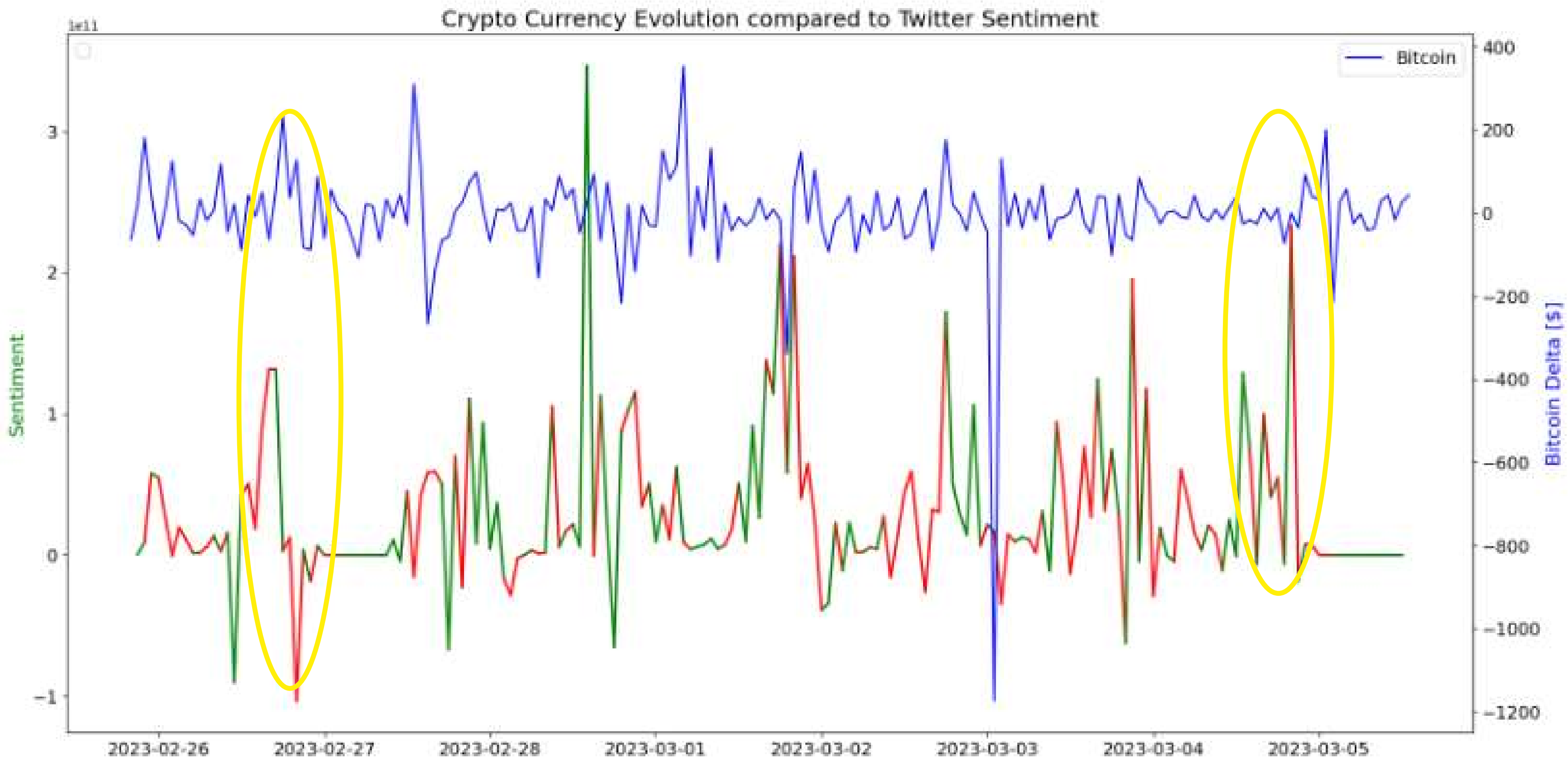


MERGING DATASETS

Upon gathering the data, I conducted **alignment on both the Bitcoin tweets and cryptocurrency data** by using predefined one-minute time windows. This ensured that the data points from both sources were synchronized to the same time intervals. The aligned data was then stored for subsequent processing.



SENTIMENT ANALYSIS
SENTIMENT VS PRICE CHANGE (DELTA)- SLOPE AND MAGNITUDE



Toronto Metropolitan University

CLASSIFICATION MODELS FOR SENTIMENT PREDICTION
FEATURE SELECTION

Term Frequency Matrix – Unigram, Bigram and Combined

```
Unigram
Count of 0 : 189953938
Count of 1 : 116293
Count of 2 : 8897
Count of 3 : 1795
Count of 4 : 142
Count of 5 : 39
Count of 6 : 55
Count of 7 : 1
Count of 8 : 5
Count of 9 : 0
Count of 10 : 0
Count of 11 : 0
Count of 12 : 0
Count of 13 : 0
```

```
Bigram
Count of 0 : 516760987
Count of 1 : 130418
Count of 2 : 1014
Count of 3 : 83
Count of 4 : 4
Count of 5 : 1
Count of 6 : 0
Count of 7 : 0
Count of 8 : 0
Count of 9 : 0
Count of 10 : 0
Count of 11 : 0
Count of 12 : 1
Count of 13 : 0
```

```
Combined
Count of 0 : 706714925
Count of 1 : 246711
Count of 2 : 9911
Count of 3 : 1878
Count of 4 : 146
Count of 5 : 40
Count of 6 : 55
Count of 7 : 1
Count of 8 : 5
Count of 9 : 0
Count of 10 : 0
Count of 11 : 0
Count of 12 : 1
Count of 13 : 0
```

CLASSIFICATION MODELS FOR SENTIMENT ANALYSIS

Classification Model	Accuracy	Precision	Recall	F1-Score	Predicted sentiment
Naive Bayes	79%	80%	79%	79%	Neutral
Support Vector Machines (SVM)	86%	86%	86%	86%	Neutral
Random Forest	86%	86%	86%	85%	Neutral
Logistic Regression	86%	86%	86%	85%	Neutral
Gradient Boosting	85%	85%	85%	84%	Neutral

RESULTS AND CONCLUSIONS BASED ON CLASSIFICATION MODELS FOR SENTIMENT PREDICTION

The differences in performance among the models were relatively small.
The Support Vector Machines (SVM) model consistently performed well across various metrics, achieving the highest accuracy, recall, and competitive precision and F1-score.
The Random Forest and Logistic Regression models also demonstrated competitive performance.

Toronto Metropolitan University

CROSS VALIDATION

Classification Model	Cross-Validation Scores	Accuracy
Naive Bayes	79%	79%
Support Vector Machines (SVM)	86%	86%
Random Forest	86%	86%
Logistic Regression	84%	85%
Gradient Boosting	84%	84%

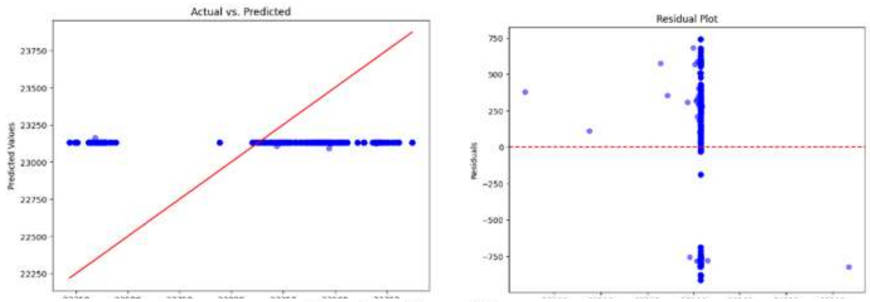
SVM and Random Forest demonstrate strong performance with high accuracy based on cross-validation results.

HYPERTUNNING PARAMETERS

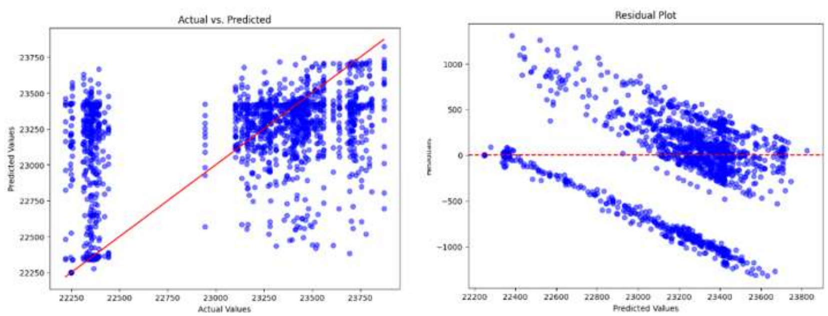
```
Model: Naive Bayes
Best Parameters: {'alpha': 1.0}
Cross-Validation Accuracy: 0.7736638954869359
Accuracy: 0.7639240506329114

Model: Support Vector Machine
Best Parameters: {'C': 10.0}
Cross-Validation Accuracy: 0.8809735710634715
Accuracy: 0.8848101265822785

Model: Random Forest
Best Parameters: {'n_estimators': 300}
Cross-Validation Accuracy: 0.8649886747446806
Accuracy: 0.8613924050632912

Model: Logistic Regression
Best Parameters: {'C': 10.0}
Cross-Validation Accuracy: 0.8732179009190496
Accuracy: 0.8816455696202532

Model: Gradient Boosting
Best Parameters: {'n_estimators': 300}
Cross-Validation Accuracy: 0.8779677430670395
Accuracy: 0.8822784810126583
```

REGRESSION MODELS FOR PRICE PREDICTION

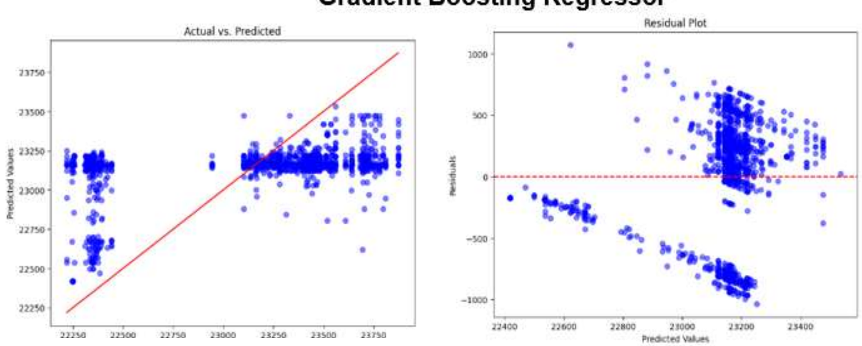
Multiple Linear Regression



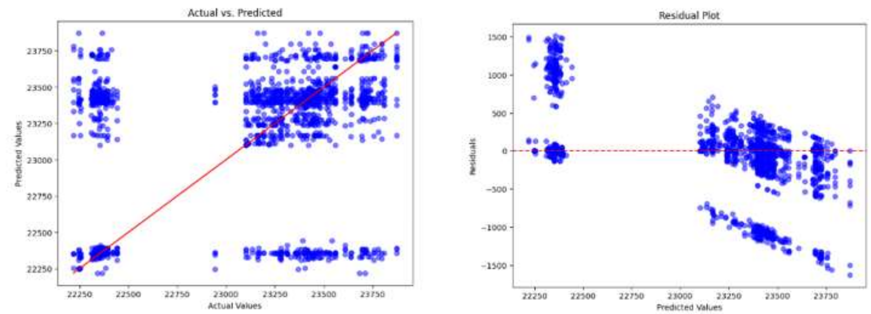
Random Forest Regressor



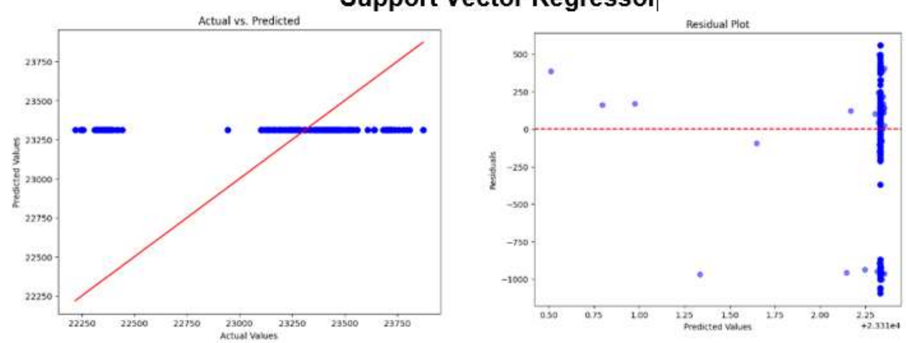
Gradient Boosting Regressor



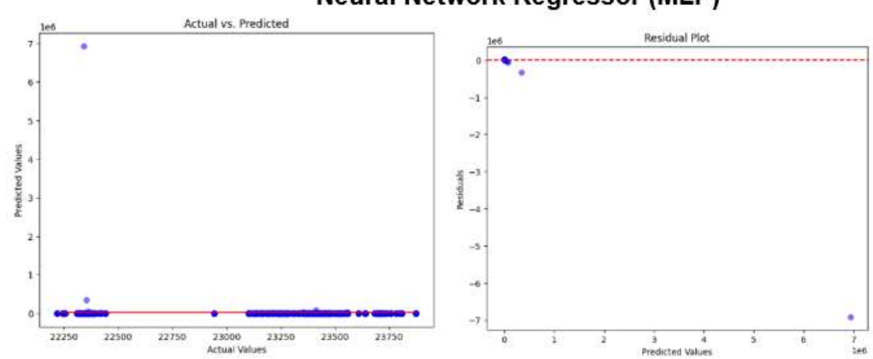
Decision Tree Regressor



Support Vector Regressor



Neural Network Regressor (MLP)



COMPARISON TABLES

Regression Model	Mean Squared Error	R-squared	Mean Absolute Error	Root Mean Squared Error	Predicted Close Price
Multiple Linear Regression	263642.15	-0.0002768	434.52	513.46	23131.47
Decision Tree Regressor	330171.61	-0.2526942	372.79	574.60	23457.06
Random Forest Regressor	229402.74	0.1296298	332.16	478.96	22660.31
Support Vector Regressor	297238.96	-0.1277454	399.94	545.19	23312.32
Gradient Boosting Regressor	218755.55	0.1700260	378.03	467.71	23144.45
Neural Network Regressor	312875279304.78	-1187069.7590	51811.64	559352.55	2.23

Note as per reference [19]: The best possible score is 1 which is obtained when the predicted values are the same as the actual values. R2 score of baseline model is 0 and during the worse cases, R2 score can even be negative.

The Random Forest and Gradient Boosting regressors demonstrate better predictive capabilities compared to Multiple Linear Regression and Decision Tree Regressor, with lower mean squared errors and positive R-squared values.

Toronto Metropolitan University

CROSS VALIDATION

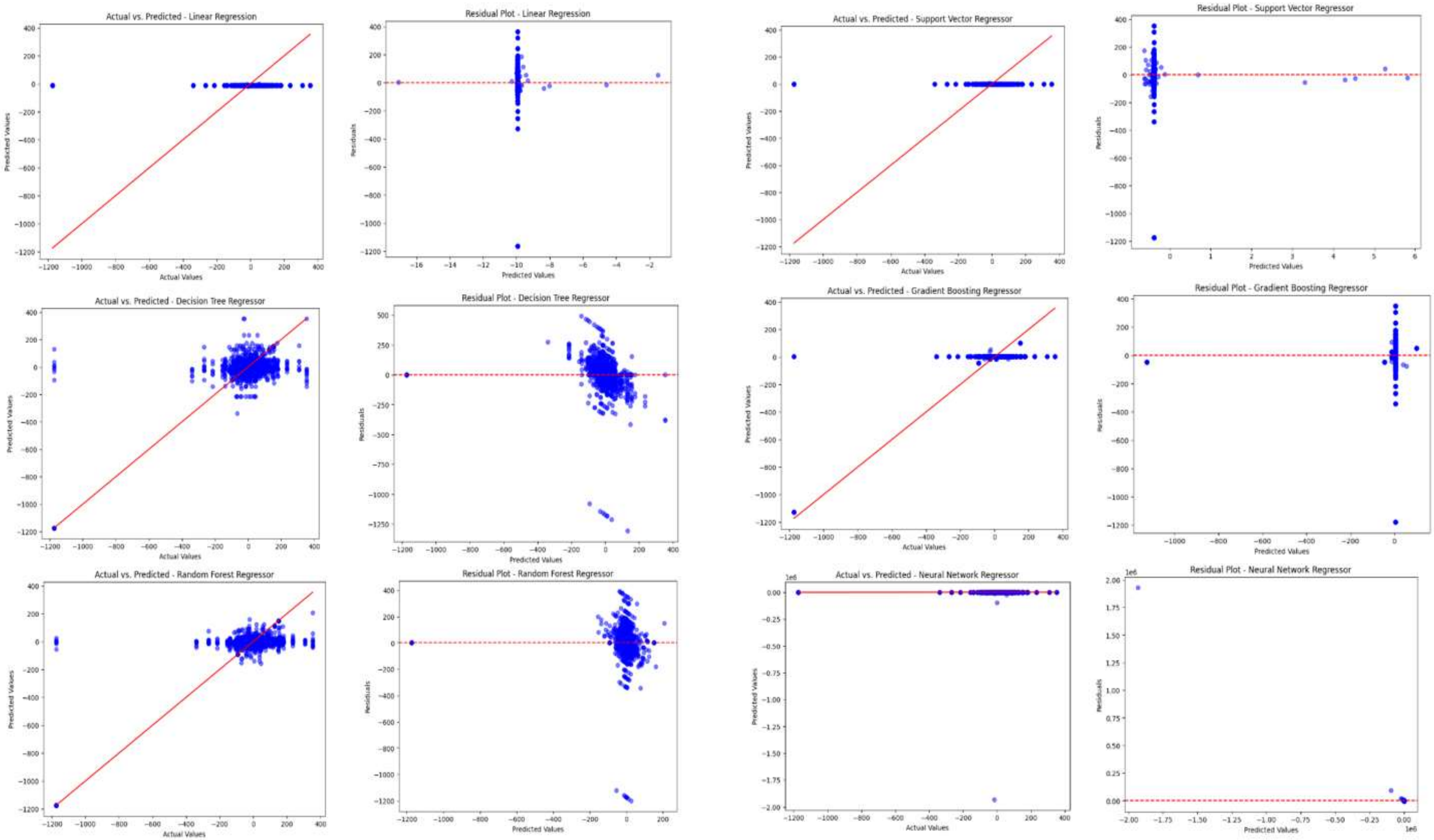
Regression Model	Mean MSE	Std MSE
Linear Regression	369821.09	306106.77
Decision Tree Regressor	465009.63	335780.03
Random Forest Regressor	358518.17	372697.44
Support Vector Regressor	347812.17	401289.32
Gradient Boosting Regressor	348887.68	310402.67
Neural Network Regressor	3.12E+1	4.70E+13

The Support Vector Regressor stands out as the best performer among the presented models, showing the lowest Mean MSE and displaying promise for accurate predictions.

HYPERTUNNING PARAMETERS

```
Model: Decision Tree Regressor
Best Model: DecisionTreeRegressor(max_depth=7, min_samples_leaf=3, min_samples_split=5)
Best Parameters: {'max_depth': 7, 'min_samples_leaf': 3, 'min_samples_split': 5}
Mean Squared Error: 249886.37207620614
R-squared: 0.05191357385549278
Mean Absolute Error: 413.05794247618445
Root Mean Squared Error: 499.8863591619661

Model: Random Forest Regressor
Best Model: RandomForestRegressor(max_depth=7, min_samples_split=5, n_estimators=50)
Best Parameters: {'max_depth': 7, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 50}
Mean Squared Error: 249886.91762665412
R-squared: 0.05191150399882072
Mean Absolute Error: 412.8730317998603
Root Mean Squared Error: 499.886904836138

Model: Support Vector Regressor
Best Model: SVR(C=0.1, epsilon=0.001)
Best Parameters: {'C': 0.1, 'epsilon': 0.001}
Mean Squared Error: 296035.0250691949
R-squared: -0.12317765310490625
Mean Absolute Error: 399.4075468116839
Root Mean Squared Error: 544.091008076034

Model: Gradient Boosting Regressor
Best Model: GradientBoostingRegressor(n_estimators=50)
Best Parameters: {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 50}
Mean Squared Error: 249884.83969784752
R-squared: 0.05191938780648764
Mean Absolute Error: 413.1128964010035
Root Mean Squared Error: 499.88482643289694

Model: Neural Network Regressor

 Best Model: MLPRegressor(alpha=0.01, hidden_layer_sizes=(200, 100))
 Best Parameters: {'alpha': 0.01, 'hidden_layer_sizes': (200, 100)}
 Mean Squared Error: 249892.94011968057
 R-squared: 0.05188865423777744
 Mean Absolute Error: 412.56464838672906
 Root Mean Squared Error: 499.89292865540773

 Model: Linear Regression
 Best Model: LinearRegression()
 Best Parameters: {}
 Mean Squared Error: 249886.37207620472
 R-squared: 0.05191357385549822
 Mean Absolute Error: 413.0579424764286
 Root Mean Squared Error: 499.88635916196466
```

REGRESSION MODELS FOR PRICE CHANGE (DELTA) PREDICTIONS



COMPARISON TABLE

Regression Model	Mean Squared Error	R-squared	Mean Absolute Error	Root Mean Squared Error	Predicted Change Prediction (Delta)
Linear Regression	2.72E+04	0.000017	77.29	164.91	-9.94
Decision Tree Regressor	1.61E+04	0.408726	71.75	126.80	16
Random Forest Regressor	1.42E+04	0.478842	64.99	119.05	8.42
Support Vector Regressor	2.73E+04	-0.003254	76.81	165.17	-0.39
Gradient Boosting Regressor	1.40E+04	0.484404	64.94	118.41	3.66
Neural Network Regressor	2.37E+09	-87042.17517	1415.69	48653.96	-0.25

The **Random Forest Regressor and Gradient Boosting Regressor** demonstrated relatively better performance in predicting the change in price (delta) compared to other models.

CROSS VALIDATION

Regression Model	Mean MSE	Std MSE
Multiple Linear Regression	26602.32	34442.20
Decision Tree Regressor	29497.64	34444.12
Random Forest Regressor	26945.52	34605.31
Support Vector Regressor	26258.12	34451.38
Gradient Boosting Regressor	26563.17	34572.29
Neural Network Regressor	1.82E+14	3.44E+14

The **Multiple Linear Regression, Support Vector Regressor, and Gradient Boosting Regressor** models showed relatively lower mean squared error (MSE) values, indicating better performance in predicting Bitcoin price changes. However, the Neural Network Regressor had significantly higher MSE values and performed poorly in this task. Further optimization and experimentation may be needed to improve model performance.

HYPERTUNNING PARAMETERS

```
Model: Decision Tree Regressor
Best Model: DecisionTreeRegressor(max_depth=5, min_samples_split=5)
Best Parameters: {'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 5}
Mean Squared Error: 249886.37207620614
R-squared: 0.05191357385549278
Mean Absolute Error: 413.05794247618445
Root Mean Squared Error: 499.8863591619661

Model: Random Forest Regressor
Best Model: RandomForestRegressor(max_depth=7, min_samples_leaf=3, n_estimators=50)
Best Parameters: {'max_depth': 7, 'min_samples_leaf': 3, 'min_samples_split': 2, 'n_estimators': 50}
Mean Squared Error: 249887.65206258508
R-squared: 0.051908717497375245
Mean Absolute Error: 412.8897577630469
Root Mean Squared Error: 499.88763943768913

Model: Support Vector Regressor
Best Model: SVR(C=0.1, epsilon=0.001)
Best Parameters: {'C': 0.1, 'epsilon': 0.001}
Mean Squared Error: 296035.0250691949
R-squared: -0.12317765310490625
Mean Absolute Error: 399.4075468116839
Root Mean Squared Error: 544.091008076034

Model: Gradient Boosting Regressor
Best Model: GradientBoostingRegressor(n_estimators=50)
Best Parameters: {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 50}
Mean Squared Error: 249884.83969784752
R-squared: 0.05191938780648764
Mean Absolute Error: 413.1128964010036
Root Mean Squared Error: 499.88482643289694

Model: Neural Network Regressor

Best Model: MLPRegressor(alpha=0.1, hidden_layer_sizes=(200, 100))
Best Parameters: {'alpha': 0.1, 'hidden_layer_sizes': (200, 100)}
Mean Squared Error: 249884.8371738839
R-squared: 0.051919397382582666
Mean Absolute Error: 413.3885683452907
Root Mean Squared Error: 499.8848239083518

Model: Linear Regression
Best Model: LinearRegression()
Best Parameters: {}
Mean Squared Error: 249886.37207620472
R-squared: 0.05191357385549822
Mean Absolute Error: 413.0579424764286
Root Mean Squared Error: 499.88635916196466
```

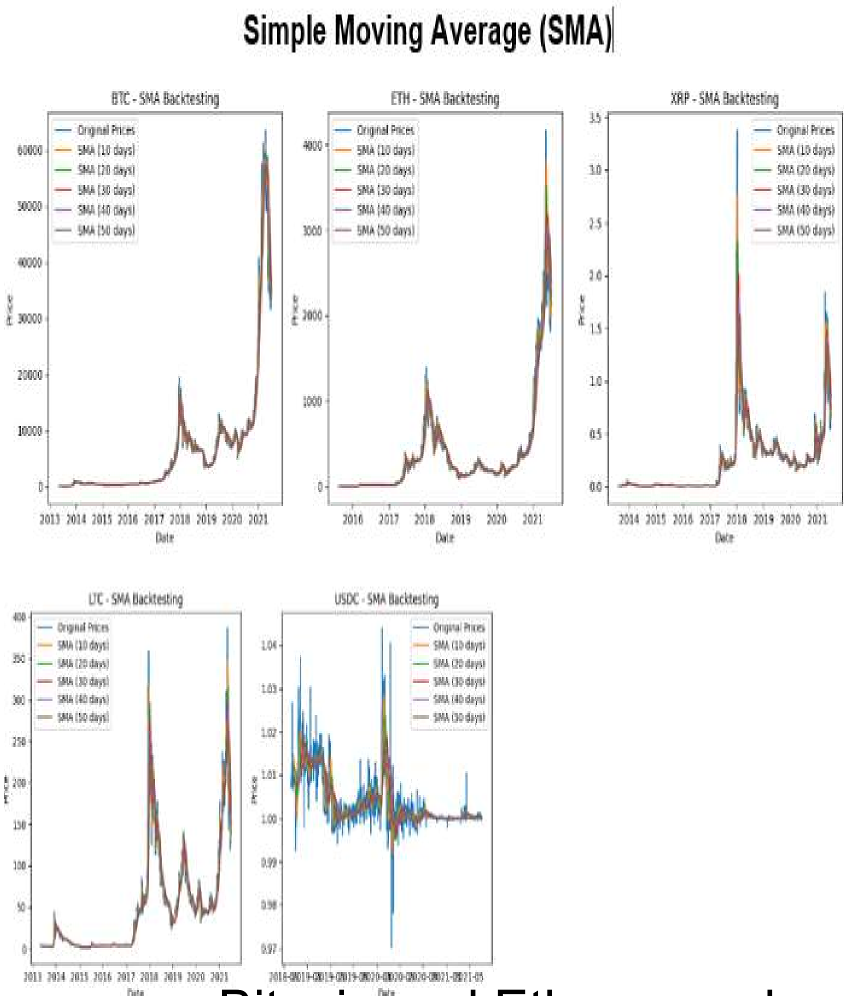
After hyperparameter tuning, slight improvements were observed in model performance. However, all models showed limited capability to explain the variance in Bitcoin price changes, as indicated by the low R-squared values. Further refinement and incorporating historical price data may enhance predictive power for future price prediction.

Toronto Metropolitan University

TIME SERIES MODELS FOR FIVE CRYTOCURRENCIES PRICE PREDICTION

- Simple Moving Average (SMA)
- Autoregressive Integrated Moving Average (ARIMA)
- Prophet
- Deep Learning Techniques - LSTM model

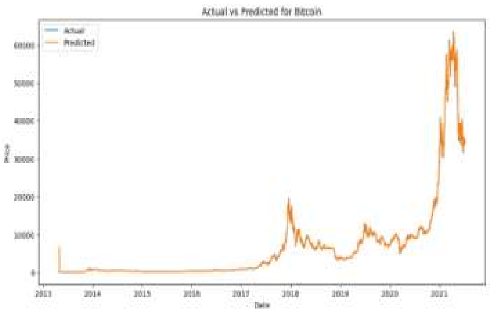
Simple Moving Average (SMA)

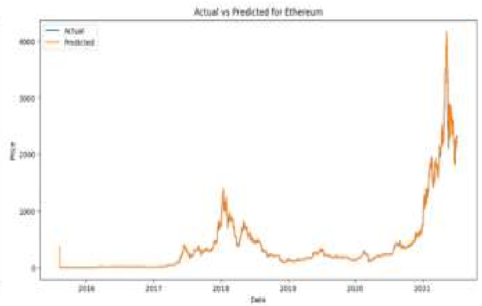


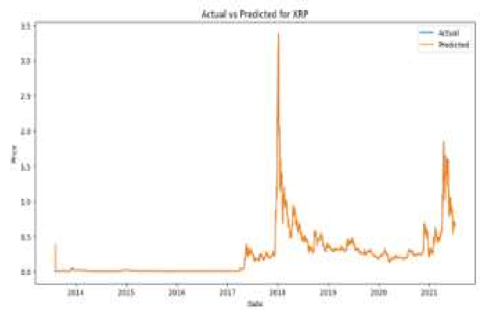
Coin	MSE (30 day)	MAE (30 day)	RMSE (30 day)	MAPE (30 day)	MSE (60 day)	MAE (60 day)	RMSE (60 day)	MAPE (60 day)
BTC	3.53E+	739.6	1880.01	9.79	8.71E+	1216.11	2951.51	14.8
ETH	20026.5	60.8	141.515	15.46	40000	89.75	200	22.34
XRP	0.022	0.048	0.15	14.88	0.037	0.068	0.19	21.
LTC	359.3	7.9	18.95	12.95	650.76	11.62	25.51	19.3
USDC	2.49E-	0.002	0.0049	0.26	2.64E-	0.0028	0.0051	0.2

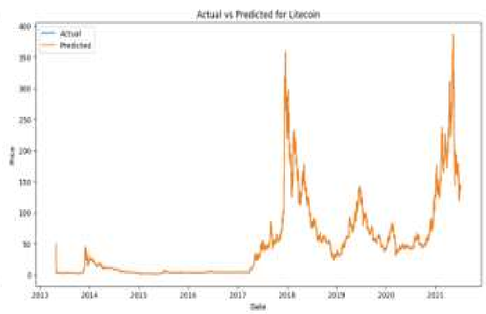
Bitcoin and Ethereum show higher errors, **XRP, Litecoin, and USD Coin** exhibit relatively better predictive performance with lower errors and higher accuracy.

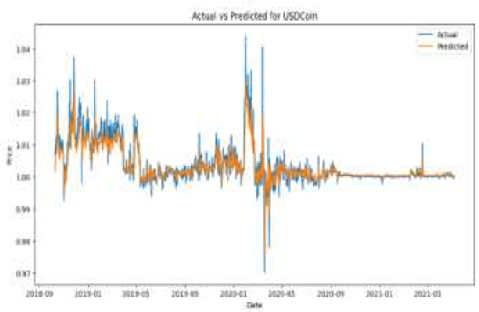
Autoregressive Integrated Moving Average (ARIMA)









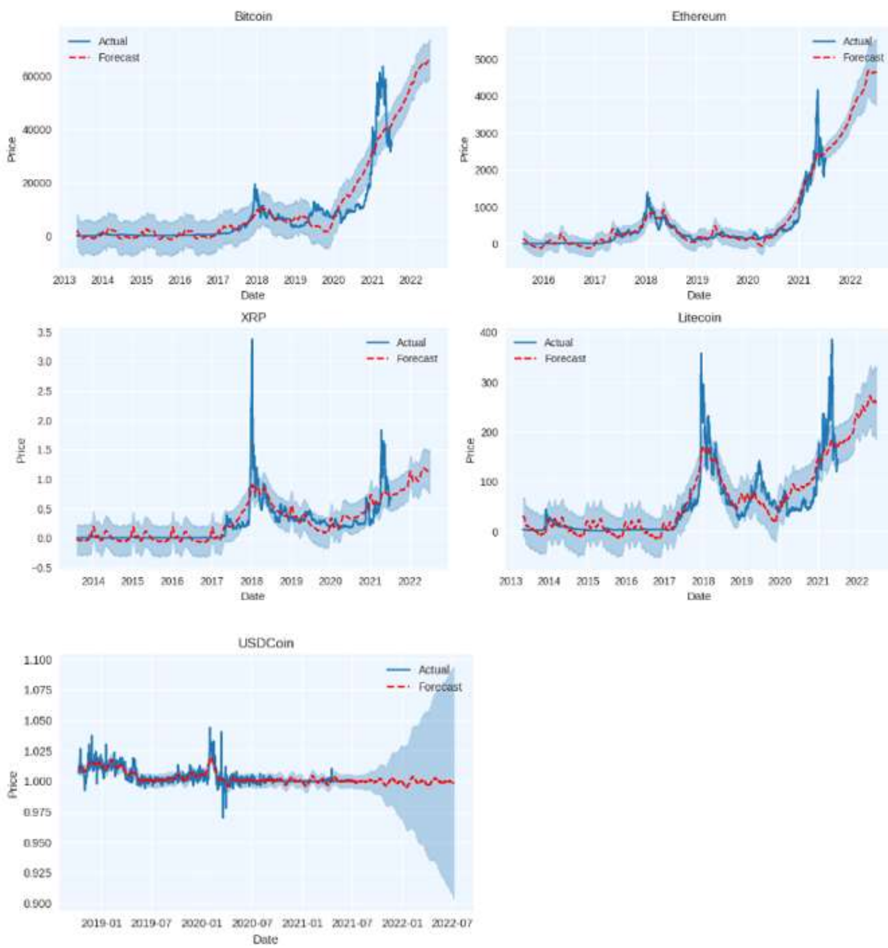


Coin	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	Mean Absolute Percentage Error
Bitcoin	209.93	357706.50	598.08	4.47
Ethereum	17.21	2332.79	48.29	32.38
XRP	0.014	0.0018	0.043	24.18
Litecoin	2.33458	37.48	6.12	6.05
USDCoin	0.0024	1.76E-05	0.0041	0.23

The ARIMA model performs differently for various cryptocurrencies, with **Ethereum and USDCoin** showing relatively accurate predictions.

Toronto Metropolitan University

Prophet

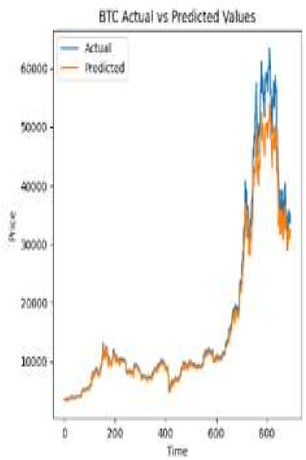


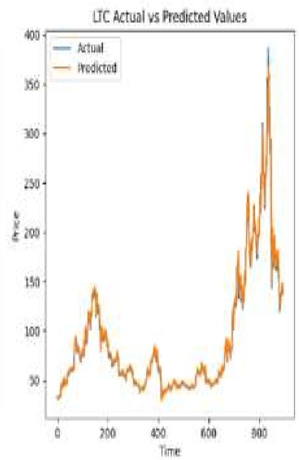
Coin	Mean Squared Error	Root Mean Squared Error	Mean Absolute Error	R2 Score
Bitcoin	2.28E+07	4770.50	2895.87	0.82
Ethereum	3.17E+04	177.91	108.10	0.91
XRP	3.87E-02	0.19	0.11	0.66
Litecoin	8.69E+02	29.47	19.794	0.78
USDCoin	1.75E-05	0.0041	0.0026	0.62

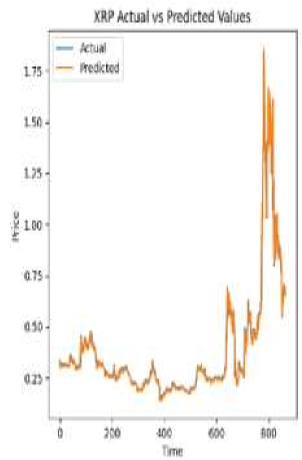
The Prophet model shows potential in predicting cryptocurrency prices, with **Ethereum** exhibiting the highest accuracy.

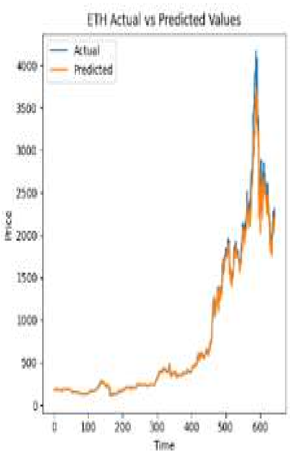
Deep Learning Techniques

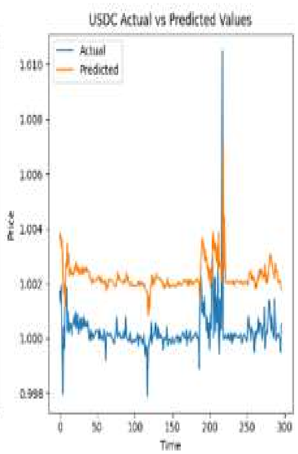
LSTM model without any dropout layers







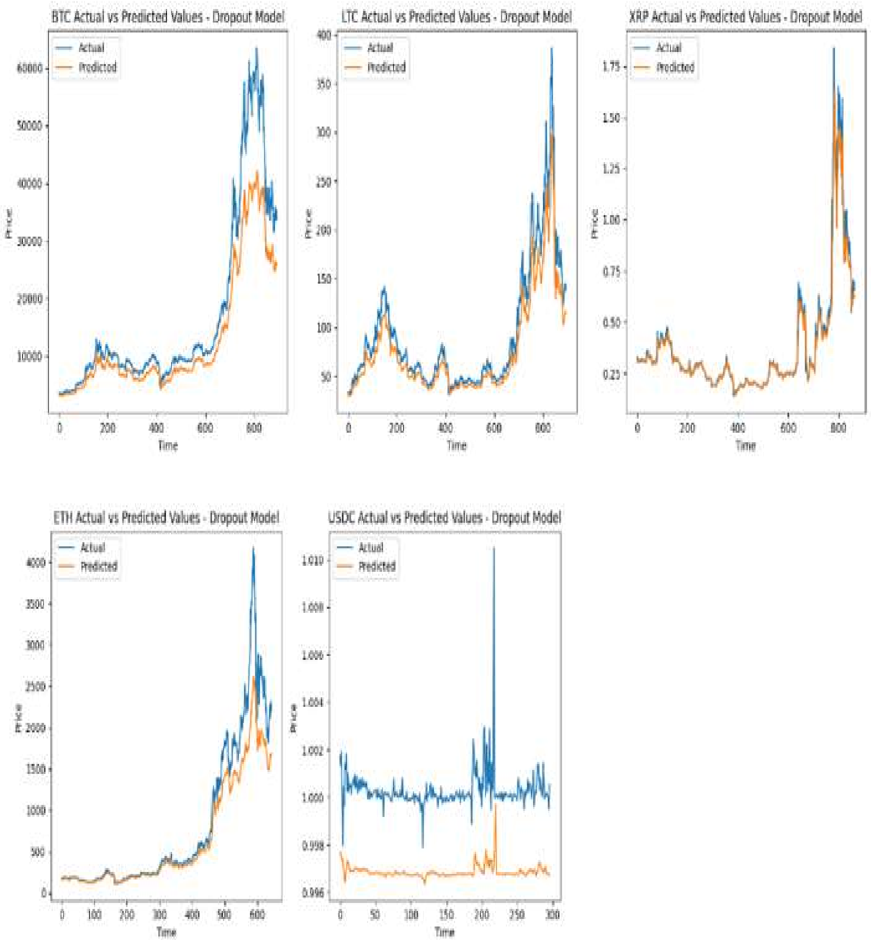




Coin	MSE	RMSE	MAE	MAPE	Corr	R2
BTC	7.75E+06	2784.63	1422.4	5.64	0.99	0.96
LTC	6.54E+01	8.08	4.28	4.27	0.99	0.98
XRP	2.10E-03	0.045	0.019	4.11	0.98	0.97
ETH	1.19E+04	109.02	46.03	4.01	0.99	0.98
USDC	4.60E-06	0.0021	0.002	0.20	0.30	-5.80

LSTM models **perform well for most cryptocurrencies**, showing strong positive linear relationships and reasonable accuracy. However, the model for USDCoin falls short, exhibiting weaker performance and lower accuracy compared to other cryptocurrencies.

Deep Learning Techniques
LSTM model with dropout layers



Coin	MSE	RMSE	MAE	MAPE	Corr	R2
BTC	44407580	6663.9	4190.13	20.44	0.99	0.82
LTC	451.03	21.23	15.17	14.65	0.99	0.88
XRP	0.0036	0.0600	0.025	4.57	0.98	0.95
ETH	107471.5	327.82	169.38	13.17	0.99	0.86
USDC	1.23E-05	0.0035	0.0034	0.341	0.22	-17.25

The LSTM model demonstrated **promising accuracy for most cryptocurrencies**, warranting further investigation, except for USDC, which performed poorly.

SELF REFLECTION

- I gained valuable insights into the significance of sentiment analysis in the cryptocurrency market and its potential impact on price movements.
- Utilizing various machine learning algorithms for sentiment prediction and regression models for price forecasts, I gained valuable insights into their strengths and limitations.
- I plan to continue exploring time series forecasting and applying my knowledge to real-world applications to contribute to the development of more accurate and efficient results.
- Maximize timely deliverables for real-time schemas.
- The insights gained and the challenges encountered will serve as a foundation for refining future projects.

Toronto Metropolitan University

LIMITATIONS

CHALLENGES WITH SENTIMENT ANALYSIS CALCULATION

Current libraries do not consider sarcasm

OVERRELIANCE ON TEXTUAL DATA

Sentiment can also be expressed through other modalities like images, videos, or audio

LIMITED AVAILABILITY OF LABELED DATA AND DATA BIAS

Presence of spam accounts, false accounts, and bots

RAPIDLY EVOLVING SENTIMENT AND CONTEXTUAL UNDERSTANDING

Market manipulation, external events, or speculative trading behavior, can contribute to this discrepancy

FUTURE WORK

- Collect tweets from **high-profile accounts** with market influence.
- Explore **removing false Twitter accounts** to improve sentiment analysis.

- Increase data volume and **expand trade date range** for better model accuracy.
- **Explore alternative machine learning** techniques for **sentiment score** computation. (as current libraries do not take into account sarcasm)

- Consider other factors **beyond investor and decision maker** sentiments for broader research in the cryptocurrency market.
- Develop a robust **big data platform for real-time analysis** and prediction.

REFRENCES

- Al Amrani, Y., Lazaar, M., & El Kadiri, K. E. (2022). Sentiment Analysis and Cryptocurrency Price Forecasting. International Journal of Advanced Computer Science and Applications, 13(10), 889-898. https://thesai.org/Downloads/Volume13No10/Paper_105_Cryptocurrency_Price_Prediction_using_Forecasting_and_Sentiment_Analysis.pdf
- E. Edgari, J. Thiojaya and N. N. Qomariyah, "The Impact of Twitter Sentiment Analysis on Bitcoin Price during COVID-19 with XGBoost," 2022 5th International Conference on Computing and Informatics (ICCI), New Cairo, Cairo, Egypt, 2022, pp. 337-342, doi: 10.1109/ICCI54321.2022.9756123. https://ieeexplore-ieee-org.ezproxy.lib.torontomu.ca/stamp/stamp.jsp?tp=&arnumber=9756123
- A. Ibrahim, "Forecasting the Early Market Movement in Bitcoin Using Twitter's Sentiment Analysis: An Ensemble-based Prediction Model," 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Toronto, ON, Canada, 2021, pp. 1-5, doi: 10.1109/IEMTRONICS52119.2021.9422647. https://ieeexplore-ieee-org.ezproxy.lib.torontomu.ca/document/9422647
- A. M. Balfagih and V. Keselj, "Evaluating Sentiment C1assifiers for Bitcoin Tweets in Price Prediction Task," 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 5499-5506, doi: 10.1109/BigData47090.2019.9006140. https://ieeexplore-ieee-org.ezproxy.lib.torontomu.ca/document/9006140
- Jacques Fleischer, Gregor von Laszewski, Carlos Theran, Yohn Jairo Parra Bautista " Time Series Analysis of Blockchain-Based Cryptocurrency Price Changes ", 2022 https://paperswithcode.com/paper/time-series-analysis-of-blockchain-based#code
- R. N, S. R. R, V. S. R and K. P. D, "Crypto-Currency Price Prediction using Machine Learning," 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2022, pp. 1455-1458, doi: 10.1109/ICOEI53556.2022.9776665. https://ieeexplore-ieee-org.ezproxy.lib.torontomu.ca/document/9776665
- C. Saravanakumar, S. K. Lakshmi, M. Prakash, M. Vijayakumar and C. Arun, "An Efficient Ensemble Model for Forecasting Time Series Analysis of Crypto Currency Data Using Machine Learning," 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 2021, pp. 1-5, doi: 10.1109/ICSES52305.2021.9633939. https://ieeexplore-ieee-org.ezproxy.lib.torontomu.ca/document/9633939
- R. Nikita, S. J. Subhashini, Y. B. S, V. K. Vavle and D. A, "A Review on Digital Coin Investing Predictor," 2022 International Conference for Advancement in Technology (ICONAT), Goa, India, 2022, pp. 1-6, doi: 10.1109/ICONAT53423.2022.9726092. https://ieeexplore-ieee-org.ezproxy.lib.torontomu.ca/document/9726092
- Pan, Linxi. (2023). Cryptocurrency Price Prediction Based on ARIMA, Random Forest and LSTM Algorithm. BCP Business & Management. 38. 3396-3404. 10.54691/bcpbm.v38i.4313. https://www.researchgate.net/publication/369430933_Cryptocurrency_Price_Prediction_Based_on_ARIMA_Random_Forest_and_LSTM_Algorithm
- Yulin Liu, Luyao Zhang "Cryptocurrency Valuation: An Explainable AI Approach", 2022 https://paperswithcode.com/paper/cryptocurrency-valuation-an-explainable-ai
- Zhengyao Jiang, Jinjun Liang "Cryptocurrency Portfolio Management with Deep Reinforcement Learning", 2016 https://paperswithcode.com/paper/cryptocurrency-portfolio-management-with-deep
- Quentin Le Baron (2018) "How to make profits in cryptocurrency trading with machine learning" https://medium.com/smileinnovation/how-to-make-profits-in-cryptocurrency-trading-with-machine-learning-edb7ea33cee4
- Raghav Agrawal (2023) " Complete Guide to Ethereum Blockchain with Python" https://www.analyticsvidhya.com/blog/2023/01/complete-guide-to-web3-0-ethereum-blockchain-using-python/

Thank You

Toronto Metropolitan University