Universiteit van Amsterdam

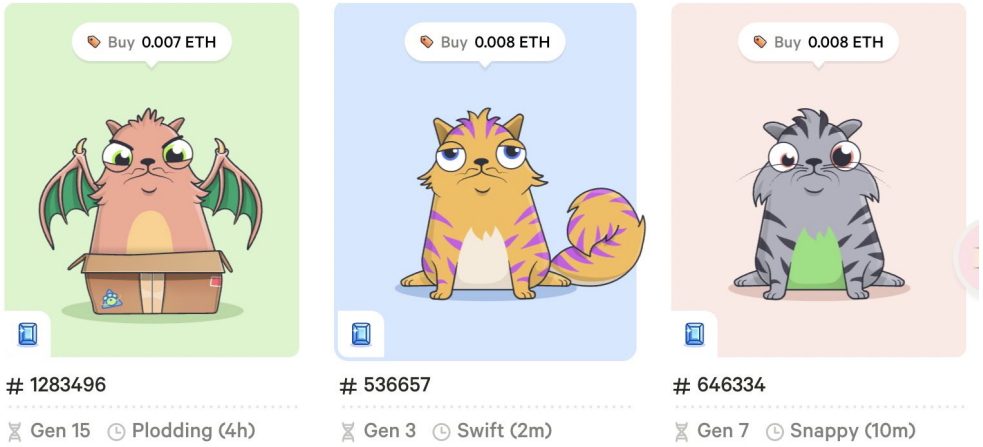
VU VRIJE UNIVERSITEIT AMSTERDAM

Unraveling the NFT chaos using data

Smruti Inamdar
Supervisor : Dr, (Z)Zhiming Zhao

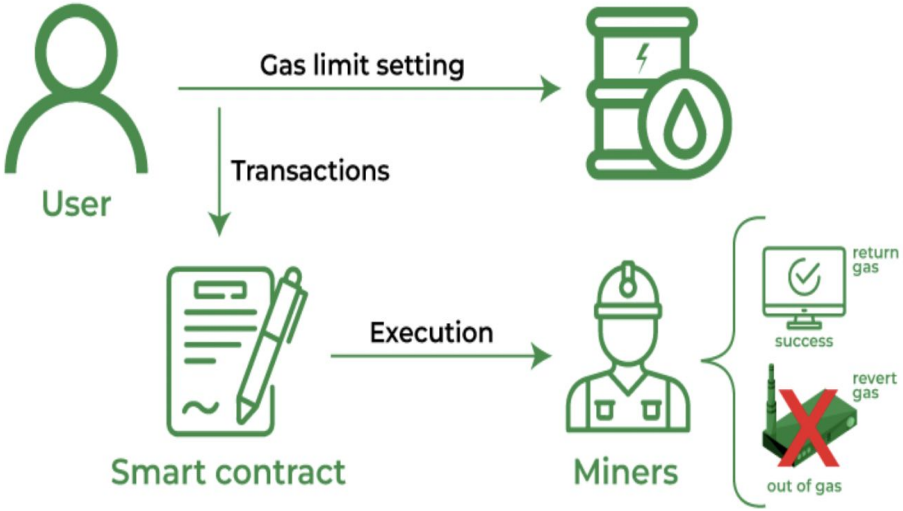
What are NFTs and why should we care?

- Breakthrough innovation in the Blockchain space.
- Garnered popularity after the debut of CryptoKitties in 2017.
- Use cases include Gaming, Art, Photography ,etc.
- Creators receive royalties for every sale.
- Authenticity verified, ownership
- Reached extreme popularity in 2021.
- NFT Bubble burst in 2022.
- Critical to establish empirical understanding of market dynamics.



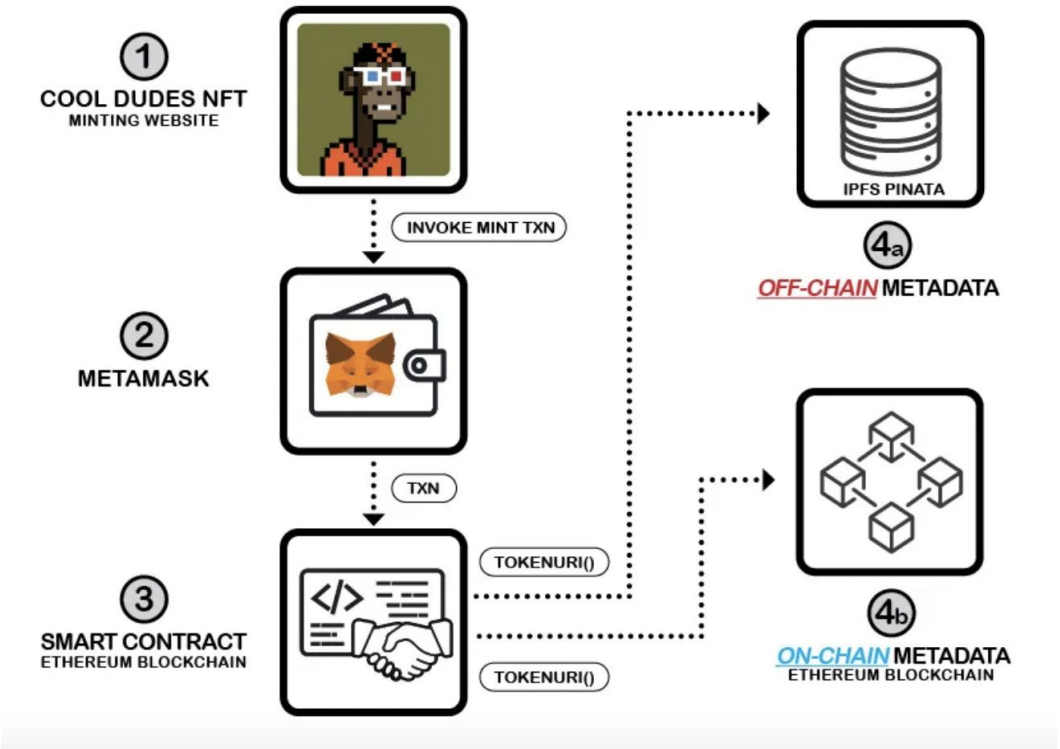
Introduction to Blockchain

- Public ledger comprised blocks,
- Each block is a set of transactions.
- They are known for decentralisation, immutability and consensus.
- Bitcoin is a peer-to-peer electronic cash system, which is a type of digital or virtual currency.
- Ethereum is a blockchain whose native currency is Ether.
- Ethereum is a distributed state machine which uses Proof of Work consensus algorithm.
- Gas fee (gwei) is a transaction fee, which measures computational resources for conducting operations.



Introduction to NFTs

- Cryptographic token used to verify ownership of goods such as digital art, photography, and gaming merchandise.
- NFTs are generated, exchanged, and stored on the Ethereum blockchain.
- Blockchain stores a link to the asset, not the actual digital asset.
- Smart contracts allow the creation of unique token.
- Pure assets and are unique in nature.



Research Questions

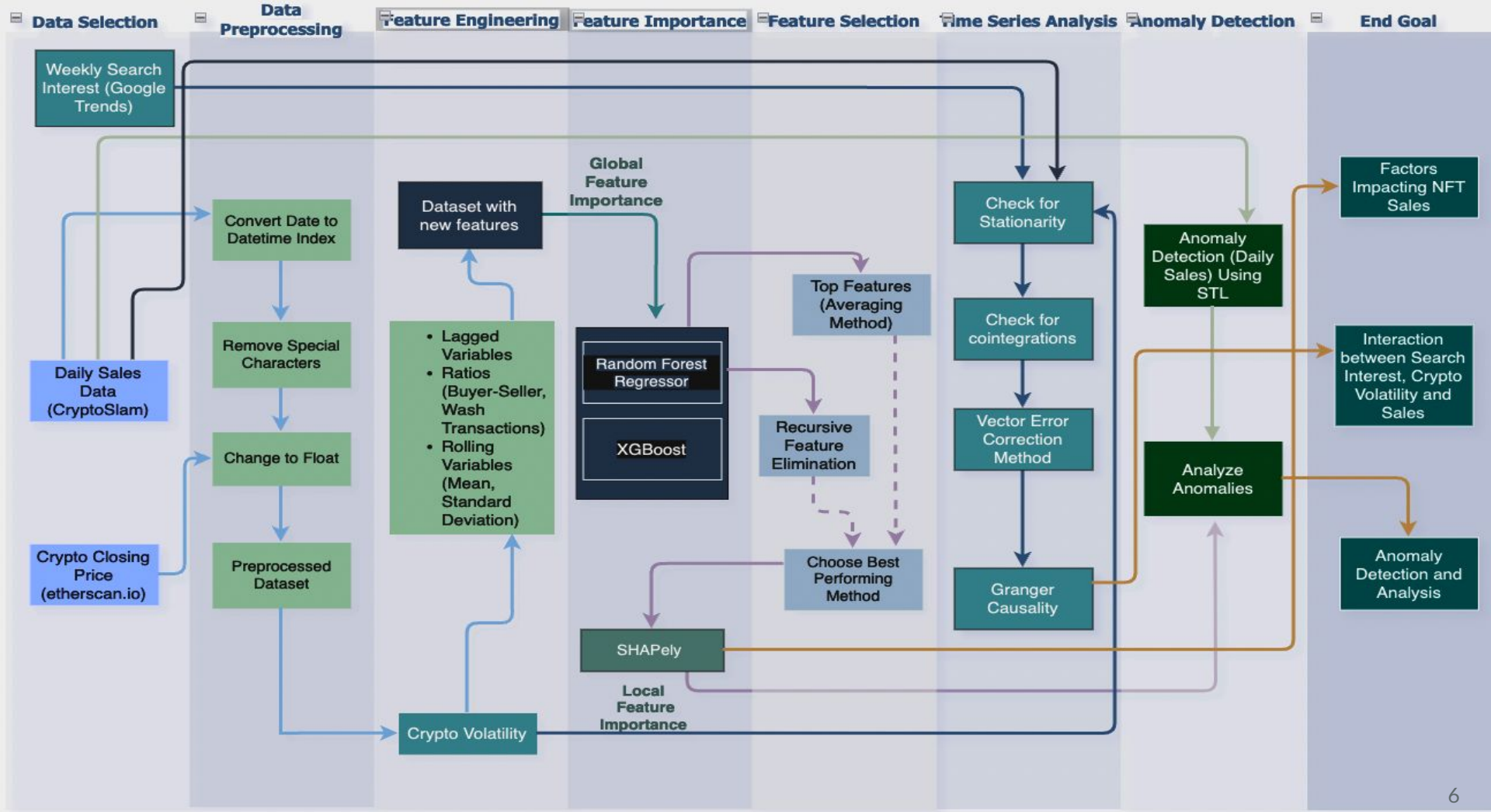
RQ : How do external market forces and overall public opinion together shape the trends and patterns in NFT sales?

RQ1 What are the most important factors driving NFT sales?

RQ2 How to identify factors impacting sales across different time windows?

RQ3 How do market sentiment, crypto volatility and NFT market (Sales) interact with each other?

RQ4 To what extent could anomalies in daily NFT sales be attributed to specific internal or external factors?


Data Selection
Data Preprocessing
Feature Engineering
Feature Importance
Feature Selection
Time Series Analysis
Anomaly Detection
End Goal
Weekly Search Interest (Google Trends)
Daily Sales Data (CryptoSlam)
Crypto Closing Price (etherscan.io)
Convert Date to Datetime Index
Remove Special Characters
Change to Float
Preprocessed Dataset
Crypto Volatility
Dataset with new features
• Lagged Variables
• Ratios (Buyer-Seller, Wash Transactions)
• Rolling Variables (Mean, Standard Deviation)
Global Feature Importance
Random Forest Regressor
XGBoost
Top Features (Averaging Method)
Recursive Feature Elimination
Choose Best Performing Method
SHAPely
Local Feature Importance
Check for Stationarity
Check for cointegrations
Vector Error Correction Method
Granger Causality
Anomaly Detection (Daily Sales) Using STL
Analyze Anomalies
Factors Impacting NFT Sales
Interaction between Search Interest, Crypto Volatility and Sales
Anomaly Detection and Analysis

RQ1: What are the most important factors driving NFT sales?

- Feature importance techniques elucidate which variables predominantly steer sales.
- Regression task as Daily Sales is a continuous variable.
- Two ensemble methods - Random Forest and XGBoost are used.
- Both are supervised machine learning methods.

Feature Engineering

- Process of selecting, transforming, creating, or modifying features from raw data.
- Historical data : Sales lag, ETH vol lag, BTC vol lag,etc.
- Change percentage variables,
- Ratios : Buyer Seller Ratio, Wash Transaction ratio
- Averages : Sales avg, Trade Profit avg.

Volatility

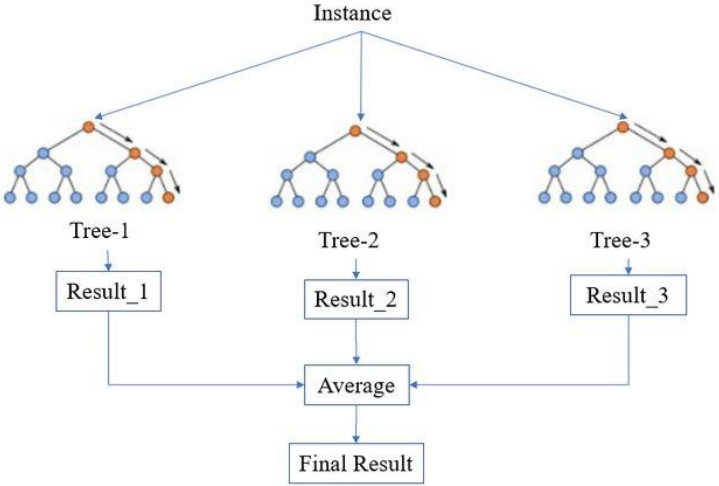
- Statistical gauge of the degree of variation in the asset value over time.
- Sales volatility
- Crypto volatility : BTC and ETH.
- Returns : Gain or loss of investment.

$$\rho_i = \log \frac{p_i^c}{p_{i-1}^c}, \qquad v = \sigma(\rho_T)\sqrt{T},$$

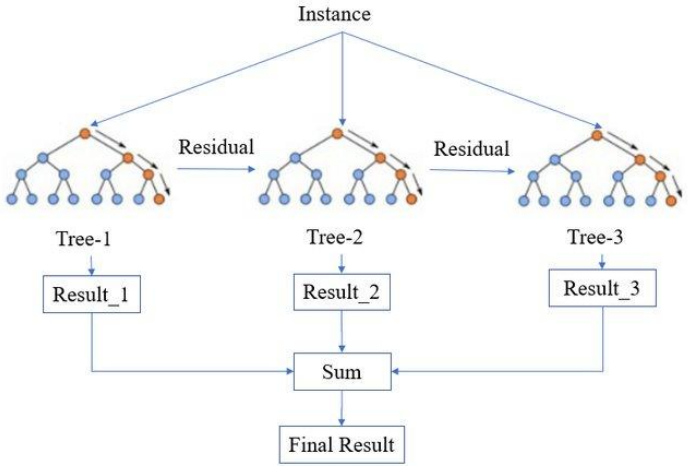
- Volatility of other variables ; Trade Profit, Gas Price.etc.

Random Forest vs XGBoost

- Bagging technique constructs multiple decision trees during training time .
- Prediction is the mean of the features across all the trees.
- Each tree in the forest is unique due to the forest's "random" nature and random selection of features for each split.



- Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library.
- "Boosting" is done by improving a single weak model by combining it with a number of other weak models in order to generate a collectively strong model.
- XGBoost utilizes gradient boosting to optimize a differentiable loss function,



Feature Selection

- After creating new features, it is important to select the best features for predictions.
- **Average Method :** Find avg importances of Random Forest and XGBoost.
- **Recursive Feature Elimination** : fits a model and recursively removes the weakest features.

Average Method	RFE
Sales avg	**Sales avg**
Sales lag 1	**Sales lag 1**
Sales volatility	**Sales volatility**
Trade Profit volatility	**Trade Profit volatility**
Supply	**Supply**
Trade Profit	**Trade Profit**
Total bitcoin vol	**Total eth vol**
Wash trade %	**Wash trade %**
Buyer Seller lag 1	**Gas price change %**
Trade Profit avg	**ETH Price**

Features obtained using Feature Selection

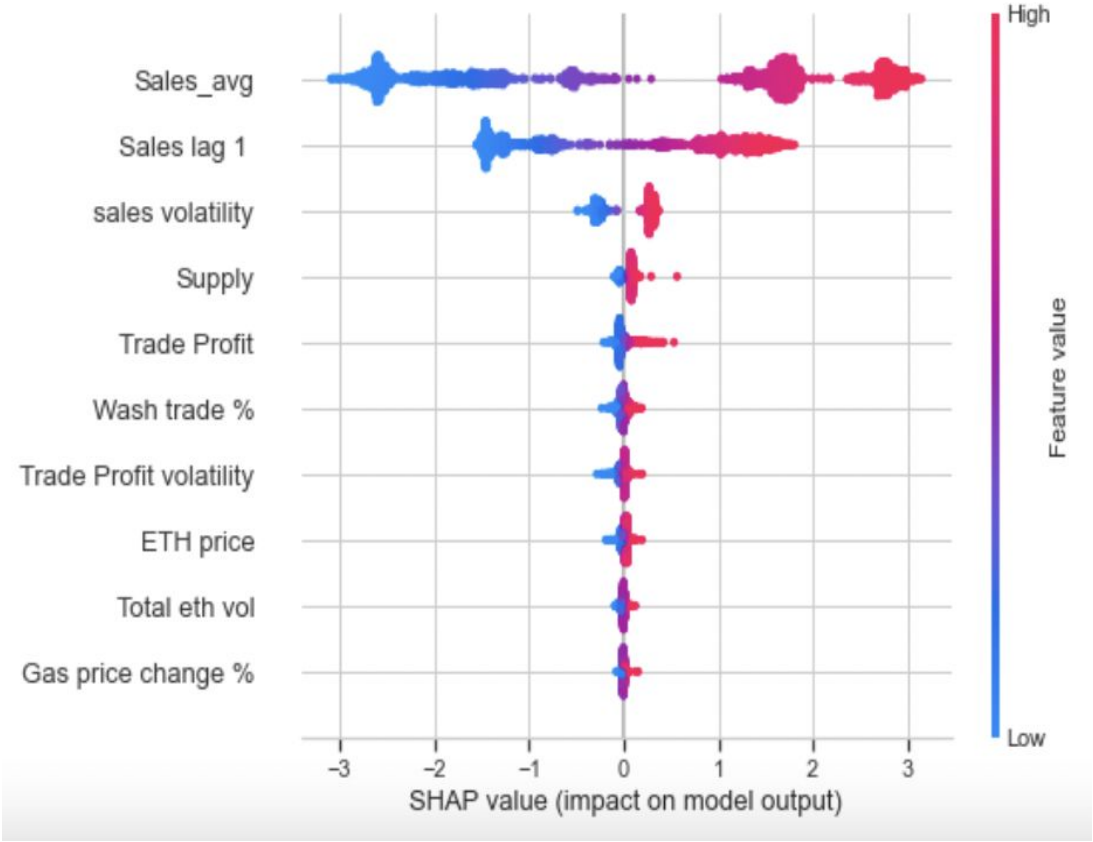
Evaluation

- Experiments are conducted for different set of features.
- Major metrics are MAPE score (error%) and R2 score (Test score), which determines the quality of fit.
- Random Forest regressor with features from RFE give the best performance.
- This model is used for local feature importance analysis.

Model 1	Random Forest			
	Minimal FE	All Features	Feature Avgs	RFE
R2 Score Train	0.9833	0.99	0.99	0.99
R2 Score Test	-1.00	-0.44	0.47	0.515
MAPE	0.30	0.27	0.164	0.157
Model 2	XGBoost			
	Minimal FE	All Features	Feature Avgs	RFE
R2 Score Train	0.99	0.9996	0.99	0.99
R2 Score Test	-0.87	-0.11	0.25	0.285
MAPE	0.315	0.229	0.19	0.185

Feature Importance (Global)

- SHAPely explainer is used to explain the Random Forest model with features from RFE.
- Intrinsic sales metrics, namely Sales avg, Sales lag1, and Sales volatility, emerge as critical determinants.
- Complementing these sales metrics are other significant features like Supply, trade profit, Wash trade %, trade profit volatility, ETH price, Total ETH volume, and gas change%.



RQ2 : How to identify factors impacting sales across different time windows?

- SHAPely's local feature importance captures relevance of features at various temporal granularities.
- SHAP explainer is coupled with model with the best performance,i.e Random Forest and RFE.
- Multiple time intervals like monthly, quarterly, biannually and yearly are considered.
- By resampling and aggregating SHAP values throughout multiple durations, a thorough image of feature significance over time is captured.

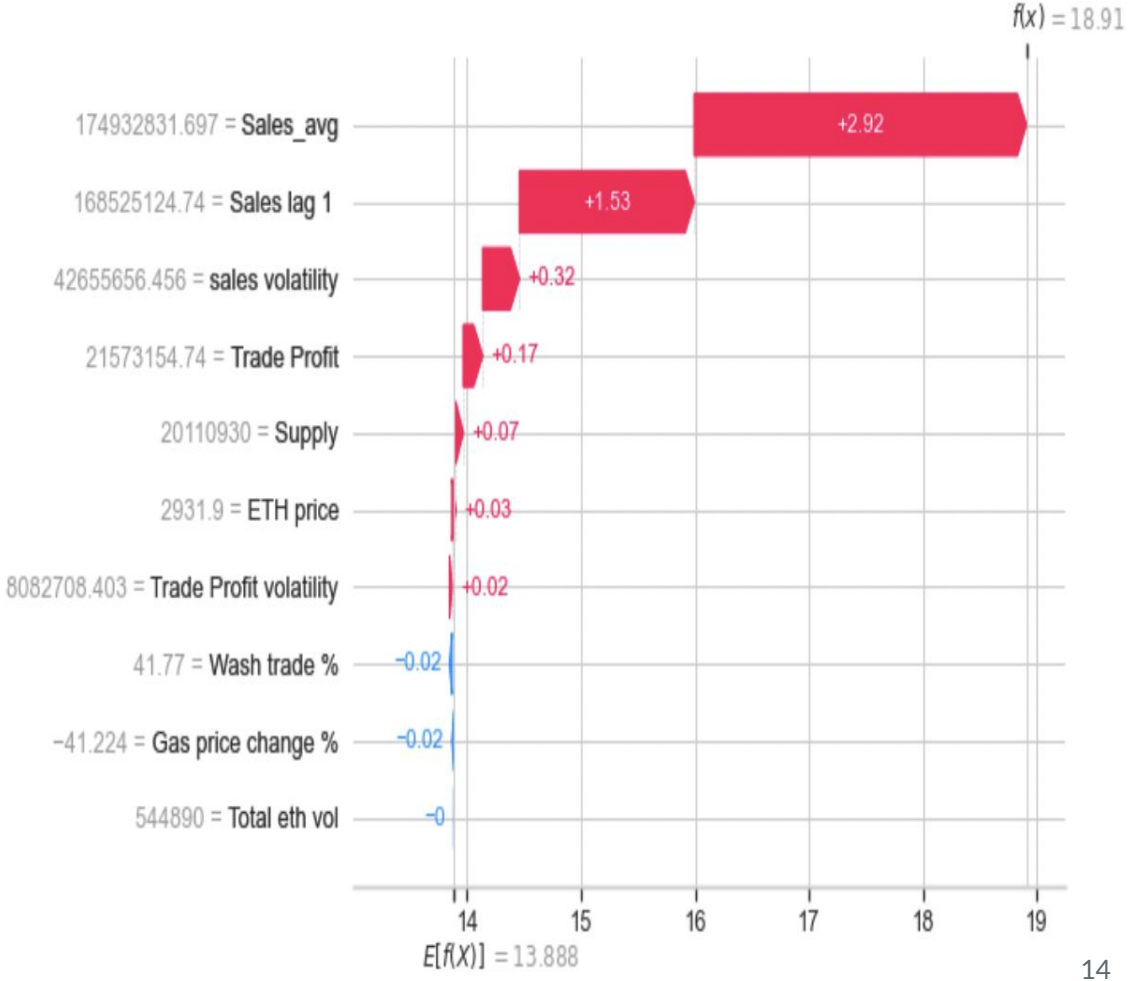
Local Feature Importance (SHAPely)

-

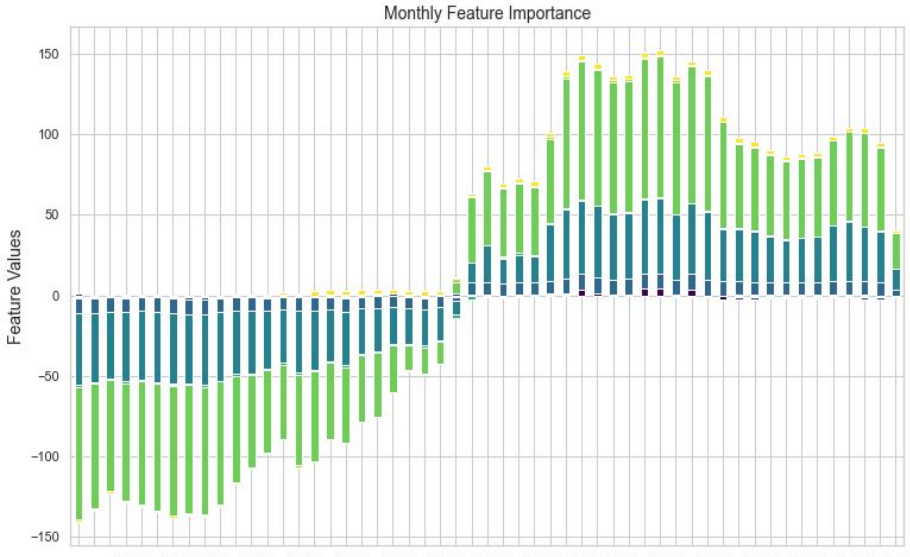
- Local feature importance refers to the individual contribution of each feature to a particular prediction.

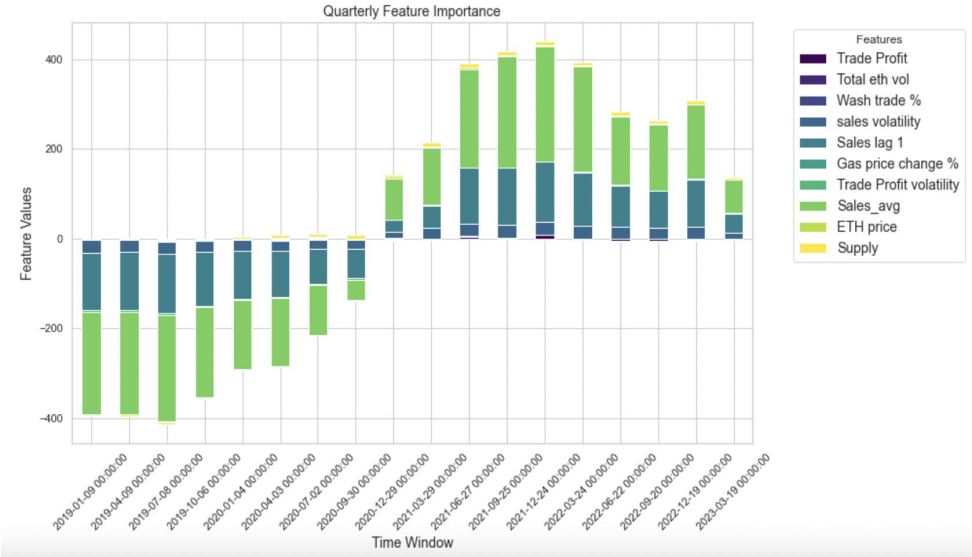
- Random Forest and XGBoost are not interpretable.

- For a given feature and an input data point, SHAP assigns a feature importance value, thus aiding model explainability

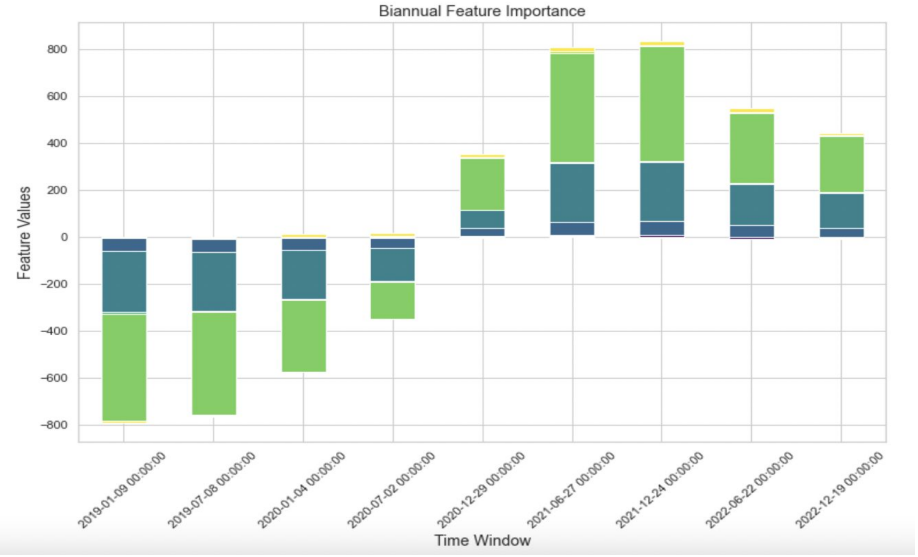


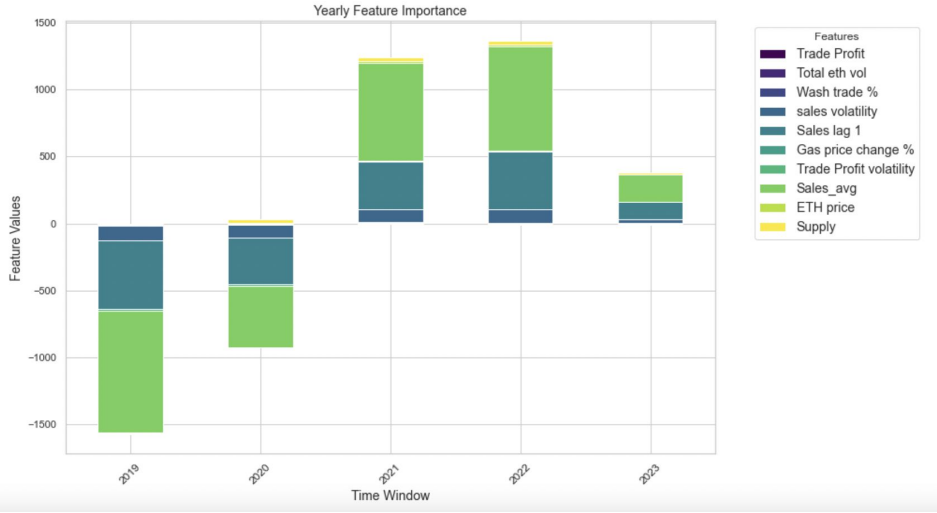
Monthly and Quarterly Feature Importance





Bi-Annual and Annual Feature Importance



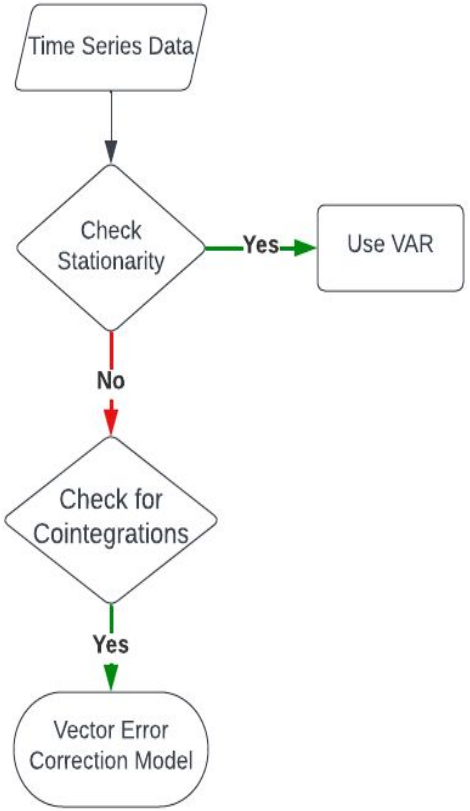


RQ3 : How do market sentiment, crypto volatility and NFT market (Sales) interact with each other?

- Data collected from Google Trends. Weekly sales avg and crypto volatility are computed.
- Google trends keywords are grouped into Positive, Negative, Neutral - based on sentiment. Basic keywords - Crypto, Ethereum, Bitcoin

Model Selection for Time Series Data

- Vector Autoregression (VAR) and Vector Error Correction Model (VECM) multivariate time series models.
- VAR models have limitations, such as their sensitivity to the choice of lag order and the requirement of the stationarity assumption.
- Vector Error Correction Model (VECM) is used when the time series is non-stationary and cointegrations are present.
- The identification of cointegrations is carried out using the Johansen Cointegration test .
- Assume 0 < r < p, where r denotes the number of cointegrating relationships and p is the total number of variables



01	**Augmented Dickey-Fuller Test**	H0 : Time series in not stationary. H1: Time series is stationary.
02	**Johansen Cointegration Rank (Trace Test)**	H0 : Number of cointegrating vectors is r. H1: Number of cointegrating vectors is greater than r.
03	**Johansen Cointegration Rank (Max eigenvalue Test)**	H0 : Number of cointegrating vectors is r. H1: Number of cointegrating vectors is r+1.
04	**Durbin Watson Test**	H0 : There is no correlation among the residuals. H1: The residuals are autocorrelated.
05	**Granger Causality**	H0 : Lagged X does not Granger-cause Y. H1: Lagged X Granger-causes Y.

Lag	AIC	BIC	FPE	HQIC
0	66.71	67.57*	9.373e+28	67.06*
1	66.61	68.23	8.502e+28	67.26
2	66.41	68.77	6.949e+28	67.36
3	66.39	69.52	6.903e+28	67.65
4	66.31*	70.18	6.393e+28*	67.87

Table 1: VECM Order Selection (* highlights the minimums)

```
Sales (USD) : 2.06
Neutral Phrases : 2.06
Basic : 2.05
Negative : 2.04
Positive : 2.08
Eth Volatility : 2.0
Btc Volatility : 1.92
```

Table 4: Durbin-Watson Test Statistic

r_0	r_1 test	Test Statistic	Critical Value
0	7	395.3	150.1
1	7	276.7	117.0
2	7	177.2	87.77
3	7	100.7	62.52
4	7	39.26	41.08

Table 2: Trace test Statistics

r_0	r_1 test	Test Statistic	Critical Value
0	1	118.6	55.82
1	2	99.48	49.41
2	3	76.51	42.86
3	4	61.44	36.19
4	5	21.64	29.26

Table 3: Max Eigenvalue test Statistics

Vector Error Correction Model

- Vector Error Correction Model is a cointegrated VAR model.
- VAR is a type of autoregressive model that models each variable as a function of its past values, with the predictors being lags of the series.
- The resulting VAR from VECM representation has more efficient coefficient estimates.

Granger Causality

- Statistical hypothesis test for determining whether one time series is useful in forecasting another
- "Causality" in this context is a misnomer.
- VECM for Granger causality tests accounts for both the short-term dynamics and the long-term relationship
- The p-value less than assigned threshold (0.05) indicates that lagged X Granger-causes Y.

VECM - Granger Causality Results

- BTC Volatility is not Granger-caused by ETH Volatility.
- ETH Volatility is not Granger-caused Negative sentiment. ,
- Sales (USD) and ETH Volatility are Granger-caused by Positive sentiment.
- Sales (USD) Granger-causes crypto volatility.
- Volatility in Bitcoin could lead to Ethereum volatility, not vice-versa.

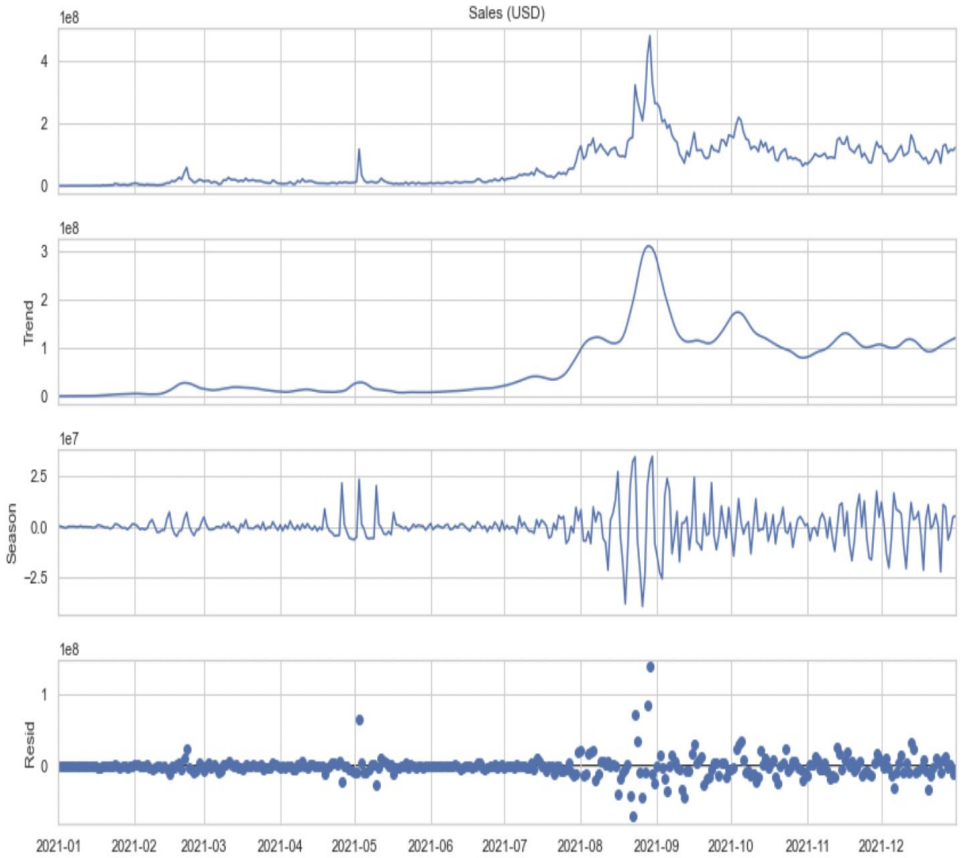
	Sales (USD)	Neutral Phrases	Basic	Negative	Positive	ETH Volatility	BTC Volatility
Sales (USD)	1	0.46	0.47	0.86	0	0.88	0.3
Neutral Phrases	0.45	1	0.19	0.96	0	0.42	0.52
Basic	0.28	0.41	1	0.12	0.22	0.19	0.28
Negative	0.16	0.06	0.23	1	0.04	0.28	0.16
Positive	0.02	0.01	0.25	0.19	1	0.07	0.34
Eth Volatility	0	0.04	0	0.28	0	1	0.02
Btc Volatility	0	0.02	0	0.13	0.17	0.03	1

RQ4: To what extent could anomalies in daily NFT sales be attributed to specific internal or external factors?

- Anomaly is an observation or a sequence of observations that deviates from the overall distribution of data.
  - Focus on contextual anomalies.
  - Sales is treated as univariate data.

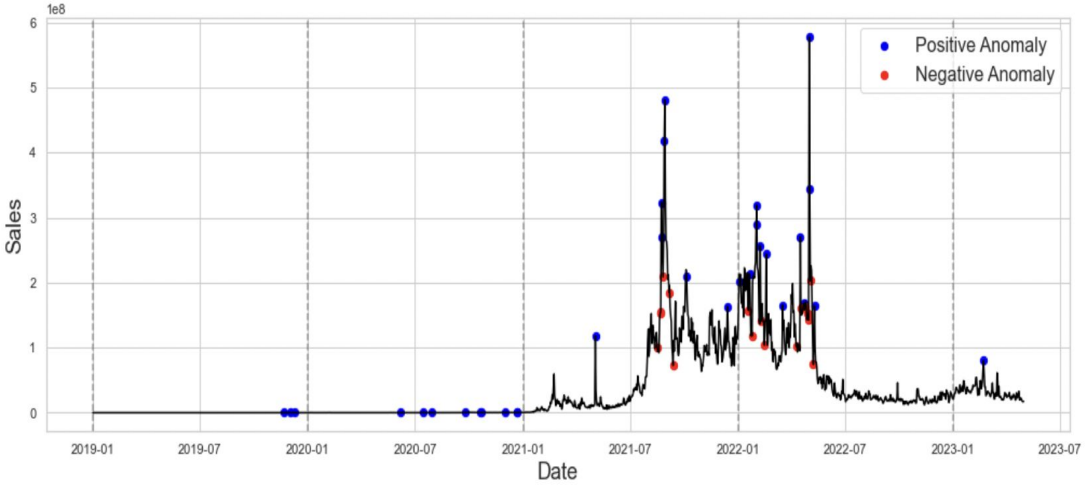
STL Decomposition

- Seasonal-Trend decomposition using LOESS (STL) decomposes a seasonal time-series into seasonal, trend, and residual components.
- Sales (USD) is treated as univariate data.
- The 'noise' or irregular movements that cannot be attributed to the trend or seasonality is captured by the residuals.
- Residuals are a good indicator to find outliers.
- Time series is broken into yearly data.



Finding Anomalies

- Residuals do not follow a normal distribution, hence Standard Deviation cannot be used to study variability.
- Mean Absolute Deviation requires symmetry in data.
- The constant 1.1926, while derived from the normal distribution, serves as a general-purpose scaling factor for different distributions.
- Positive and negative anomalies are spikes and dips in sales respectively.



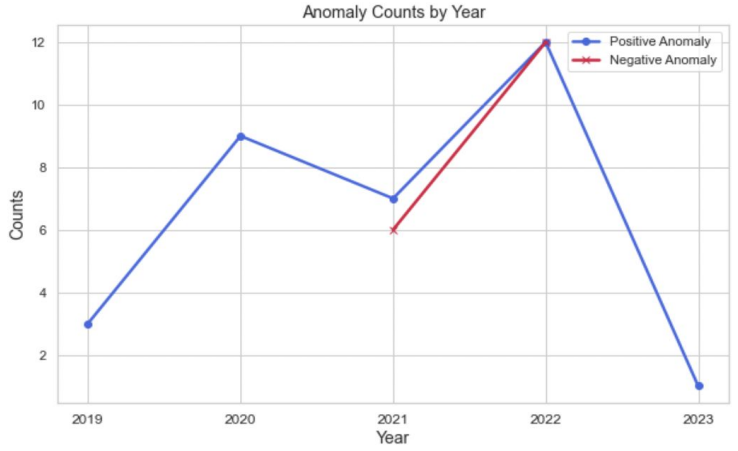
$$S_n = c \times \text{Median}_i \left\{ \text{Median}_j \left| x_i - x_j \right| \right\}$$

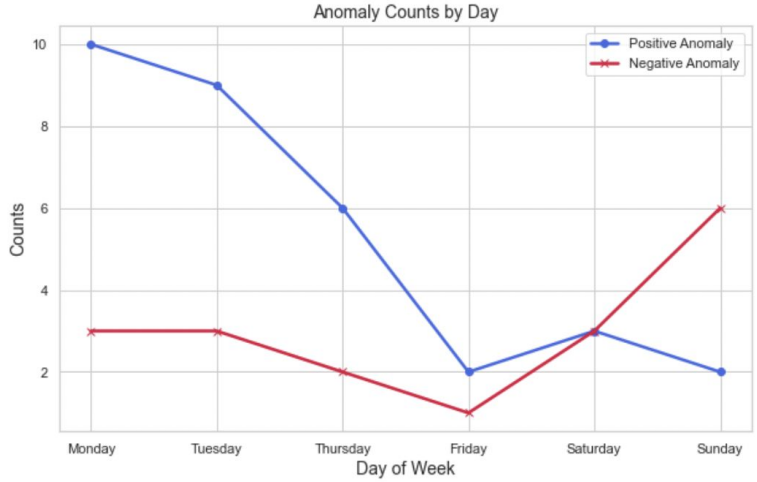
$$\tau_u = \tilde{x} + 6 \times S_n$$

$$\tau_l = \tilde{x} - 6 \times S_n$$

Analysis of Anomalies

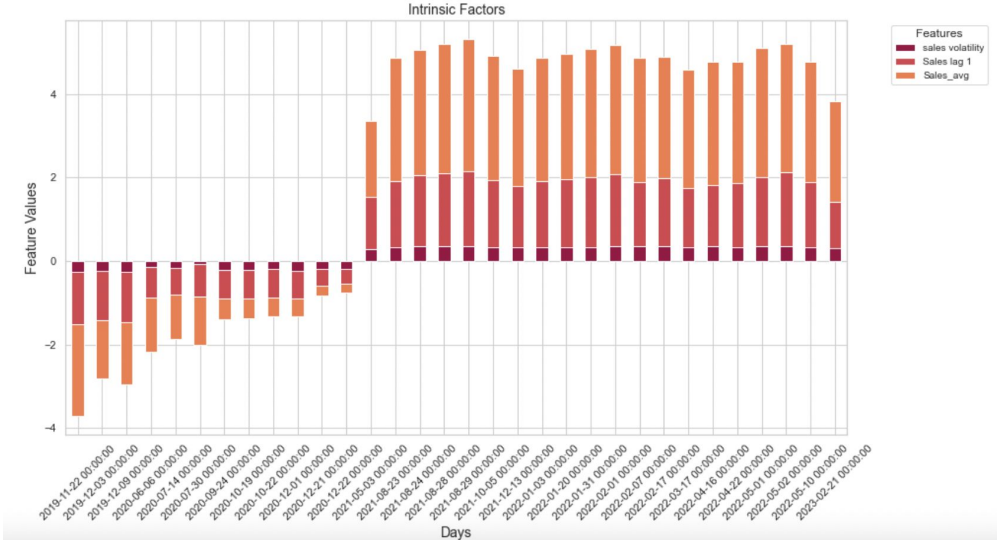
- A total of 50 anomalies are found, with 32 Positive and 18 Negative.
- Absence of anomalies on Wednesday could suggest a mid-week stability.
- The maximum number of positive and negative anomalies occur on Monday and Sunday respectively.
- Negative anomalies present only in 2021 and 2022.
- 2022 marks the peak for both positive and negative anomalies.

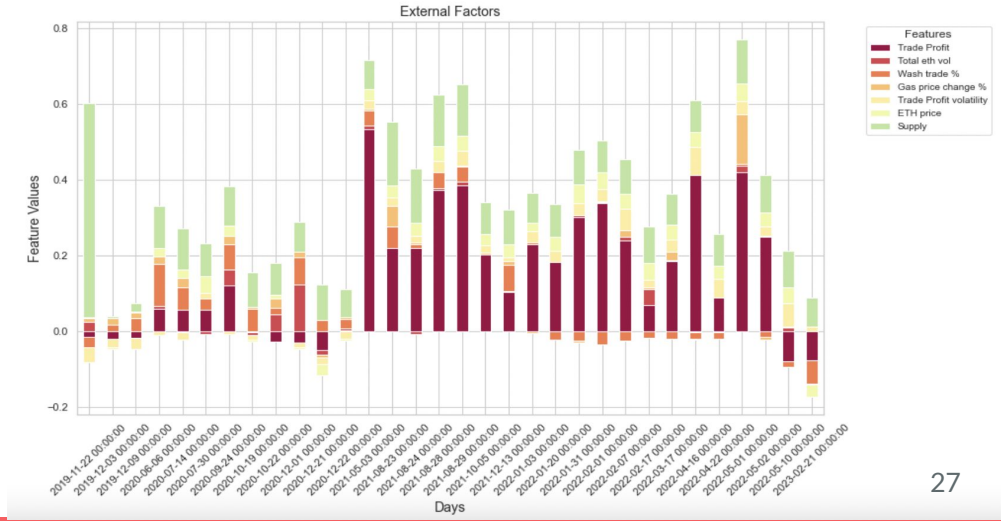




Positive Anomalies

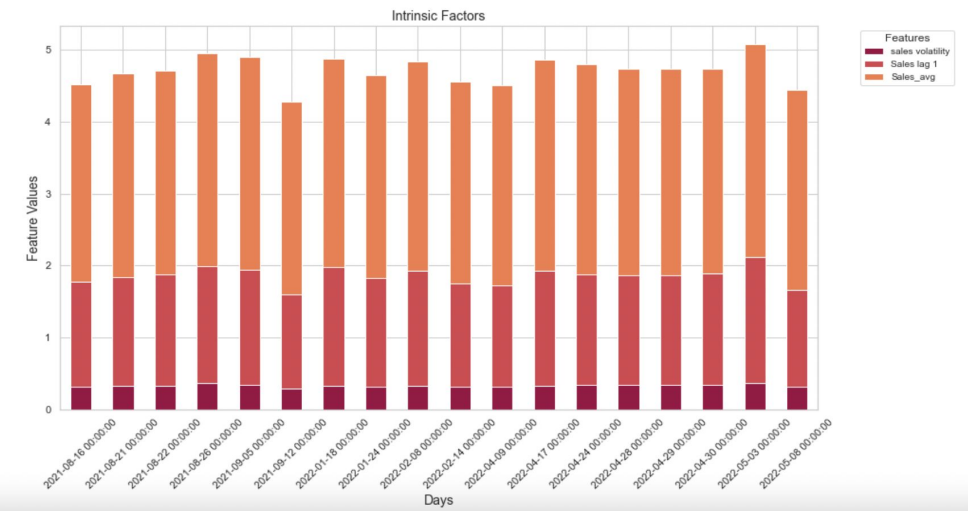
- Intrinsic Sales Average and Sales lag are primary factors impacting positive anomalies, with Sales volatility being a notable factor.
- The first positive anomaly in 2019 sees a high positive impact by Supply. Supply is consistent with its positive impact.
- External factors like Trade profit, Total ETH volume and Gas price change % have a high positive impact 2021 onwards.
- ETH price has a positive impact only in 2021 and 2022.
- Wash trade % has a negative impact 2022 onwards.

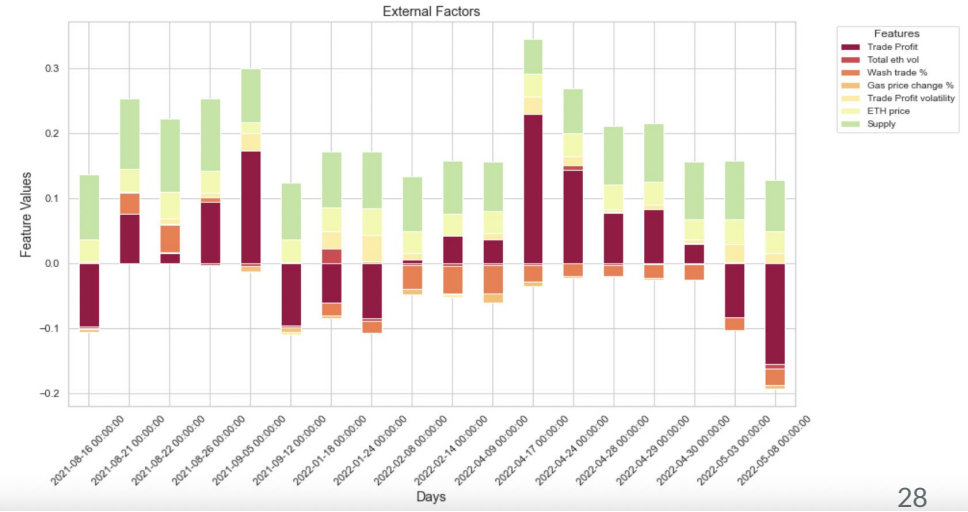




Negative Anomalies

- Intrinsic Sales Average and Sales lag are primary factors leading to negative anomalies, with Sales volatility being a notable factor.
- All negative anomalies occur between 2021 and 2022.
- External variables Trade profit and Wash Trade% have both positive and negative impacts on Sales.
- ETH Price and Supply consistently have positive impact.
- Gas price change% has a slight negative impact.





Important Findings

1. Random Forest and Recursive Feature Elimination give features and model with best R2 and MAPE scores.
2. Sales average (weekly), lagged sales (1-day) and sales volatility (weekly) are the most important features to determine sales.
3. They are followed by 'Supply', 'Trade Profit' and 'Wash Trade%'. Similar pattern is seen with different time periods.
4. Sales (USD) and ETH Volatility is Granger-caused by Positive sentiment. Sales (USD) Granger-causes crypto volatility. Volatility in Bitcoin could lead to Ethereum volatility, not vice-versa.
5. 50 anomalies are detected - 32 positive and 18 negative.
6. The most important features are Sales average and Sales lag. Sales Volatility, Trade Profit and Supply are third most important features.
7. For negative anomalies, Sales average and Sales lag and Sales Volatility are the top 3 undisputed features. They are followed by Supply, Trade Profit and Ethereum price.

Limitations and Future Work

1. Absolute Feature Importance is used. Relative feature importance could highlight the nuanced variables.
2. STL is not validated.
3. Variables in the given dataset might not be the most explainable features to predict sales.

Thank You