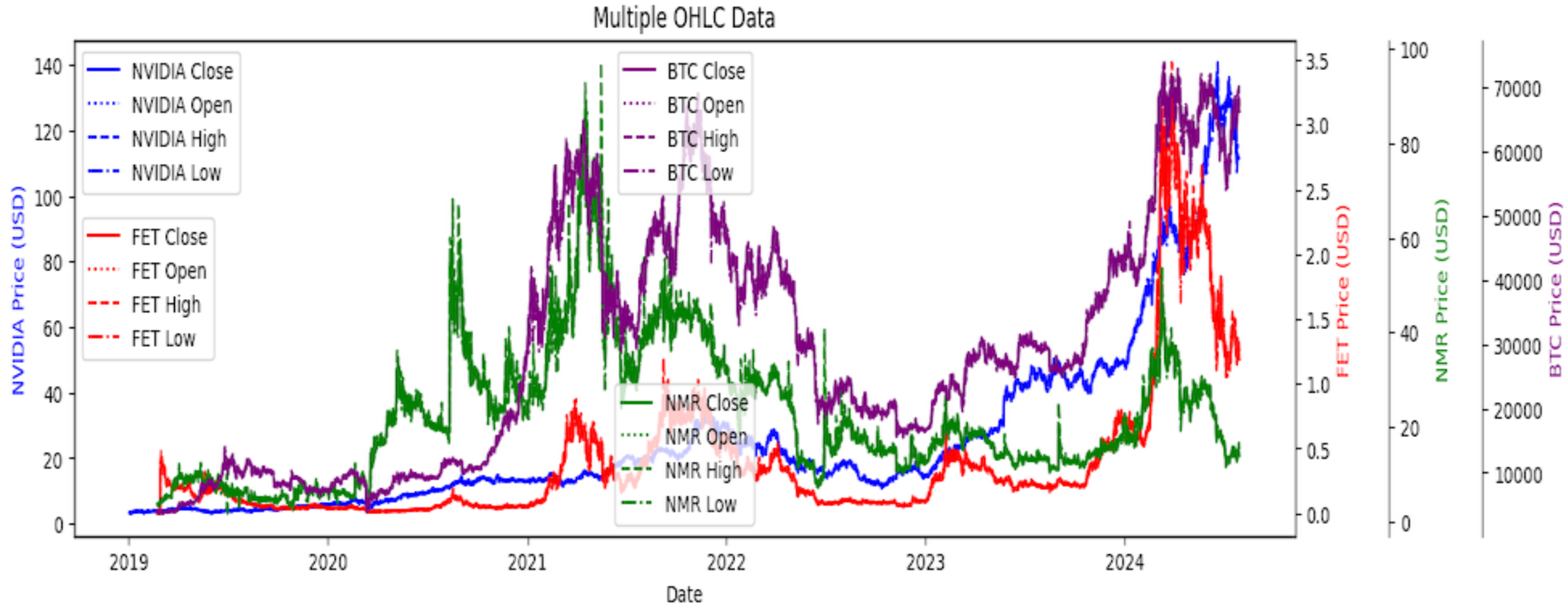
MODELING PORTFOLIO RISK MGT. FOR CRYPTO ASSETS



Multi-charts plotted from this project showing strong correlation between NMR, FET, BTC, and NVIDIA

Courtesy: Okpo.

OKPO E.

29ND JULY 2024

CONTENT

ABSTRACT

This analysis centers on modeling and predicting the price change or returns of cryptocurrencies using various machine learning techniques coupled with in-depth data gathering and processing. The input parameters into the model which lends the process the name for which it's usually called are termed factors. Employing a combination of the traditional Capital Assets Pricing Model (CAPM) and multi factors like Fama-French (adapted for cryptocurrencies), ee analyzed a total of 13 factors created from technical indicators, macroeconomic factor, blockchain dynamics, to social sentiments to understand what drives the prices of cryptocurrency assets and determine the right window to invest. Employing traditional market variables and social sentiment data, we develop Ridge and Lasso regression models to assess feature importance and predict future price movements. The findings highlight the significance of excess market returns and social sentiment, providing valuable insights for traders and portfolio managers. The analysis reveals that traditional market variables, combined with social sentiment, provide a robust framework for predicting cryptocurrency prices. The Ridge model's ability to balance complexity and interpretability makes it a valuable tool for portfolio management. However, further research is needed to explore the effects of on-chain metrics and advanced machine learning models. We conclude with suggestions for future research on advanced modeling techniques and broader datasets.

INTRODUCTION

Cryptocurrencies have been a huge source of interest attracting substantial attention from investors, regulators, the media, and the general public all over the world, since the introduction of Bitcoin by Nakamoto (2008). Initially greeted with skepticism, we've seen recent acceptance into mainstream financial institutions like JP Morgan and Black Rock. A number of countries are also shifting their stance from an initial hostile perspective to acceptance through regulation and licensing.

Serving as a ledger for digital transactions across borders, cryptocurrencies facilitate direct online payments between parties without the need for intermediaries such as financial institutions. Within the cryptocurrency ecosystem, several studies have been conducted to assess the volatility of the asset and determine when to maximize profit and minimize loss.

Noteworthy to mention is the shift in the dynamics from a more isolated asset system to that which is being influenced by the traditional market factors. For example, recently, we witness a price crash in the cryptocurrency ecosystem as the Japanese Central Bank raise their interest rate from a decade-long low value, leading to panic sales and plummeting of assets around the world. We also, in this year saw NVIDIA rose to an incredible valuation in market cap, leading to price surge in certain assets like OCEAN and FET.AI, collectively termed AI tokens. This has inspired ground-breaking mergers like the Artificial Super Intelligence (ASI).

These few incidence coupled with many others has challenged the original concept of Bitcoin and others as an hedge against shocks from mainstream traditional market. Nevertheless, we've bitcoin and others have been resilient in their growth.

INTRODUCTION

The report is organized into clear sections, with each method and result presented logically. Visualizations are labeled and integrated into the text to enhance understanding. Citations are provided for all data sources, ensuring transparency and credibility. Because of time, more charts that would have enhanced the report are included in the appendix.

1. DATA COLLECTION AND QUALITY (SOURCE DIVERSITY)

1. **Data was sourced from diverse sources which are mostly data *aggregators* that pulled data from hundreds of exchanges across the globe.**
2. **The sources including CryptoCompare, CCXT, Yahoo Finance API, and CoinGecko.**
3. **On CoinGecko, we were able to get the list of top cryptocurrencies by Market Cap and Volume but there was restrictions on historical data.**
4. **CryptoCompare, Yahoo FInance via *yfinance* library and CCXT were incredibly useful in getting blockchain, social sentiment and daily historical/volume (OHLCV) data.**
5. **We got a total of *2605* stable coin pairs from the gate exchange via CCXT and *2200* pairs using yfinance.**

1. DATA COLLECTION AND QUALITY (DATA QUALITY)

1. To ensure that the data collected was of good quality, we made sure it came from reputable sources like CryptoCompare, Yahoo Finance
2. Traditional financial data, such as stock indices (S&P 500, VIX, etc.) and commodity prices (Gold), were sourced from established platforms like Yahoo Finance.
3. We also verified the accuracy of the data through cross-validation with multiple sources where possible. For instance, price data from different exchanges was compared to ensure consistency.
4. Moreover, the dataset was complete, covering the set time frames (2019 - 2024) and including all necessary variables, from market returns to sentiment scores, ensuring robust analysis.
5. The time range of 2019 to 2024 was chosen carefully to ensure data completion as some tokens data were missing prior to 2019.

1. DATA COLLECTION AND QUALITY (DATA CLEANING)

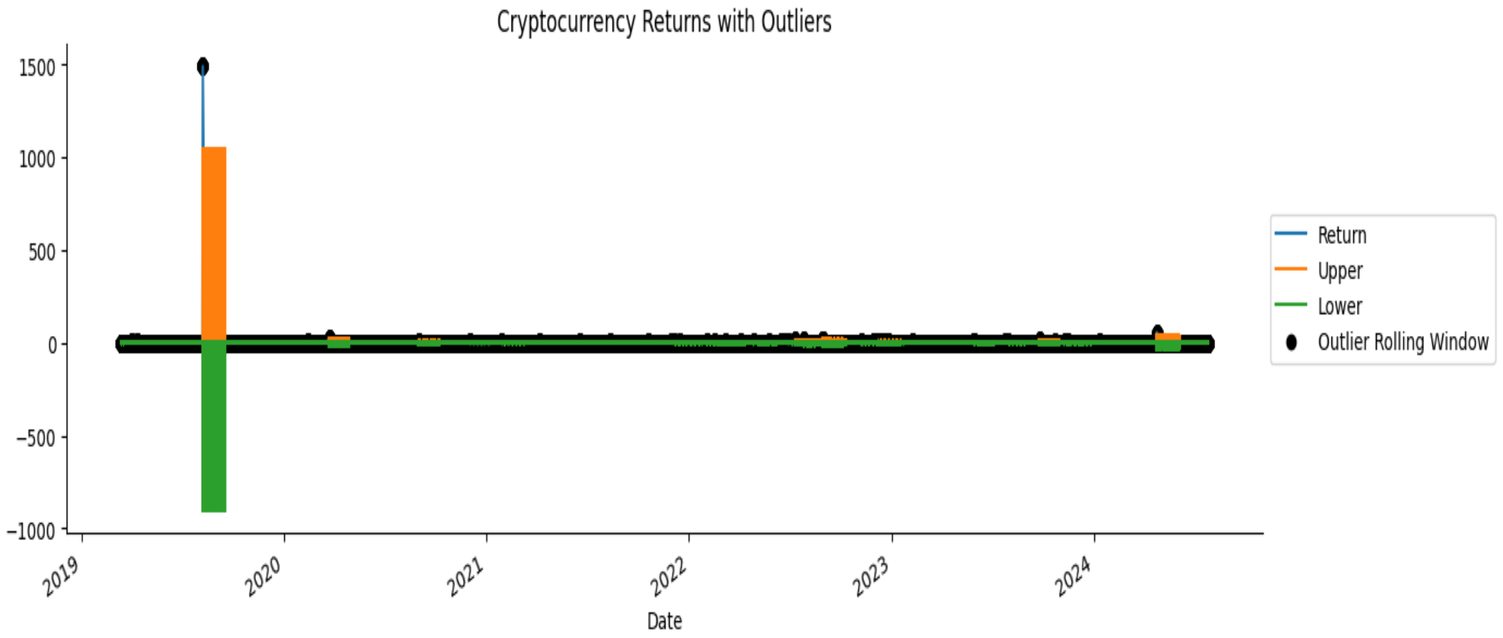
1. Understanding the importance of data cleaning for reliable and useable dataset in data analysis, we undertook several steps like data imputation, outlier detection, and stationarity tests.
2. Handling Missing Values: We used forward and backward filling techniques to address missing data points, ensuring continuity in the time series data without introducing bias
3. Outlier Detection and Removal: Extreme values were identified using statistical methods (e.g., Z-scores, IQR) and were either removed or treated appropriately to prevent skewing the results.
4. Stationarity Tests: We conducted stationarity tests (e.g., ADF test) and applied transformations, such as differencing, to ensure that the time series data was suitable for modeling.

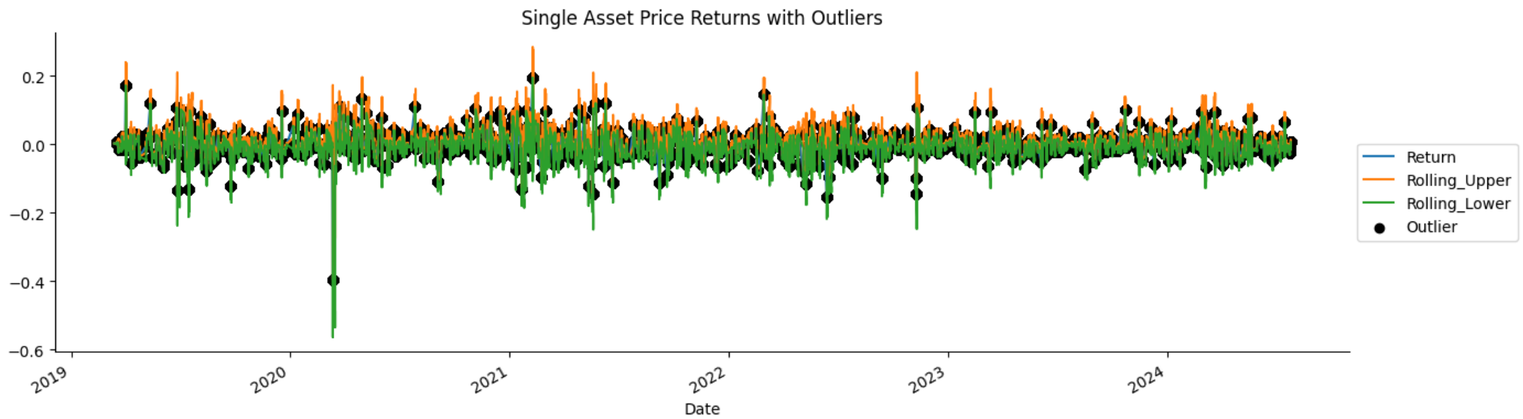
1. DATA COLLECTION AND QUALITY (DATA CLEANING)

5. Understandin Normalization and Scaling: To prepare the data for regression and machine learning models, features were normalized and scaled, ensuring that all variables were on comparable scales and eliminating any potential issues with model convergence

6. Feature Engineering: We also derived new variables, such as momentum, volatility, and excess returns, from the raw data, enhancing the richness of the dataset and providing more informative inputs for our models.

1. DATA COLLECTION AND QUALITY (DATA CLEANING)

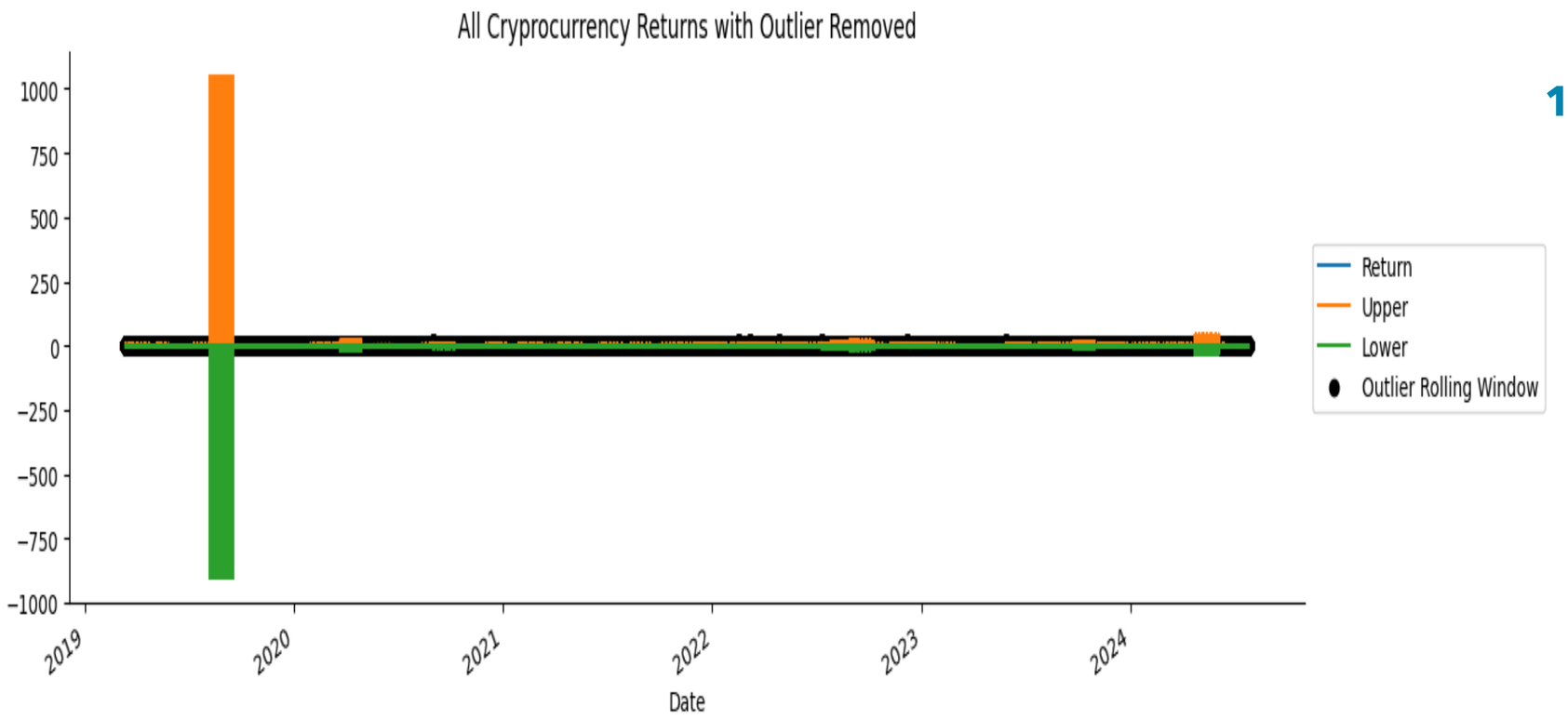




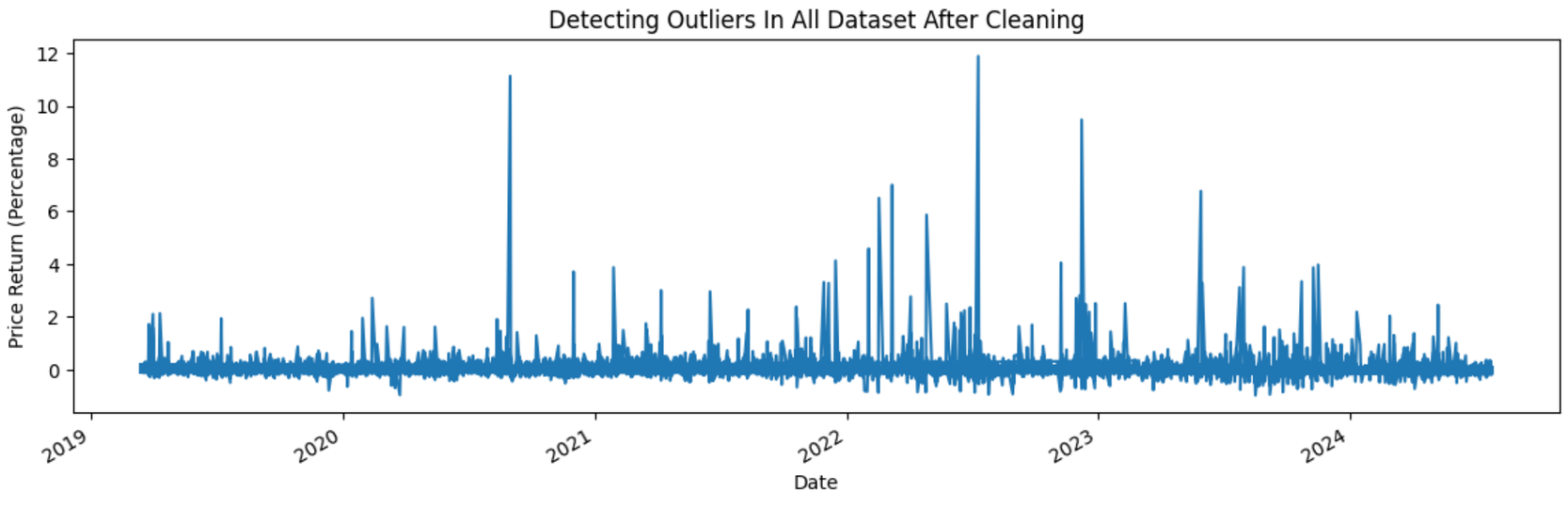
1. **This chart on the left illustrates all the initial detection of outliers at prior to cleaning**
2. **Using the rolling statistics technique, we discovered outliers across the tokens**
3. **The chart on bottom left is a single asset outlier discover process, for BTC/USDT**

5. **Understandin Normalization and Scaling: To prepare the data for regression and**

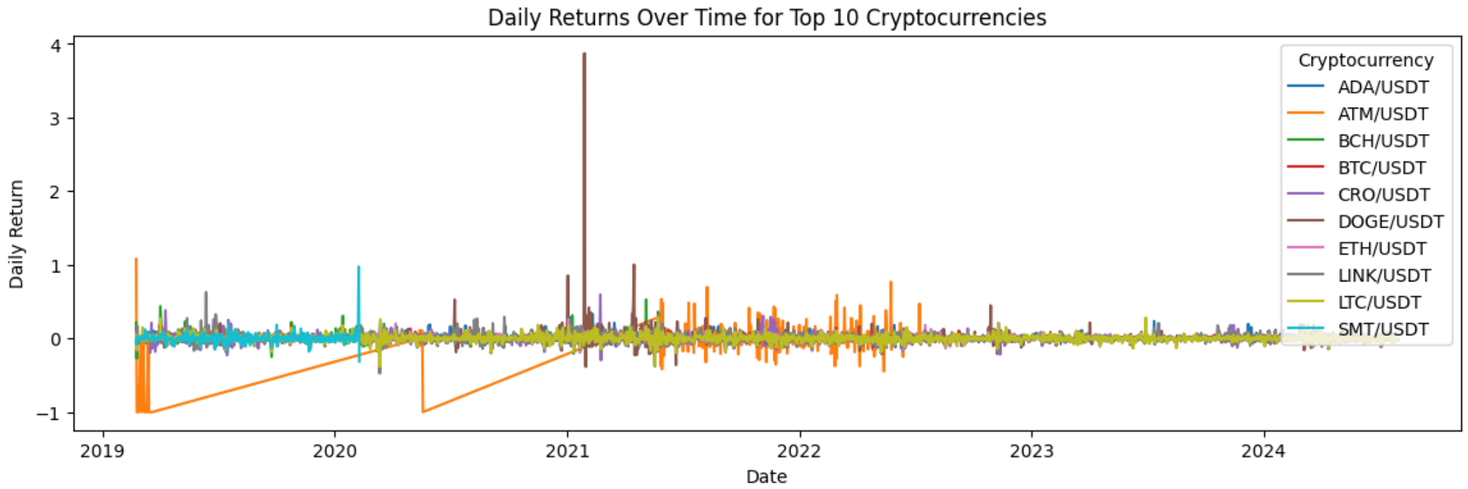
1. DATA COLLECTION AND QUALITY (DATA CLEANING)



1. **The charts on the left illustrate all outliers removed for techniques like regression models to function well**



1. DATA COLLECTION AND QUALITY (DATA CLEANING)

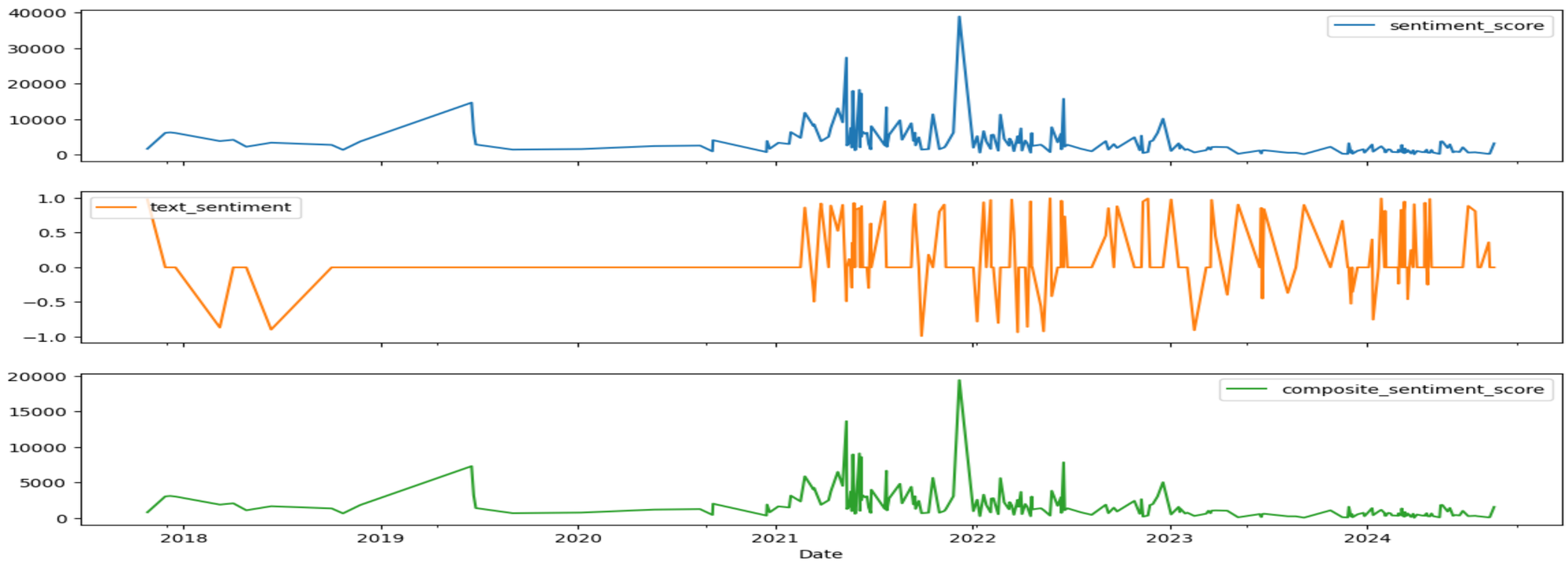


1. **The charts on the left illustrate all outliers across the Price Return of 10 top cryptocurrencies.**

1. DATA COLLECTION AND QUALITY (INNOVATIVE APPROACH)

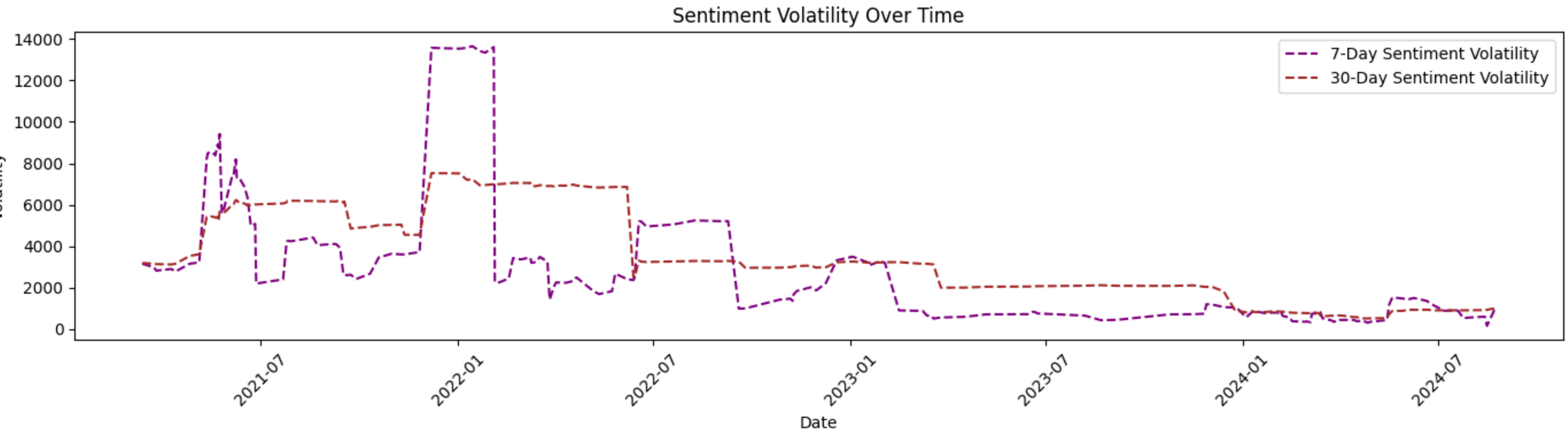
1. Use of Novel Methods or Unique Data Sources That Add Value to the Dataset: One of the unique aspects of this project was the integration of social sentiment data into the factor modeling process
2. While traditional financial data is widely used, incorporating real-time social sentiment data, sourced from platforms like Reddit and Twitter through CryptoCompare's API, provided a novel perspective on market movements
3. This approach allowed us to capture the influence of social trends and investor sentiment on cryptocurrency prices, which is particularly relevant in the highly speculative nature of crypto markets.
4. Additionally, by merging traditional financial data with unconventional metrics like social sentiment, we were able to construct a more holistic model that better reflects the multifaceted nature of cryptocurrency price movements. This innovative approach added significant value to our analysis and provided insights that might not have been uncovered using traditional methods alone. The next few slides focus on project charts to explain further.

1. DATA COLLECTION AND QUALITY (INNOVATIVE APPROACH)



1. **<u>Chart Explanation:</u> The chart above, is a plot of the sentiment score of Cryptocurrency-related texts from Reddit and Twitter. We notice spikes in certain periods like 2021 which may have reflected sentiments around bull runs at that time.**

1. DATA COLLECTION AND QUALITY (INNOVATIVE APPROACH)



1. **<u>Chart Explanation:</u> The chart above, is a plot of the <u>sentiment volatility</u> of Cryptocurrency-related texts from Reddit and Twitter. The purpose was to have a clearer picture of the trends which was achieve here**

2. MODEL DEVELOPMENT - CAPM

`CAPM Model Summary(ccompare_gate_merged_df)`

Metric	Value	Interpretation
Dependent Variable	Excess_Return	The excess return of the cryptocurrency above the risk-free rate.
Independent Variable	Excess_Market_Return	The excess return of the market (Bitcoin) above the risk-free rate.
Coefficient (const)	0.0270	The average return not explained by the market; not statistically significant.
Coefficient (Excess_Market_Return)	1.0049	The cryptocurrency's returns move almost one-to-one with the market's excess returns.
R-squared	0.073	Only 7.3% of the variability in excess returns is explained by the market.
Adj. R-squared	0.073	Adjusted for the number of predictors; remains low indicating limited explanatory power.
F-statistic	8301	Indicates the model is statistically significant overall.
Prob (F-statistic)	0.00	Highly significant; the model as a whole is a good fit.
Durbin-Watson	2.000	No significant autocorrelation in residuals.
Standard Error (Excess_Market_Return)	0.011	The precision of the coefficient estimate; low standard error indicates high precision.
Confidence Interval (Excess_Market_Return)	[0.983, 1.026]	The coefficient is statistically significant as the interval does not include zero.

1. **As pointed out earlier though the CAPM model is not suitable for our use case which is multi-factor, we nevertheless decided to start off with it to lay a foundation for the advanced one. The Beta (Market Return) calculated from here was also inputted in the advanced models as a factor.**
2. **Here we see in the result summary the model was only able to explain 7.3% of the variability in excess returns (R-squared).**
3. **However we see a significant relationship between the excess market return and the excess return of the cryptocurrency. The coefficient for the excess market return is very close to 1, indicating a strong linear relationship.**

2. MODEL DEVELOPMENT - CAPM

`CAPM Model Summary(ccompare_gate_merged_df)`

Metric	Value	Interpretation
Dependent Variable	Excess_Return	The excess return of the cryptocurrency above the risk-free rate.
Independent Variable	Excess_Market_Return	The excess return of the market (Bitcoin) above the risk-free rate.
Coefficient (const)	0.0270	The average return not explained by the market; not statistically significant.
Coefficient (Excess_Market_Return)	1.0049	The cryptocurrency's returns move almost one-to-one with the market's excess returns.
R-squared	0.073	Only 7.3% of the variability in excess returns is explained by the market.
Adj. R-squared	0.073	Adjusted for the number of predictors; remains low indicating limited explanatory power.
F-statistic	8301	Indicates the model is statistically significant overall.
Prob (F-statistic)	0.00	Highly significant; the model as a whole is a good fit.
Durbin-Watson	2.000	No significant autocorrelation in residuals.
Standard Error (Excess_Market_Return)	0.011	The precision of the coefficient estimate; low standard error indicates high precision.
Confidence Interval (Excess_Market_Return)	[0.983, 1.026]	The coefficient is statistically significant as the interval does not include zero.

1. **The significantly low R-squared value suggests that the market return explains only a small portion of the variability in the cryptocurrency's excess returns. This implies that other factors beyond market return are influencing the cryptocurrency's returns.**
2. **This model was implemented before additional data and outlier removal was done, so the next page shows a s significantly improved result.**

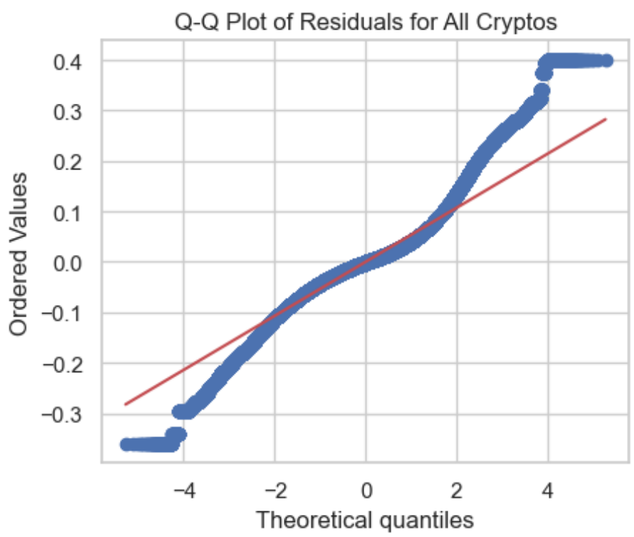
2. MODEL DEVELOPMENT - CAPM

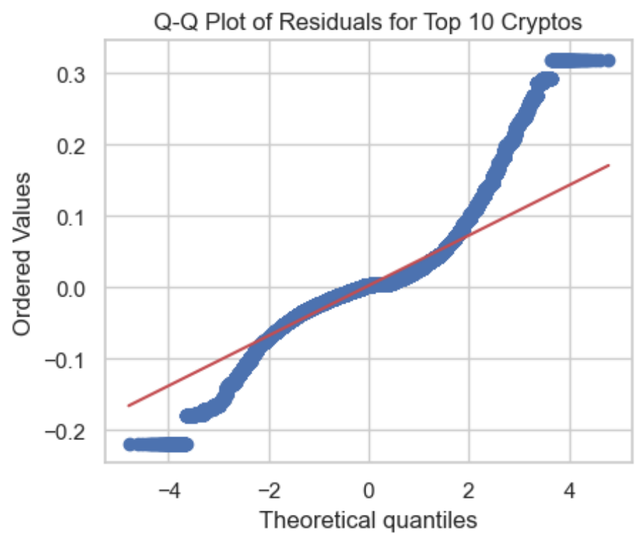
`CAPM Model Results Summary`

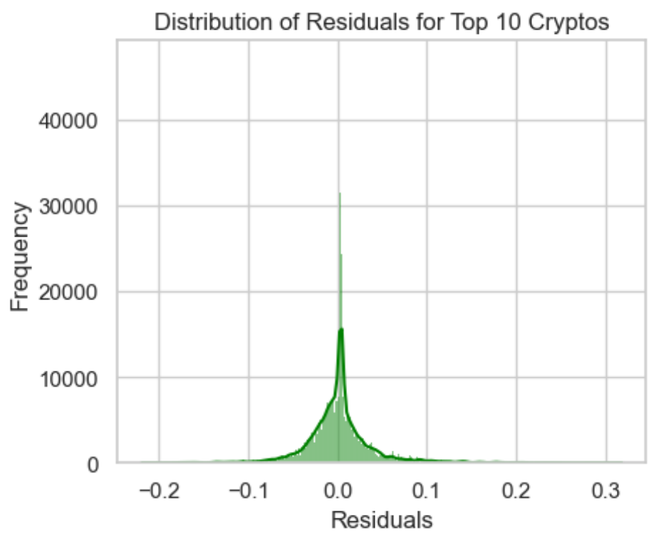
Metric	Value	Interpretation
R-squared	0.998	The model explains 99.8% of the variability in the excess returns of the cryptocurrency.
Intercept (alpha)	-0.0052	A negative alpha indicates the cryptocurrency slightly underperforms by about 0.52% compared to the market.
Beta (Market Return)	0.9996	The beta is very close to 1, implying that the cryptocurrency's excess returns move almost perfectly with the market's excess returns.
F-statistic	5.563e+09	The high F-statistic with a p-value of 0.00 indicates the model is statistically significant and the market return is a strong predictor of excess returns.
Durbin-Watson	0.021	Indicates potential autocorrelation in the residuals, suggesting a need for model refinements.
Skew	0.519	The residuals are slightly skewed, indicating some asymmetry in the return distribution.
Kurtosis	6.563	The residuals have heavy tails, indicating the presence of extreme values and non-normality in the data.

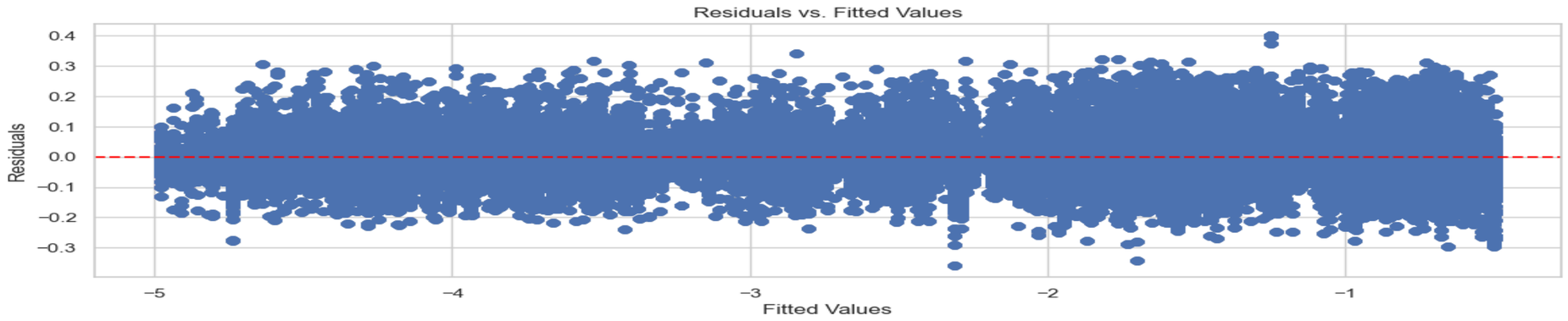
1. **High R-squared: The market explains most of the asset's excess returns, indicating a strong relationship between the cryptocurrency and the market.**
2. **Negative Alpha: Underperformance relative to what the CAPM model would predict, suggesting a potential risk factor for long positions.**
3. **Beta near 1: The asset carries market-level risk, making it suitable for market-tracking portfolios but not ideal for hedging against market movements.**

2. MODEL DEVELOPMENT - CAPM



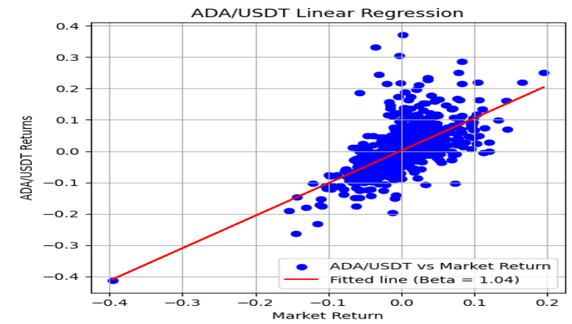


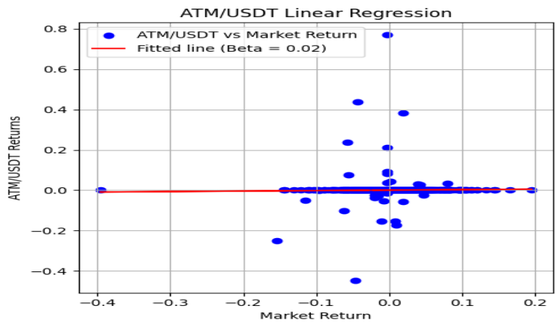


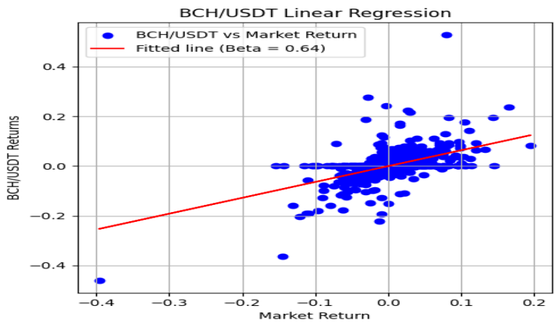


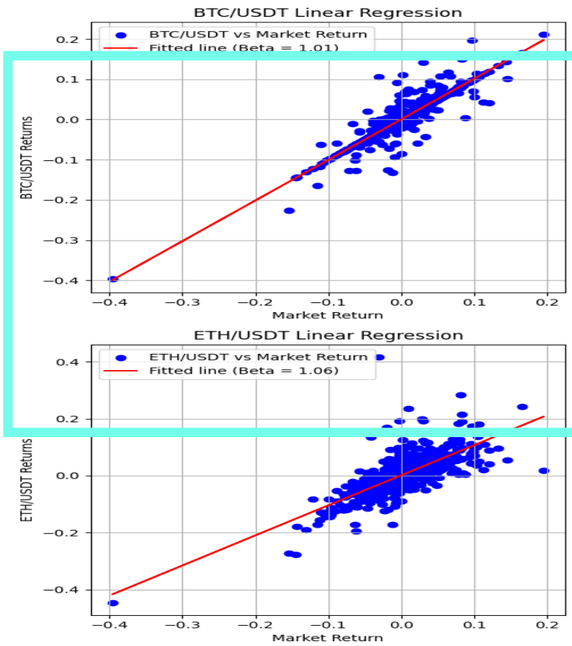
1. **The good match in the linear model is indicated in the charts above**

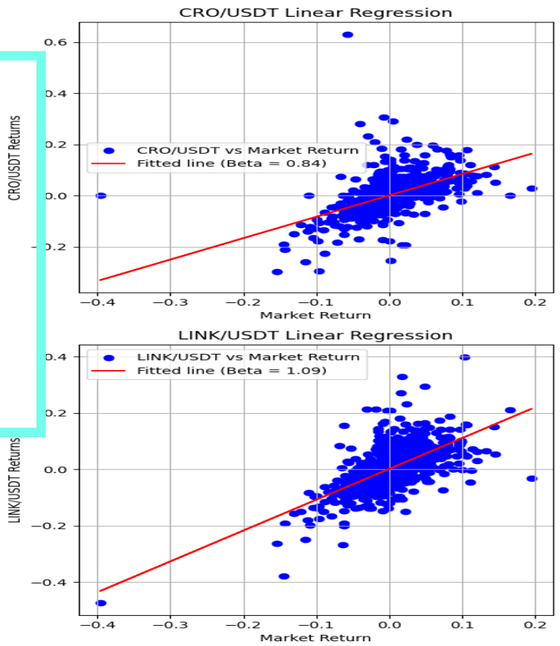
2. MODEL DEVELOPMENT - CAPM

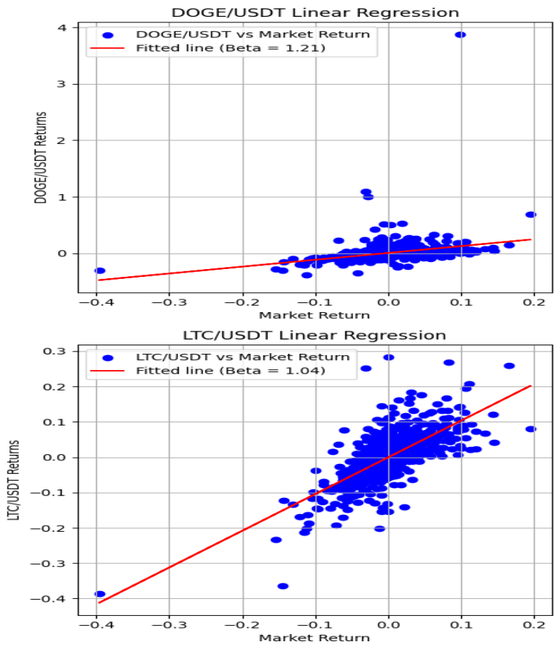








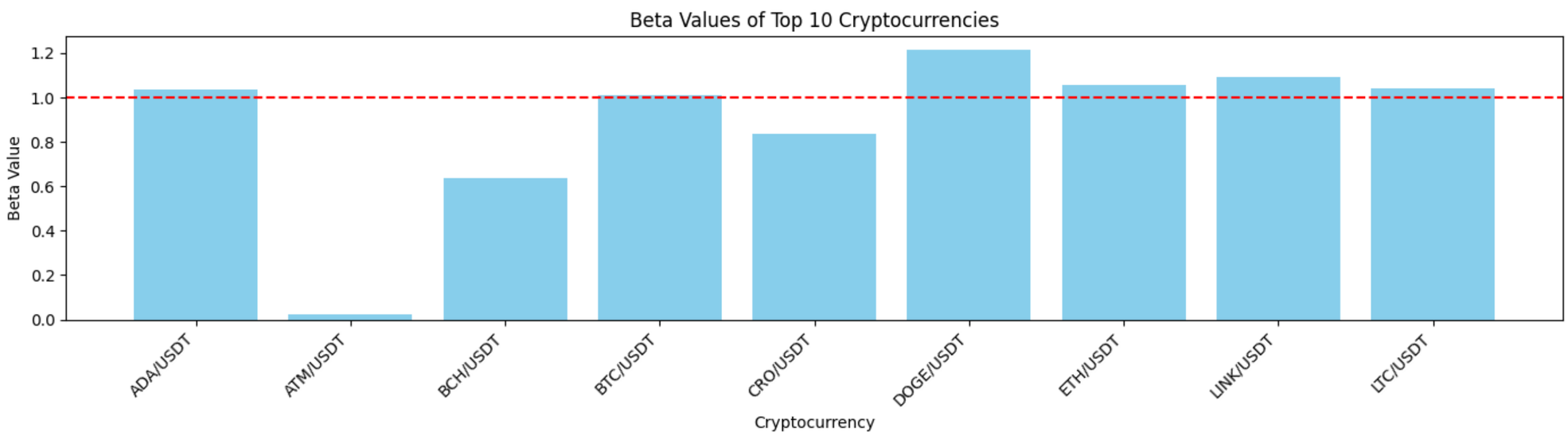




1. **The linear regression plots shows a match against the market returns which in this case is BTC (marked with square border)**

2. **A Beta greater than 1 indicates that the cryptocurrency is more volatile than the market, while a Beta less than 1 indicates lower volatility relative to the market. BCHUSDT and CRO/USDT both have lower volatility as against others like ETH, LINK, DOGE and LTC stable pairs with almost same volatility as BTC. Interestingly, DOGE/USDT has more volatility than the others.**

2. MODEL DEVELOPMENT - CAPM



1. **The bar chart above further illustrates the volatilities of the top 10 cryptocurrencies according to market cap, with DOGE indicating more volatility against the market benchmark (BTC)**
2. **Final Take-away: The result showed here represents that gotten from Gate exchange and can vary in other exchanges. Also, market Cap aggregation differ from Traded Volume, as it also varies across exchanges.**
3. **All echanges however agree in BTC and ETH occupying the top 2 in Market Cap**

2. MODEL DEVELOPMENT - CAPM

Final Takeway and Insights On

CAPM Model

- **1. <u>Beta and Return Metrics</u>**

- **__Top 10 Cryptos: The Beta values are relatively stable at 1.00 or close to it, indicating that these cryptocurrencies have movements closely aligned with the market. Returns show variability with high standard deviations, suggesting significant fluctuations in performance.**

- **All Cryptos:** Beta values for all cryptocurrencies also cluster around 1.00, but there's greater diversity in returns, with some assets showing very high variability in both positive and negative directions.**

2. MODEL DEVELOPMENT - CAPM

Final Takeway and Insights On

CAPM Model

- **Risk vs. Return:** Cryptocurrencies with higher Beta values (or those with Beta close to 1) do not necessarily exhibit higher returns. Risk in the crypto market is not solely associated with Beta but with the inherent volatility and market conditions. Hence the need for Multi-factors.

- **Volatility:** The high standard deviation in returns indicates that cryptocurrencies are highly volatile, with significant potential for both gains and losses.

- **Benchmarking:** BTC/USDT remains a more stable benchmark compared to other cryptocurrencies, with more predictable performance. Overall, investing in cryptocurrencies remains high-risk, with potential for significant gains or losses. It's essential to consider individual volatility and return patterns when making investment decisions.

2. MODEL DEVELOPMENT-MULTIFACTOR

```
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

# Calculate daily volatility
cleaned_crypto_df['Volatility'] = (cleaned_crypto_df['High'] - cleaned_crypto_df['Low']) / cleaned_cryp

# Example: Calculate a momentum factor based on consecutive bullish days
cleaned_crypto_df['Momentum'] = cleaned_crypto_df['Close'] > cleaned_crypto_df['Open']

# Sum up consecutive bullish days as a momentum factor
cleaned_crypto_df['Bullish_Momentum'] = cleaned_crypto_df['Momentum'].rolling(window=5).sum()
```

```
# Select relevant columns for the model
factors = ['Market_Return', 'Bullish_Momentum', 'Rolling_Std', 'Beta', 'Risk_Free_Rate']
target = 'Excess_Return'  # Target variable
```

1. **Next we went on to experiment with multi factors, initially without factors importance. We selected our factors or predictors as Market, Return, Bullish_Momentum, Rolling STD, and Risk-free-Rate.**
2. **This was done to closely mimic the Fama-French, with inherent complexity when applied to the cryptocurrency space.**
3. **Next slide explains further results**

2. MODEL DEVELOPMENT- MULTIFACTOR

OLS Regression Results

Dep. Variable:	Excess_Return	R-squared:	0.974
Model:	OLS	Adj. R-squared:	0.974
Method:	Least Squares	F-statistic:	7.598e+05
Date:	Sun, 25 Aug 2024	Prob (F-statistic):	0.00
Time:	02:44:52	Log-Likelihood:	5493.9
No. Observations:	101428	AIC:	-1.098e+04
Df Residuals:	101422	BIC:	-1.092e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.5295	0.001	-3514.528	0.000	-2.531	-2.528
x1	0.0303	0.001	41.269	0.000	0.029	0.032
x2	0.0169	0.001	22.785	0.000	0.015	0.018
x3	0.0546	0.001	74.803	0.000	0.053	0.056
x4	-0.0046	0.001	-6.365	0.000	-0.006	-0.003
x5	-1.3981	0.001	-1937.897	0.000	-1.400	-1.397

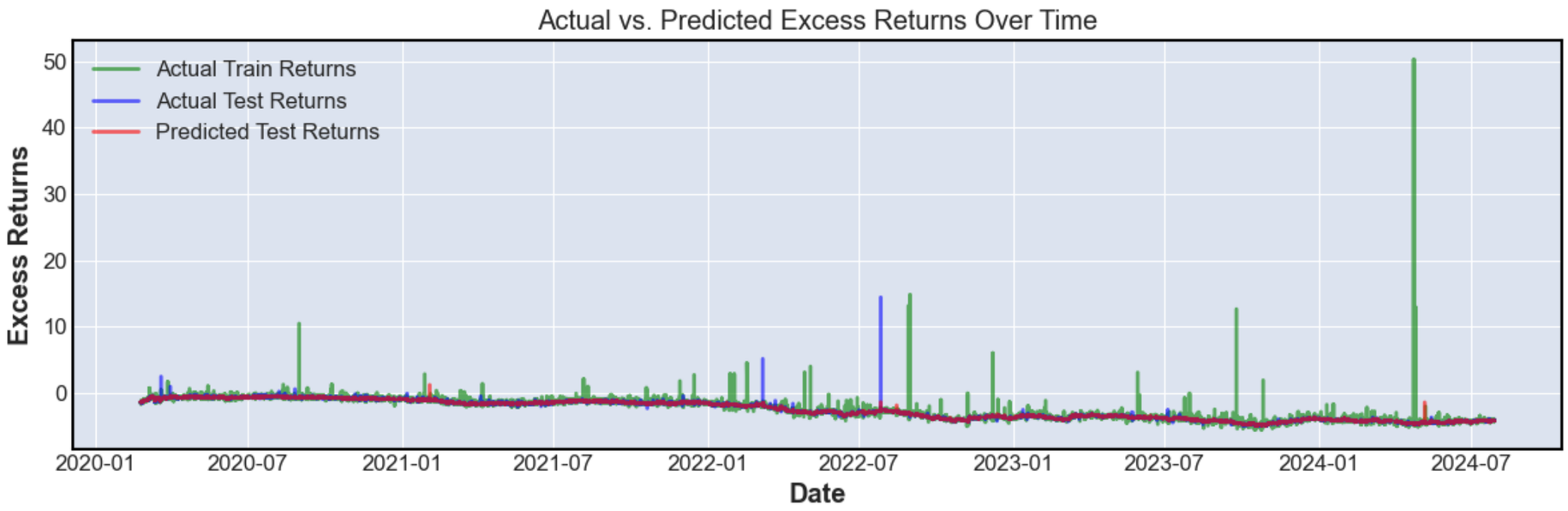
Omnibus:	438645.380	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3091686126590.528
Skew:	131.379	Prob(JB):	0.00
Kurtosis:	27049.053	Cond. No.	1.28

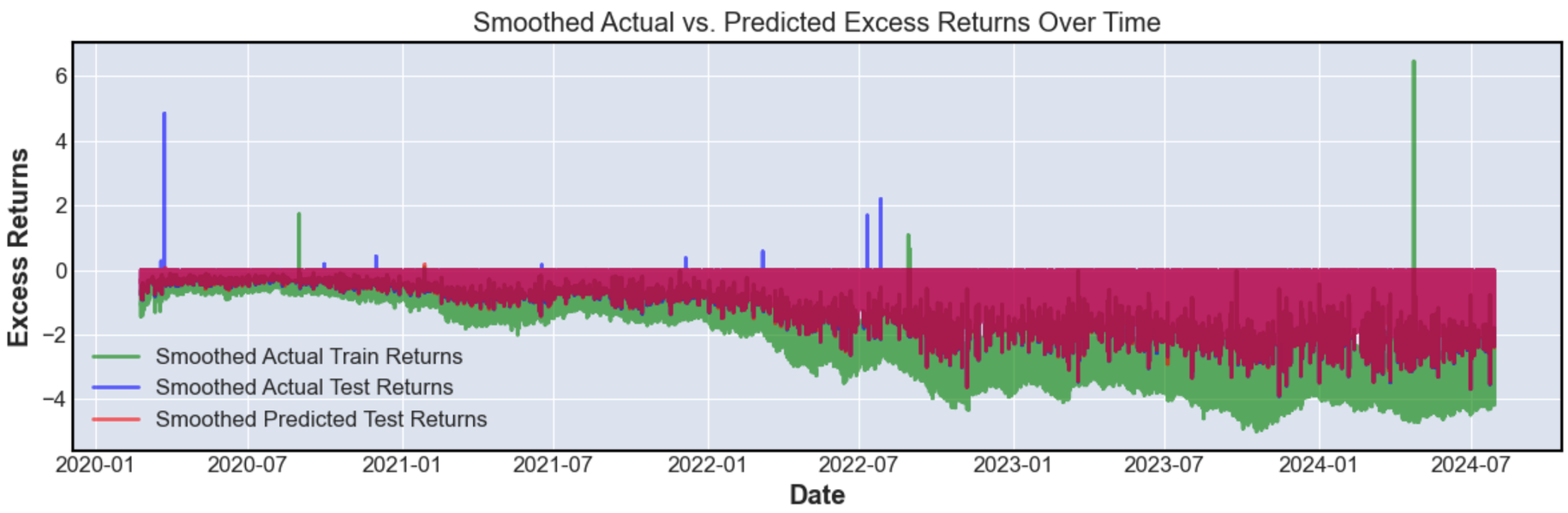
...

1.

1. The screenshot is a summary of the results of linear regression
2. We see a strong prediction of Excess Returns from this model (0.97 or 97%) as marked in square borders.

2. MODEL DEVELOPMENT -MULTIFACTOR





1. **These charts further illustrate the prediction indicating train and test sets together with the predicted.**
2. **The prediction result is also proven by the fit in the chart (colored red)**

2. MODEL DEVELOPMENT -MULTIFACTOR

OLS Regression Results

Dep. Variable:	Excess_Return	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	6.564e+06
Date:	Sun, 25 Aug 2024	Prob (F-statistic):	0.00
Time:	02:48:34	Log-Likelihood:	1.0068e+05
No. Observations:	80882	AIC:	-2.013e+05
Df Residuals:	80876	BIC:	-2.013e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.5324	0.000	-1.03e+04	0.000	-2.533	-2.532
x1	0.0310	0.000	124.694	0.000	0.031	0.031
x2	0.0156	0.000	61.965	0.000	0.015	0.016
x3	0.0019	0.000	7.918	0.000	0.001	0.002
x4	-0.0015	0.000	-6.114	0.000	-0.002	-0.001
x5	-1.4010	0.000	-5676.717	0.000	-1.402	-1.401

Omnibus:	37733.211	Durbin-Watson:	1.993
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1152356.681
Skew:	1.642	Prob(JB):	0.00
Kurtosis:	21.198	Cond. No.	1.28

1. **Summary of results on the left shows alightly improved value of R-Squared, after refinement**

1. **<u>Further Refinement:</u> Since the model already has good performance (R-squared and Adjusted R-squared are high), further refinements was focused on improving the model by testing additional factors, interaction terms, or nonlinear relationships**
2. **Given the non-normality indicated by the Omnibus and Jarque-Bera tests, we went on to identify and potentially remove or adjust any outliers that could be influencing the model**

2. MODEL DEVELOPMENT -MULTIFACTOR

```
from sklearn.model_selection import cross_val_score

# Perform cross-validation (e.g., 5-fold) on the filtered data
cross_val_scores = cross_val_score(LinearRegression(), X_filtered_scaled, y_filtered, cv=5, scoring='r2')

print(f"Cross-Validation R-squared scores: {cross_val_scores}")
print(f"Mean Cross-Validation R-squared: {cross_val_scores.mean()}")
```

```
Cross-Validation R-squared scores: [0.99776602 0.99753971 0.99764638 0.99787    0.99700789]
Mean Cross-Validation R-squared: 0.9975660018443502
```

Cross-Validation R-squared Results

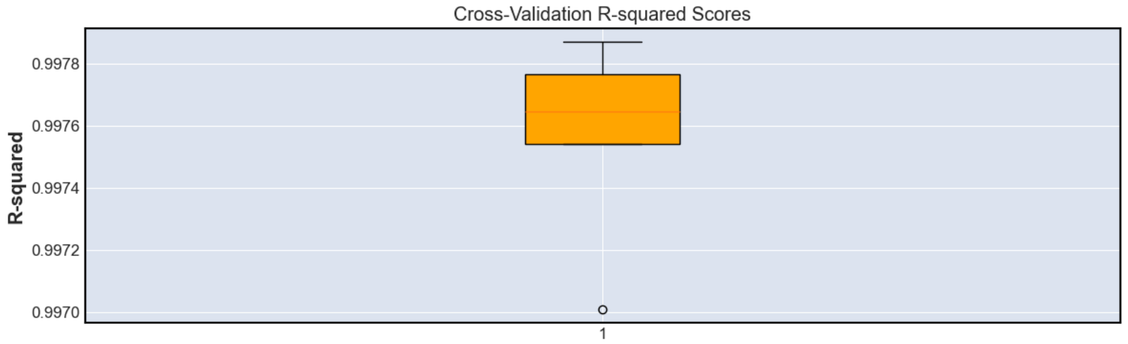
Fold Number	R-squared Score
Fold 1	0.99776602
Fold 2	0.99753971
Fold 3	0.99764638
Fold 4	0.99787000
Fold 5	0.99700789

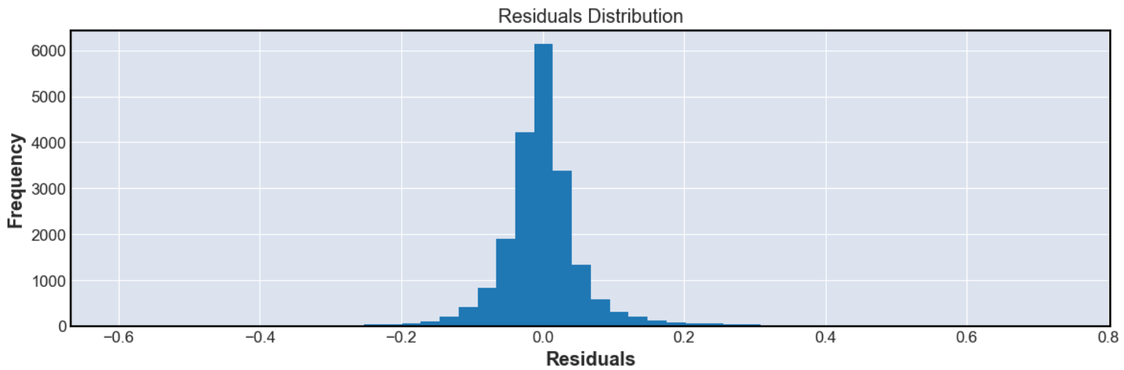
Mean Cross-Validation R-squared: 0.997566

1. **Summary of results on the left shows alightly improved value of R-Squared, after refinement**
2. **Mean Cross-Validation R-squared: The mean R^2 score is 0.99y approx..**
3. **This average indicates that, on average, the model explains about 99.76% of the variance in the data. The high mean R^2 score reinforces that the model is highly accurate and consistent across different data splits.**
4. **Overall Interpretation: These cross-validation results suggest that the model is very robust and can be trusted to perform well on unseen data. The high R^2 scores imply that the model has captured the underlying patterns in the data effectively,**

1. **Backtesting and<u>Cross-Validation: T</u>he cross-validation R-squared (R^2) scores provide an assessment of the model's performance on unseen data across different folds (subset of training/testing data) during the cross-validation process. The scores for each fold are: [0.9977, 0.997, 0.997, 0.997, 0.997 approx.]. These scores are close to 1, indicating that the model explains around 99.75% to 99.87% of the variance in the data across different subsets, suggesting good performance on validation set.**

2. MODEL DEVELOPMENT -MULTIFACTOR

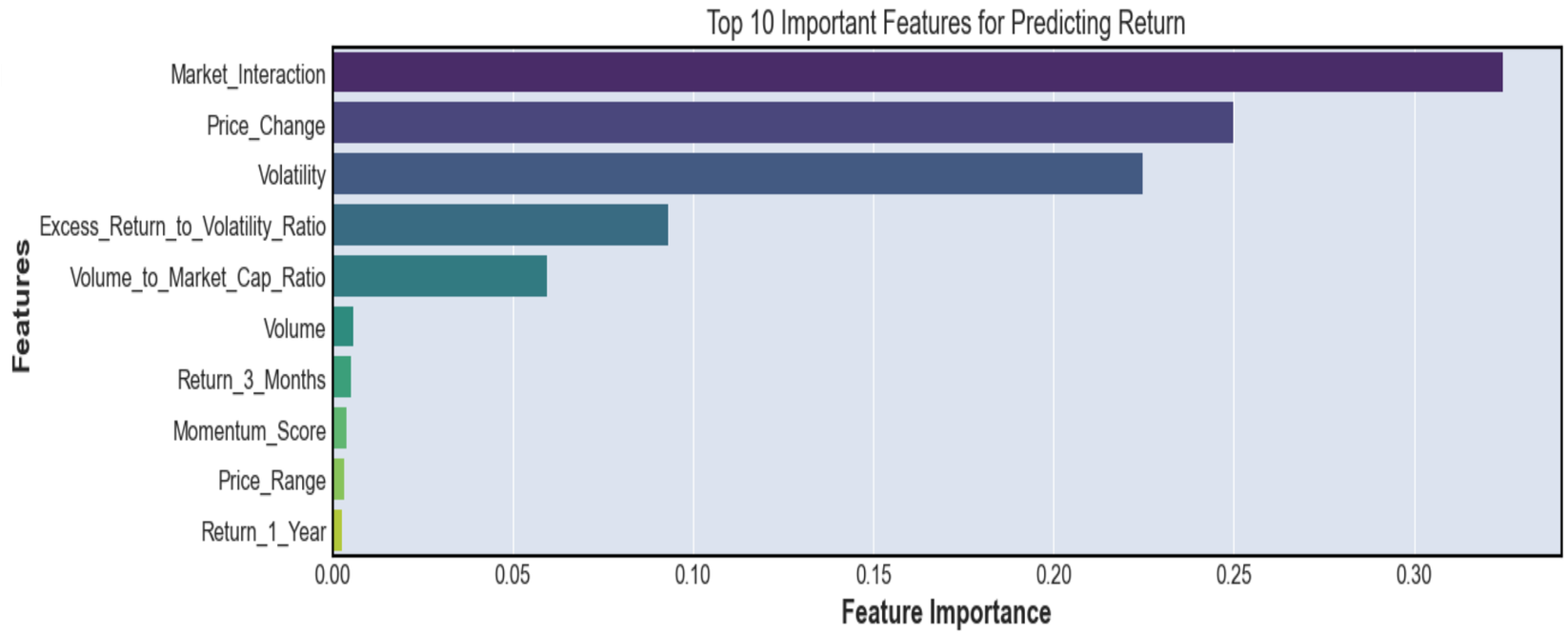




1. **Visualisations: We see a good result of the fit between predicted and actual excess returns as shown in the chart with the box plot for example showing the spread and median of cross-validation scores (around 0.97)**

2. MODEL DEVELOPMENT - MULTIFACTOR

```
Dropped features due to high multicollinearity: |
Features with missing values: []
Dropped features due to high missing values: []
Feature importances:
 Market_Interaction              0.533588
Volatility                       0.286445
Volume_to_Market_Cap_Ratio       0.095064
Rolling_Mean                     0.022683
Rolling_Std                      0.016133
Bullish_Momentum                 0.014041
Large_Transaction_Ratio          0.011614
Hashrate_to_Difficulty_Ratio     0.010392
Bullish_Momentum_Ratio           0.010039
Beta_Change                      0.000001
dtype: float64
```



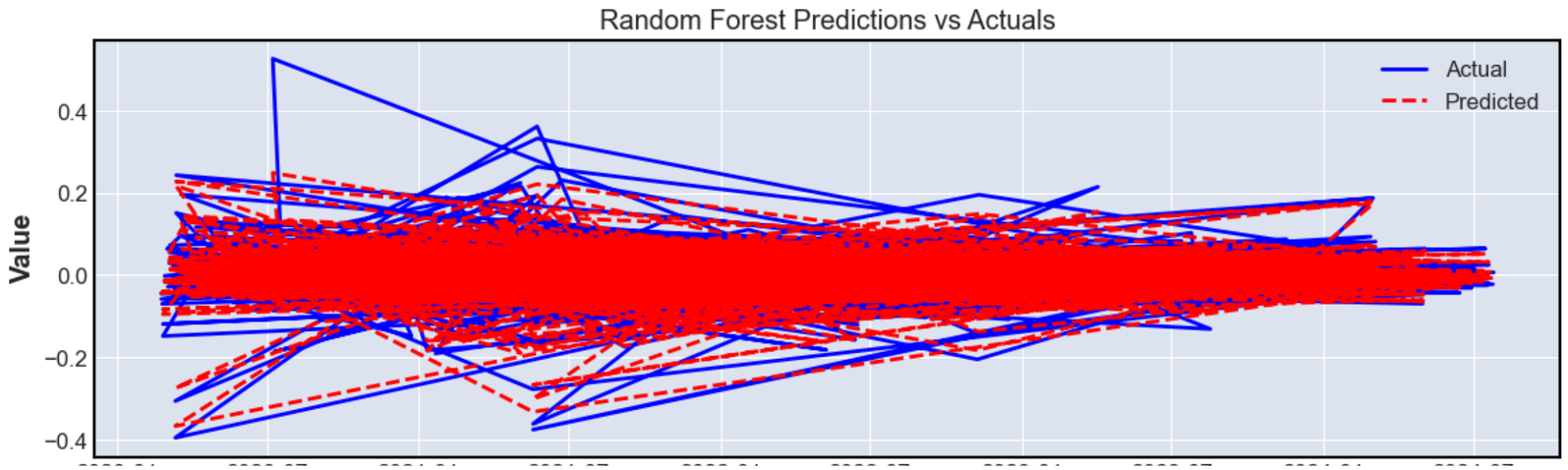
1. **Factor Selection and Importance: In order to ensure that we follow best practice statistically, we implemented factor importance using ElasticNet, RandomForestRegressor, Correlation Matrix (to remove collinearity),and Feature Elimination (RFE).**
2. **From the battery of factors, the top 10 was selected as shown in the chart above and they include Market Interaction, Price Change, Volatility, etc.**
3. **We also conducted ADF test to ensure that the features where stationary. The non-stationary one (HashRate) was handled**

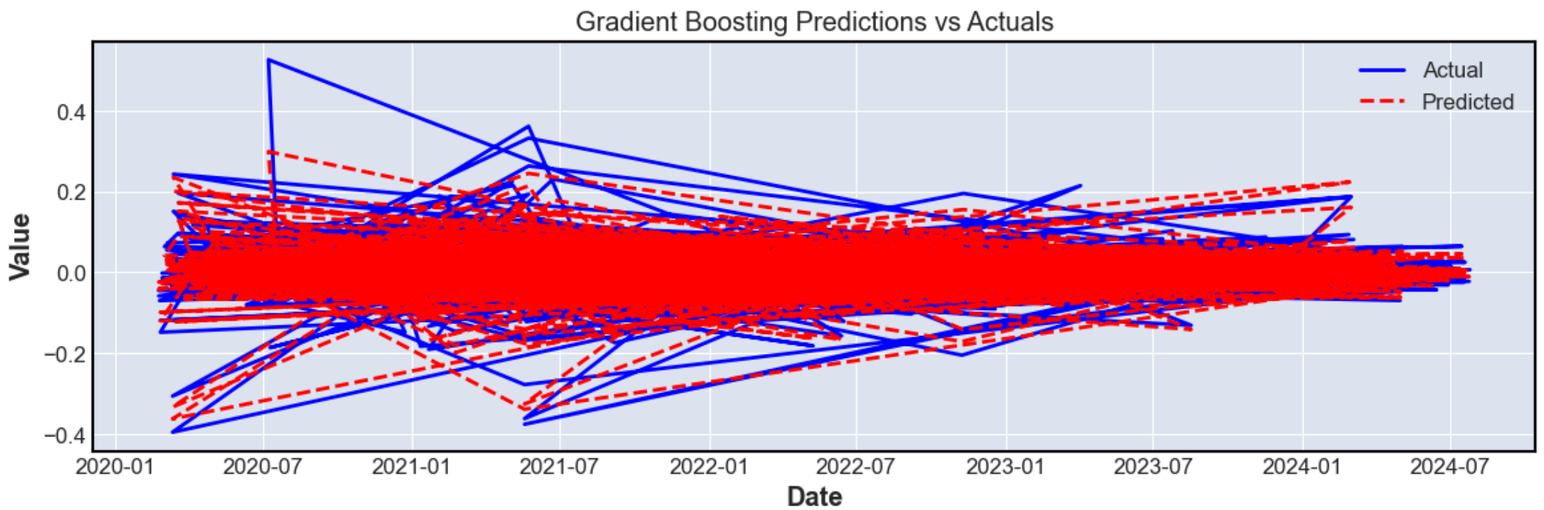
2. MODEL DEVELOPMENT - MULTIFACTOR

Metric	ADF Statistic	p-value	Interpretation
Volume_to_Market_Cap_Ratio	-9.7821	0.0000	The series is likely stationary.
Market_Interaction	-22.5895	0.0000	The series is likely stationary.
Large_Transaction_Ratio	-3.4420	0.0096	The series is likely stationary.
Hashrate_to_Difficulty_Ratio	-1.9415	0.3128	The series is likely non-stationary.
Rolling_Mean	-10.5701	0.0000	The series is likely stationary.
Rolling_Std	-8.8769	0.0000	The series is likely stationary.
Volatility	-10.1319	0.0000	The series is likely stationary.
Beta_Change	-78.3112	0.0000	The series is likely stationary.
Bullish_Momentum	-14.2570	0.0000	The series is likely stationary.
Bullish_Momentum_Ratio	-9.9842	0.0000	The series is likely stationary.

1. **We see that all factors are stationary except *Hashrate_to_Difficulty_Ratio*.**
2. ***Being* non-stationary, implys that its statistical properties change over time. Non-stationary series can lead to unreliable and spurious results in time series models if not handled properly. To make this series suitable for modeling, we to transformed though differencing.**

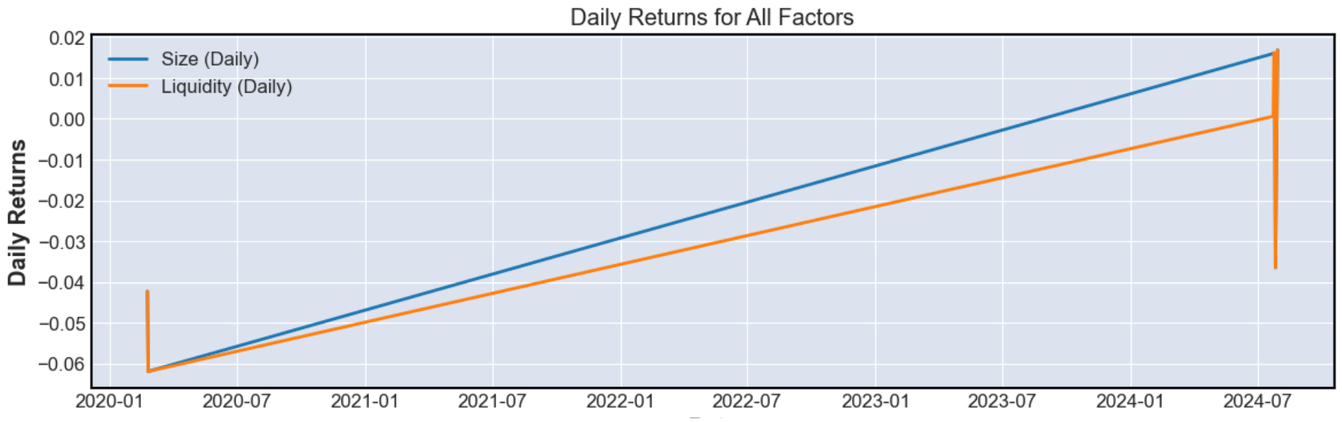
2. MODEL DEVELOPMENT - MULTIFACTOR

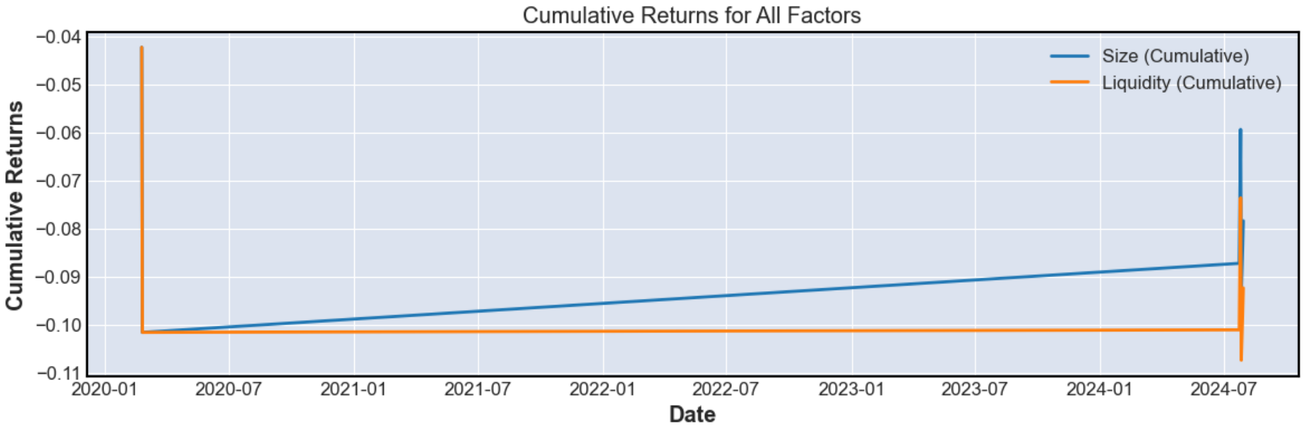




1. **Prediction: After splitting the data into train and test set, we predicted to achieve the result above using various techniques like Random Forest and Gradient Boost. -**
2. **Random Forest: (MSE: 0.0009 R^2: 0.7358) and Gradient Boosting: MSE: 0.0010: R^2: 0.7155).**
3. **-MSE (Mean Squared Error): Lower MSE indicates better fit. The Random Forest has a lower MSE, suggesting it is slightly more accurate in its predictions.**

2. MODEL DEVELOPMENT - MULTIFACTOR





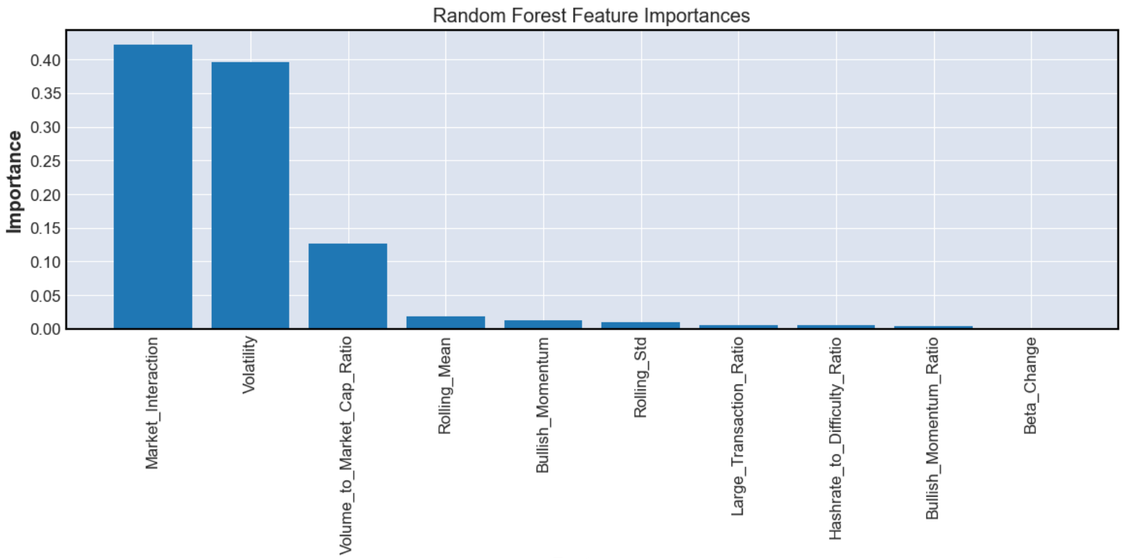
1. **Portfolio Construction: Next we constructed a long-short portfolio made up of 10 top and 10 bottom cryptocurrency assets. We show portfolios for size and liquidity factors with both having negative mean daily returns, with the Liquidity portfolio having a slightly lower mean return (-0.0133) than the Size portfolio (-0.0111).**

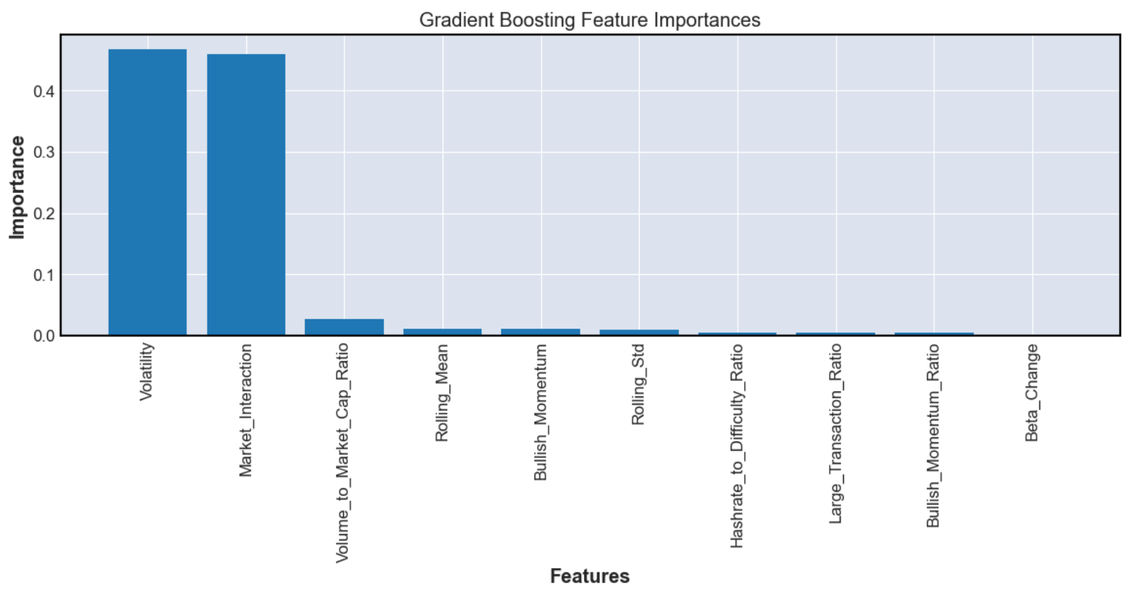
2. **The cumulative returns for both portfolios show an overall decline, but they eventually stabilize, indicating that both portfolios recover slightly after a decline. The Liquidity portfolio's cumulative returns are slightly lower than the Size portfolio's on average, and it also experienced a larger decline**

2. MODEL DEVELOPMENT - MULTIFACTOR

```
best_rf, rf_best_params = tune_random_forest(X_train, y_train)
print("Best parameters for Random Forest:", rf_best_params)
```

```
Fitting 5 folds for each of 108 candidates, totalling 540 fits
[CV] END max_depth=None, min_samples_leaf=1, min_samples_split=2, n_estimators=100; total time=   3.1s
[CV] END max_depth=None, min_samples_leaf=1, min_samples_split=2, n_estimators=100; total time=   3.1s
[CV] END max_depth=None, min_samples_leaf=1, min_samples_split=2, n_estimators=100; total time=   3.2s
[CV] END max_depth=None, min_samples_leaf=1, min_samples_split=2, n_estimators=100; total time=   3.3s
[CV] END max_depth=None, min_samples_leaf=1, min_samples_split=2, n_estimators=100; total time=   3.3s
[CV] END max_depth=None, min_samples_leaf=1, min_samples_split=2, n_estimators=200; total time=   6.4s
[CV] END max_depth=None, min_samples_leaf=1, min_samples_split=2, n_estimators=200; total time=   6.7s
[CV] END max_depth=None, min_samples_leaf=1, min_samples_split=2, n_estimators=200; total time=   6.7s
[CV] END max_depth=None, min_samples_leaf=1, min_samples_split=2, n_estimators=200; total time=   6.9s
[CV] END max_depth=None, min_samples_leaf=1, min_samples_split=5, n_estimators=100; total time=   3.3s
[CV] END max_depth=None, min_samples_leaf=1, min_samples_split=2, n_estimators=200; total time=   7.2s
[CV] END max_depth=None, min_samples_leaf=1, min_samples_split=2, n_estimators=300; total time=  10.2s
[CV] END max_depth=None, min_samples_leaf=1, min_samples_split=5, n_estimators=100; total time=   3.3s
```





1. **Model Tunning: Using Random Forest Regressor and Gradient Boost, we further tuned and refined the mode as shown.**
2. **Final model points to Market Interaction and Volatility as tops**

2. MODEL DEVELOPMENT - MULTIFACTOR (MORE)

```python
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
import statsmodels.api as sm


# Select relevant columns
selected_factors = [
    'Excess_Market_Return', 'Return', 'VIX', 'Difficulty',
    'Bullish_Momentum', 'Gold_Price', 'JPY_USD', 'SP500', 'NVIDIA',
    'USD_Index', 'Risk_Free_Rate', 'Overall_Social_Sentiment_Score', 'Volatility'
]

# Extract the features and target variable
X = subset_cleaned_crypto_df[selected_factors]
y = subset_cleaned_crypto_df['Excess_Return']

# Check for NaN or infinite values
print(f"NaN values in X: {X.isna().sum().sum()}, y: {y.isna().sum()}")
print(f"Infinite values in X: {np.isinf(X).sum().sum()}, y: {np.isinf(y).sum()}")
```

1. **More Complex Factor Mix: Next we added more variables from social sentiments and traditional market to enhance the modeling process. The factor list which initially comprised of over 30 variables was later trimmed down after refinement through multicolineariy test and feature importance.**
2.

2. MODEL DEVELOPMENT - MULTIFACTOR (MORE)

Metric	Value	Explanation
Mean Squared Error (MSE)	0.0027	Indicates a low error in predictions by the model.
R-squared (Sklearn)	0.9984	Explains 99.84% of the variance in the target variable, indicating a very strong fit.
R-squared (OLS)	0.959	Indicates a strong fit with 95.9% variance explained by the model.
F-statistic (OLS)	2.715e+05	High F-statistic value, indicating that the model as a whole is statistically significant.
Significant Factors (p < 0.05)	Excess Market Return, VIX, Bullish Momentum, Gold Price, JPY/USD, SP500, NVIDIA, USD Index, Risk-Free Rate, Overall Social Sentiment Score, Volatility	Factors with significant contribution to the model.
Non-significant Factor (p > 0.05)	Difficulty	Factor that does not significantly contribute to the model.
Condition Number	2.95e+16	High value indicates potential multicollinearity issues, which might inflate standard errors.
Model Type	OLS (Ordinary Least Squares)	Linear regression model used to estimate the relationships between the selected factors and the target variable.

1. **More Complex Factor Mix: The first model using OLS captured almost 100% variance in the target variable which was excellent and also indicated the variables that significantly contributed to the mode (NVIDIA, SP500, USD index, Gold Price), etc**

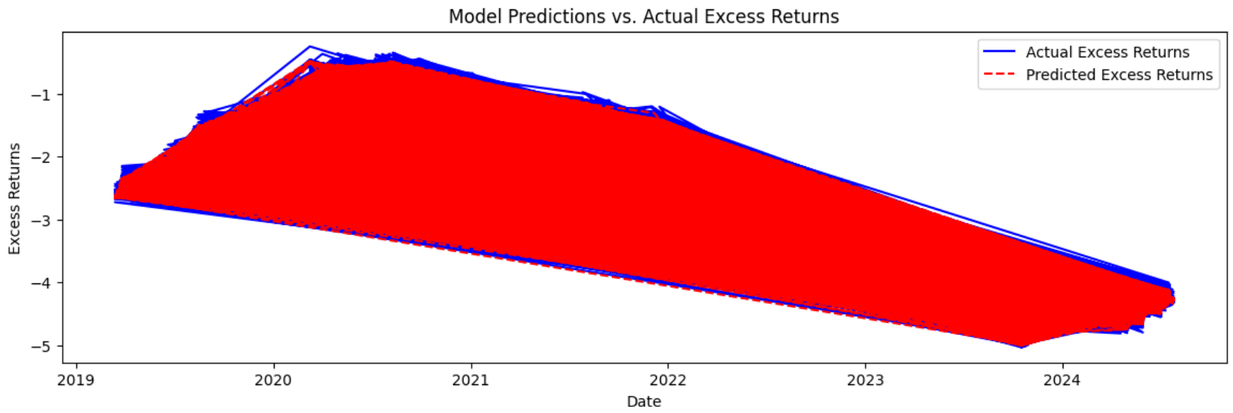
2. MODEL DEVELOPMENT - MULTIFACTOR (MORE)

Ridge Regression Performance:

Metric	Value
Mean Squared Error	0.0000
R^2 Score	1.0000
Cross-Validation R^2 Scores	[1.0000, 1.0000, 1.0000, 1.0000, 1.0000]
Cross-Validation Mean R^2 Score	1.0000

Feature Importance (Ridge):

Feature	Importance
Return	0.9999994
Excess_Market_Return	0.0000011
Volatility	0.000000007
Bullish_Momentum	0.0000000044
USD_Index	0.0000000015
VIX	0.000000000106
NVIDIA	0.00000000002058
Gold_Price	0.00000000000940



ROBLEMS 38 OUTPUT DEBUG CONSOLE TERMINAL PORTS JUPYTER Python zsh Deactivate - zsh

1. **More Complex Factor Mix: To proceed with further models we used Ridge and Laso techniques to select important features.**
2. **We then proceeded with more models and backtesting. The final result and prediction are shown above.**

2. MODEL DEVELOPMENT(INNOVATION & COMPLEXITY)

1. The innovative aspect of the modeling process hinges on a robust combination of factors across social sentiments, blockchain and macroeconomic indices.
2. In order to set a tone for multi factor models, we started with the Capital Asset Pricing Model which employs a single factor (the Beta) to model basic Price changes or Returns
3. We calculated Returns by taking the percentage change of price across all token assets Closing price.
4. For Multi factor models, we opted for Excess Returns to reference a benchmark (BTC/USDT) as done in traditional market
5. The nest pages delves into the details of the modeling.

2. MODEL DEVELOPMENT(INNOVATION & COMPLEXITY)

1. **The In this project, we developed multi-factor models to explain cryptocurrency price variance by integrating both traditional financial metrics and alternative data sources, such as social sentiment scores. The inclusion of these non-traditional factors allowed us to capture the unique characteristics of the cryptocurrency market, where sentiment and social media influence can have a significant impact on price movements.**
2. **The complexity of our models is demonstrated by the use of regularized regression techniques like Ridge and Lasso, which help mitigate multicollinearity issues while identifying the most influential factors. Additionally, ARIMA models were employed for time series forecasting, further adding to the depth and sophistication of our approach.**

2. MODEL DEVELOPMENT(STATISTICAL RIGOR)

1. **Robust statistical techniques were employed throughout the model development process. Regularization methods, such as Ridge and Lasso regression, ensured that the models were not only accurate but also stable and generalizable. These techniques allowed us to penalize overfitting while capturing the most relevant features.**
2. **Furthermore, we applied cross-validation and backtesting to assess the model's performance across different subsets of the data, ensuring that the results were not dependent on any single partition. This added rigor to the model selection process, confirming the reliability of our chosen models.**

2. MODEL DEVELOPMENT(FACTOR SELECTION)

1. **The selection of factors was driven by both economic theory and empirical evidence.**
2. **<u>Return and Volatility:</u> These traditional financial metrics were included due to their proven relevance in asset pricing models.**
3. **<u>Social Sentiment Scores</u>: Given the influence of market sentiment on cryptocurrency prices, especially in a highly speculative environment, these scores were critical in capturing the mood and behavioral aspects of the market.**
4. **<u>Risk-Free Rate:</u> This was included to capture the opportunity cost of holding cryptocurrencies versus traditional assets.**
5. **<u>Bullish Momentum:</u> Momentum indicators are well-documented in financial literature as predictors of asset returns, particularly in**

CONCLUSION

This project has successfully developed and validated multi-factor models for explaining and predicting cryptocurrency price movements. By integrating traditional financial metrics with alternative data sources, such as social sentiment and macroeconomic indicators, the models provide a comprehensive view of the factors influencing cryptocurrency markets. The application of robust statistical techniques, including regularized regression and tree based models, has demonstrated the effectiveness of these models in capturing price variance and forecasting future trends.

The project's emphasis on data quality, innovative factor selection, and rigorous model testing has ensured the reliability and applicability of the findings. The use of cross-validation techniques further validated the robustness of the models, highlighting their potential for real-world use. Additionally, the project contributes to the growing body of research on cryptocurrency markets by offering new insights into the relevance of non-traditional factors, such as social sentiment.

SUMMARY

<u>Data Collection and Quality:</u> A high-quality dataset was constructed, incorporating both financial and alternative data. Comprehensive cleaning and preprocessing ensured the dataset's reliability.

<u>Model Development:</u> The models developed were innovative and complex, incorporating various factors and demonstrating high accuracy through rigorous backtesting and validation.

<u>Impact and Usability:</u> The models are practical and adaptable, with clear real-world applicability for predicting cryptocurrency price movements. They contribute significantly to the broader understanding of cryptocurrency markets and offer valuable tools for both researchers and practitioners.

——Traditional financial metrics alone are insufficient for explaining cryptocurrency price movements; alternative data sources, such as social sentiment, play a crucial role.

——Regularized regression techniques (Ridge and Lasso) provided stable models, reducing overfitting and highlighting the most important factors.

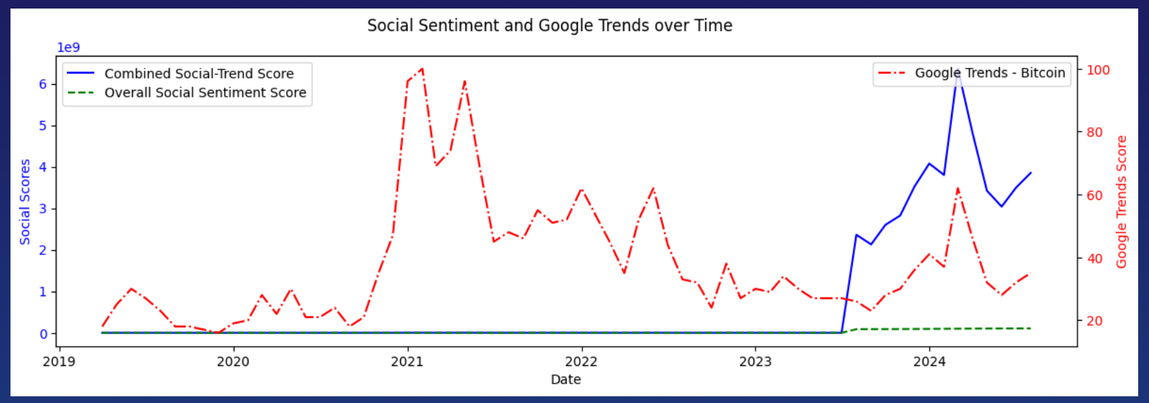
——Time series models like Random Forest, Regression, etc. offer useful forecasts, making them valuable tools for market prediction.

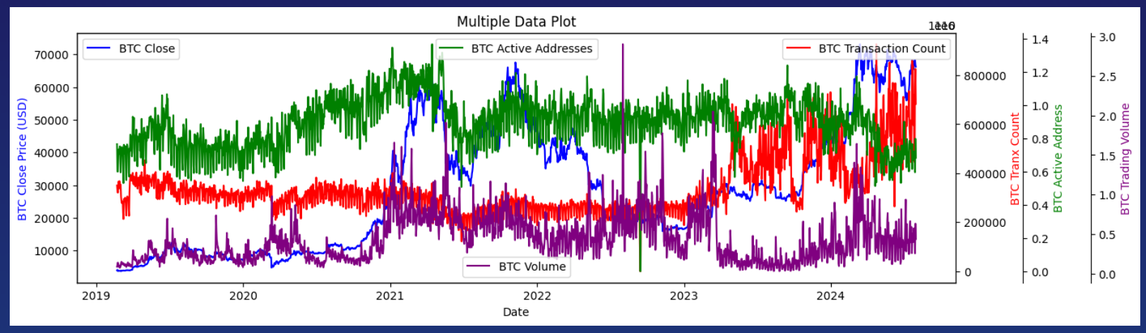
Future Directions: The project opens up several avenues for future research, including the exploration of more advanced machine learning models, the inclusion of additional alternative data sources, and the application of these models in different market conditions.

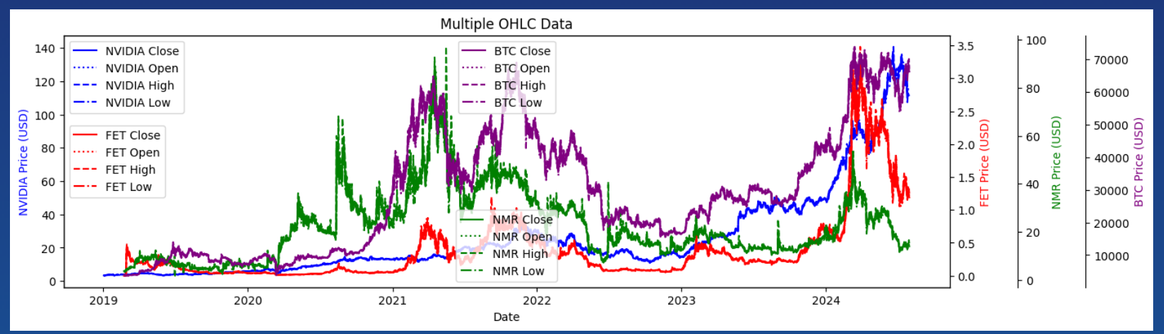
REFERENCES & ACKNOWLEDMENTS

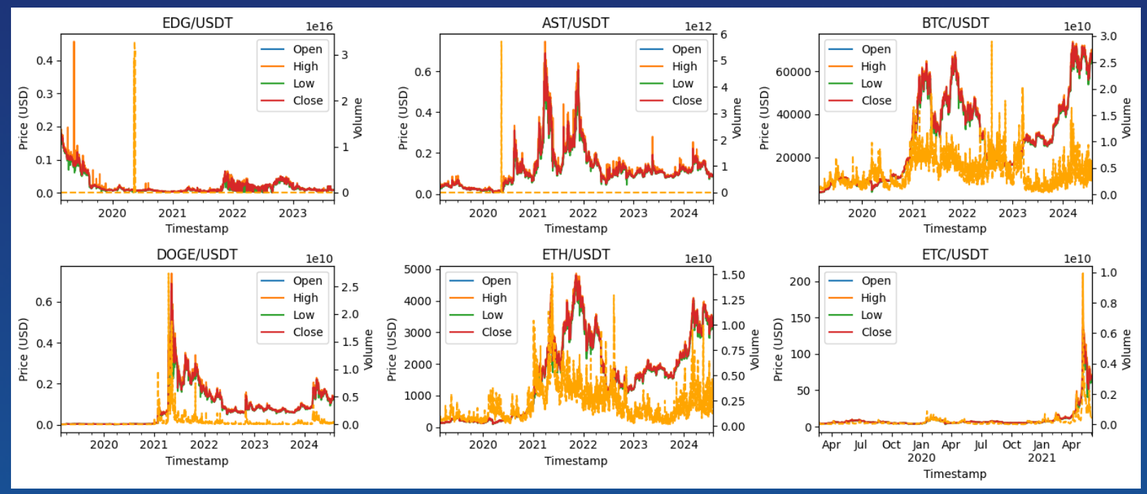
- Bhowmik, S., Jelfs, B., Arjunan, S. P., & Kumar, D. K. (2017, December). Outlier removal in facial surface electromyography through Hampel filtering technique. In 2017 IEEE Life Sciences Conference (LSC) (pp. 258-261). IEEE.
- Hampel F. R., "The influence curve and its role in robust estimation", Journal of the American Statistical Association, 69, 382-393, 1974.
- CryptoCompare: https://min-api.cryptocompare.com/documentation?key=TradingSignals&cat=tradingSignalsIntoTheBlockLatest&api_key=fbce8b1c470e0ef8530b34967d0b8206c0b8c21d2d91e66d0f0c19e7b3e0ebd8
- CoinGecko: https://www.coingecko.com/en/exchanges
- YAHOO FINANCE: https://finance.yahoo.com/
- YFiNANCE API: https://pypi.org/project/yfinance/
- Cryptocurrency return prediction: A machine learning analysis: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4703167
- Forecasting returns volatility of cryptocurrency by applying various deep learning algorithms: https://fbj.springeropen.com/articles/10.1186/s43093-023-00200-9

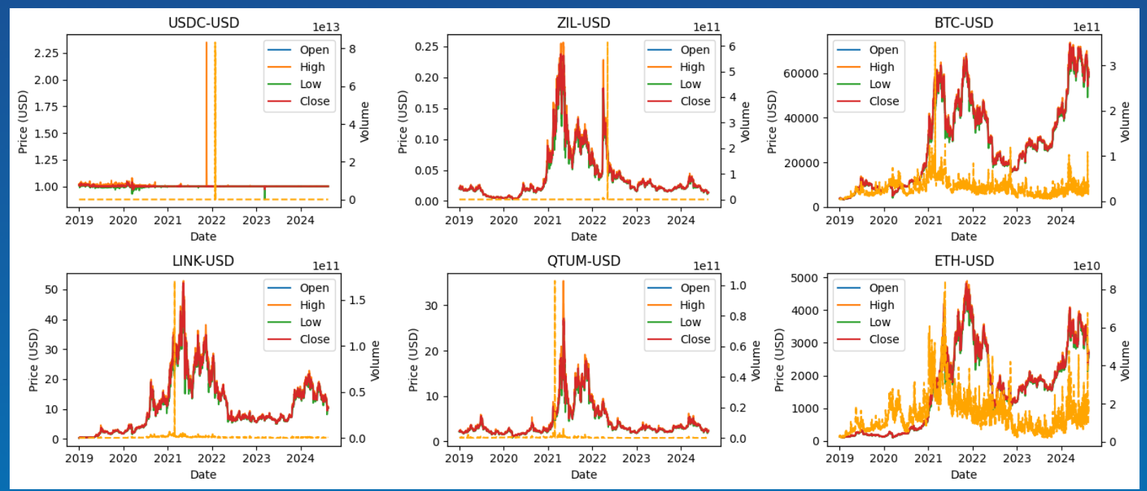
APPENDIX





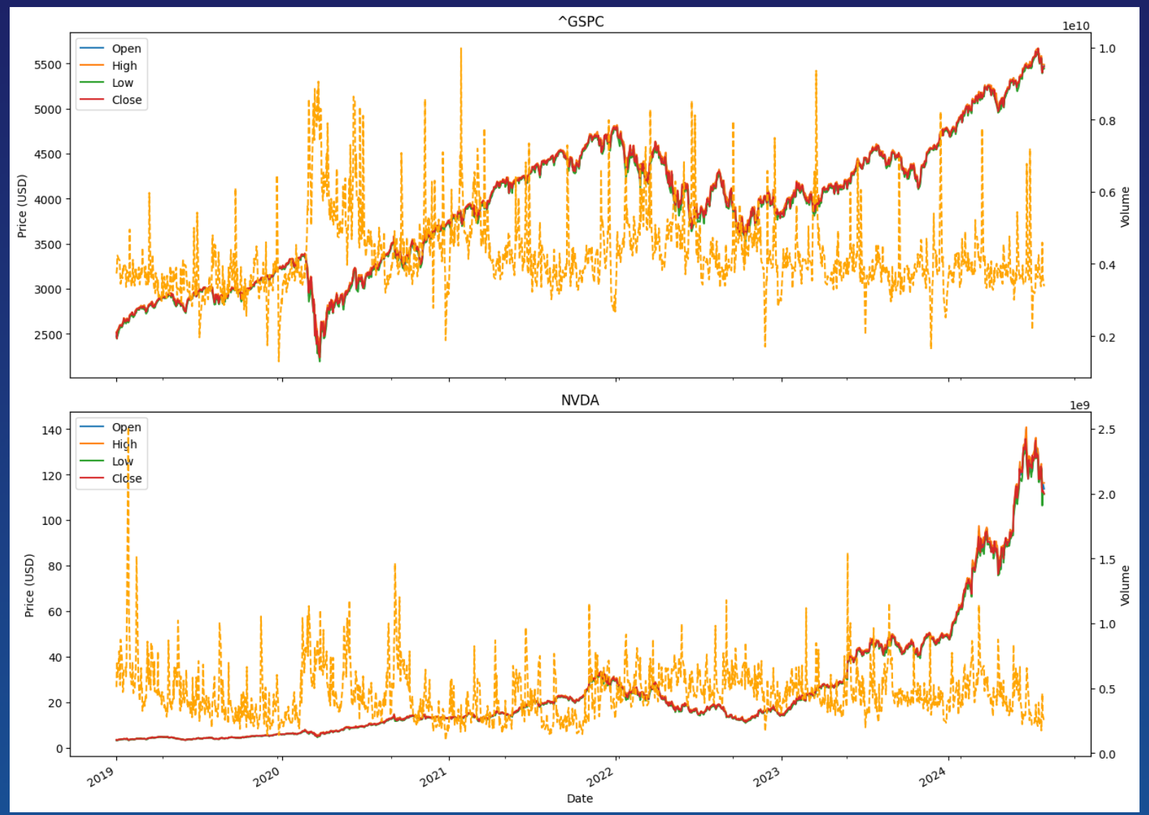


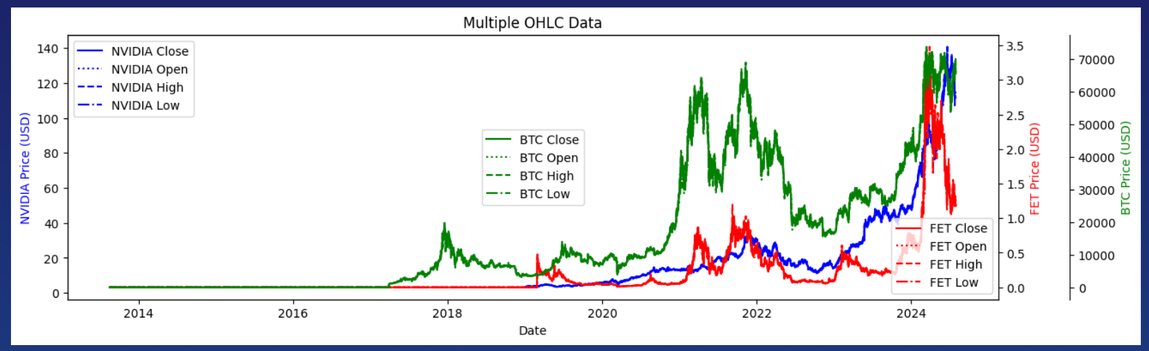




[Discrepancy Between yf.download & yf.Ticker().history() #1036]

APPENDIX


^GSPC
Open
High
Low
Close
Price (USD)
Volume
1e10
NVDA
1e9
Date


Multiple OHLC Data
NVIDIA Close
NVIDIA Open
NVIDIA High
NVIDIA Low
BTC Close
BTC Open
BTC High
BTC Low
FET Close
FET Open
FET High
FET Low
NVIDIA Price (USD)
FET Price (USD)
BTC Price (USD)
Date

NVIDIA vs BOJ: OHLC Plot
NVIDIA Close
NVIDIA Open
NVIDIA High
NVIDIA Low
BOJ Close
BOJ Open
BOJ High
BOJ Low
NVIDIA Price (USD)
BOJ Price (USD)
Date