Predicting Future Votable Supply: Machine Learning Approaches and Results

--By Chain_L and Team

Prediction of Future Votable Supply Using Machine Learning

This report outlines the methodologies and results to predict the future votable supply in the Optimism ecosystem using a Machine Learning (ML) model. We explored various approaches by leveraging the **Random Forest Regressor** and performing hyperparameter tuning using **GridSearchCV**. The features incorporated into the models include past circulating supply, votable supply, airdrop-related data, and derived metrics.

Table Of Content:

Features in the Training Dataset

Below is the general list of features included in the dataset used for training the models. These features represent a variety of metrics and derived variables that help in understanding and predicting the future votable supply:

- Date: The date corresponding to the observations.
- Votable Supply: The total supply of tokens eligible for voting.
- Circulating Supply: The total supply of tokens actively circulating in the market.
- OP_Price: The historical price of the OP token.
- airdrop_tokens: The number of tokens distributed during an airdrop event.
- airdrop_number: The sequence number of the airdrop event (e.g., 1st, 2nd, 3rd).
- airdrop_targets_delegators: Indicator of whether the airdrop targeted specific delegators (1 for Yes, 0 for No).
- Days_Since_Airdrop: The number of days elapsed since the most recent airdrop.
- effective_airdrop_tokens: The total effective tokens distributed during the airdrop, calculated as airdrop_tokens * airdrop_targets_delegators.
- airdrop_effect: The decayed impact of an airdrop over time, calculated using an exponential decay factor based on the half-life.
- airdrop_lag: The number of tokens distributed during the most recent airdrop.
- votable_lag_1: The votable supply value lagged by 1 day.
- circulating_lag_1: The circulating supply value lagged by 1 day.
- votable_lag_7: The votable supply value lagged by 7 days.
- circulating_lag_7: The circulating supply value lagged by 7 days.
- votable_lag_30: The votable supply value lagged by 30 days.
- circulating_lag_30: The circulating supply value lagged by 30 days.
- vs_rolling_mean_7: The 7-day rolling average of the votable supply.
- cs_rolling_mean_7: The 7-day rolling average of the circulating supply.
- vs_rolling_mean_30: The 30-day rolling average of the votable supply.
- cs_rolling_mean_30: The 30-day rolling average of the circulating supply.
- vs_growth_rate_1: The day-over-day percentage growth rate of the votable supply.
- cs_growth_rate_1: The day-over-day percentage growth rate of the circulating supply.
- vs_cs_ratio: The ratio of votable supply to circulating supply, expressed as a percentage.

Note: This is a general list of features used across different approaches. For each specific approach, the exact features utilized are detailed in the Dataset section under the respective approach.

Approach 1: Random Forest Regressor with Circulating Supply and Votable Supply

Objective

To predict future votable supply using a minimal dataset that includes circulating supply, votable supply, and airdrop-related data. This approach tests the model's ability to generate predictions based solely on fundamental supply-related features.

Dataset

- Circulating Supply and its derived features (e.g., rate of change, rolling averages).
- Airdrops data and its derived features (e.g., cumulative airdrops, airdrop ratios).
- Votable Supply as the target variable.

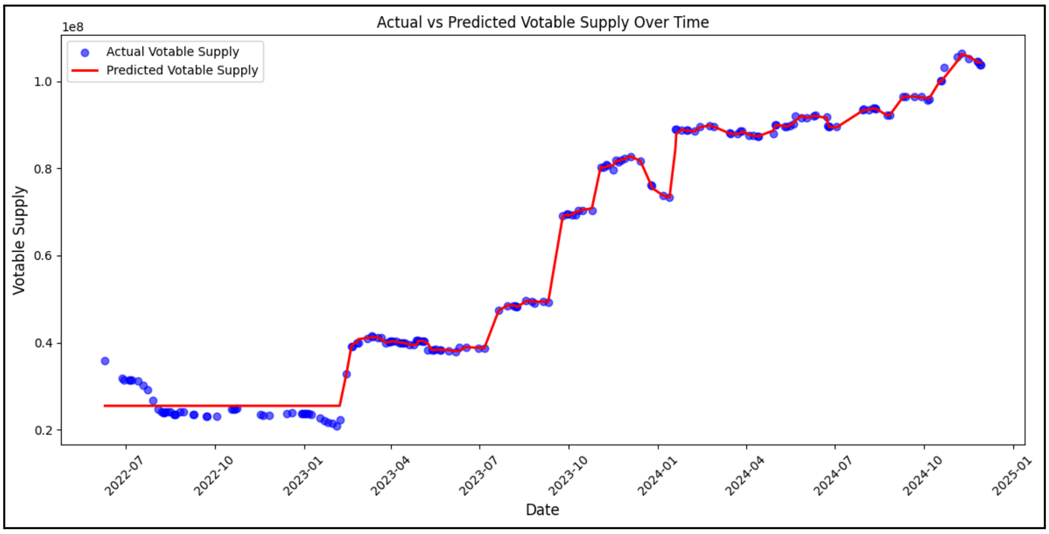
Features Used

['Circulating Supply', 'airdrop_is_active', 'Airdrop_Tokens_Distributed', 'Days_Since_Airdrop', 'Airdrop_Number', 'Airdrop_Target_Delegators', 'CS_pct_change', 'CS_rolling_mean_7', 'CS_lag_1', 'CS_lag_30']

Testing and Results

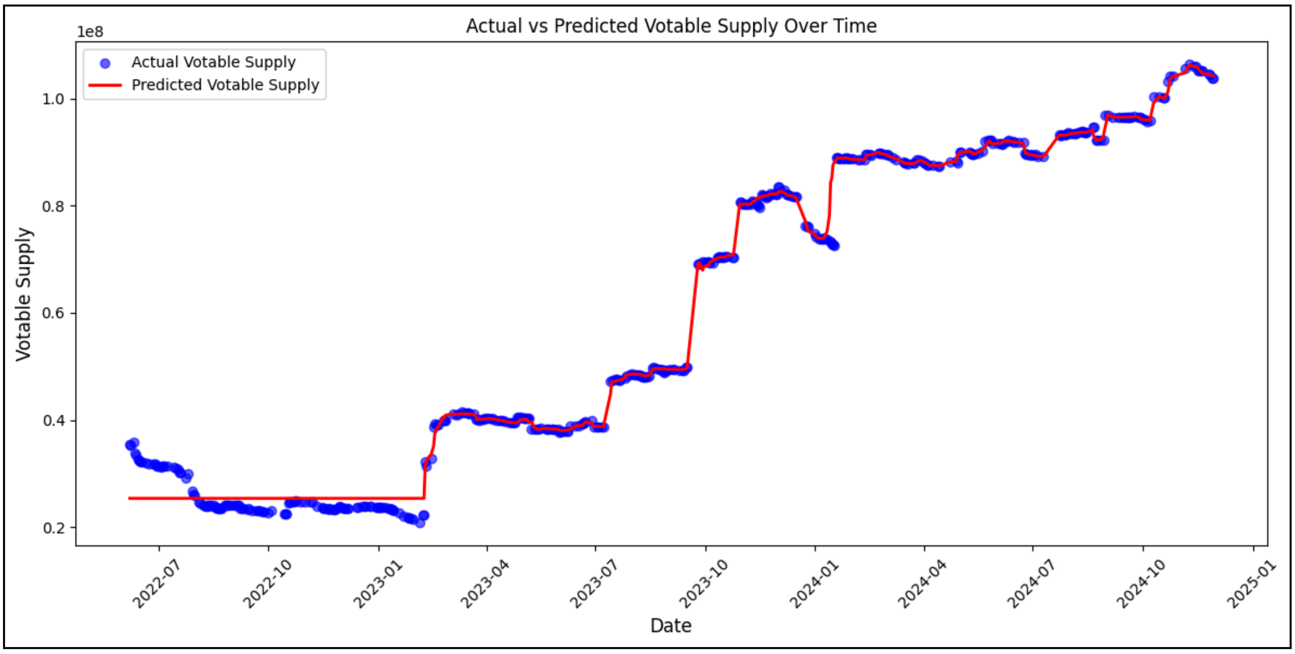
1.1 Random Split: Train 80% - Test 20%

- **Best Parameters**:
  {'max_depth': None, 'min_samples_split': 2, 'n_estimators': 500}
- **Performance Metrics**:
  - **Mean Absolute Error (MAE)**: 900,663.56
  - **Mean Squared Error (MSE)**: 3,157,501,567,739.20
  - **R² Score**: 0.9961
  - **Cross-Validation R² Scores**: [0.9945, 0.9937, 0.9910, 0.9943, 0.9907]
  - **Mean Cross-Validation R²**: 0.9928



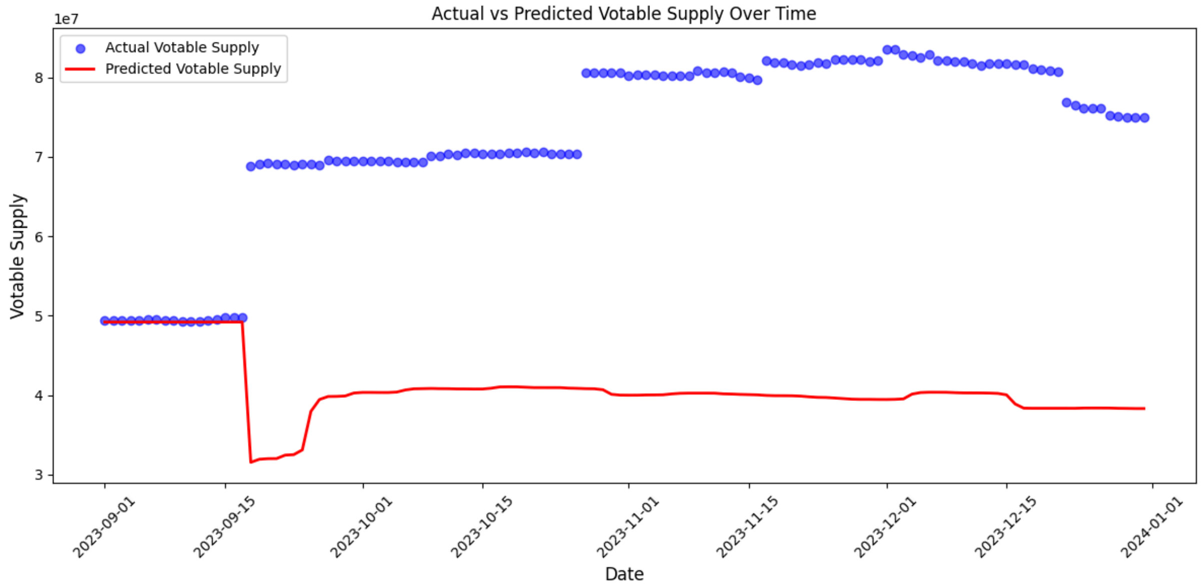
1.2 Random Split: Train 50% - Test 50%

- **Best Parameters**:
  {'max_depth': 10, 'min_samples_split': 2, 'n_estimators': 100}
- **Performance Metrics**:
  - **MAE**: 1,103,467.26
  - **MSE**: 5,401,630,196,690.40
  - **R² Score**: 0.9933
  - **Cross-Validation R² Scores**: [0.9950, 0.9895, 0.9973, 0.9942, 0.9928]
  - **Mean Cross-Validation R²**: 0.9938



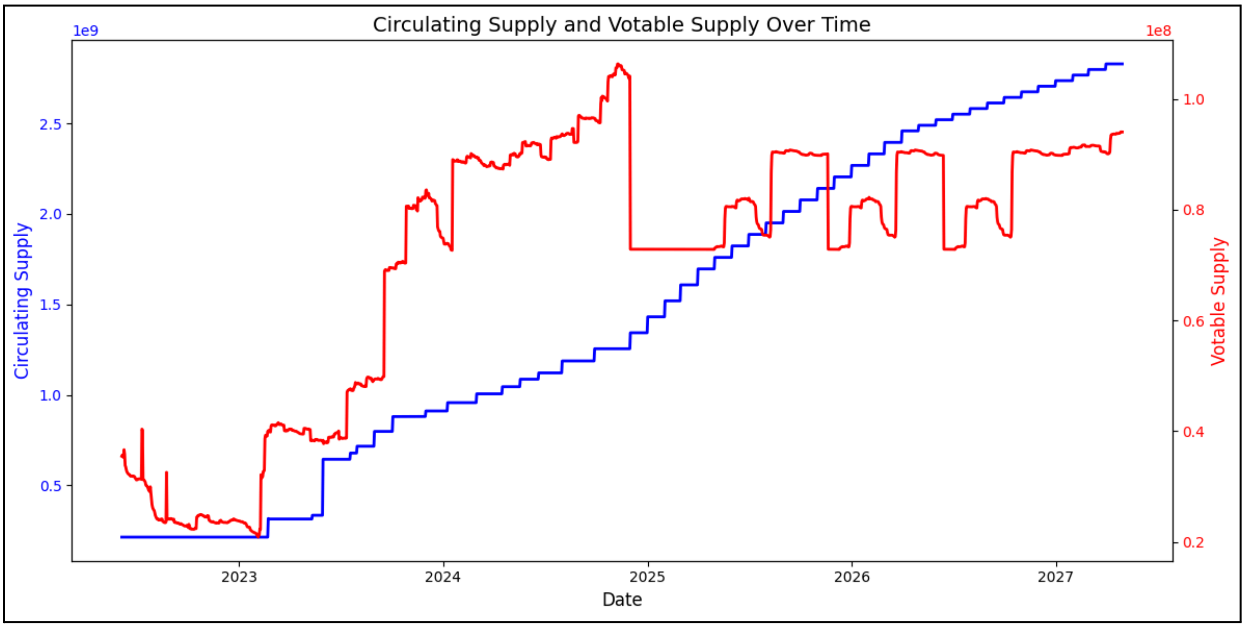
1.3 Train-Test Split Based on Specific Dates

- **Best Parameters**:
  {'max_depth': 10, 'min_samples_split': 5, 'n_estimators': 100}
- **Performance Metrics**:
  - **MAE**: 31,992,434.48
  - **MSE**: 1,211,467,366,051,261.50
  - **R² Score**: -9.65
  - **Cross-Validation R² Scores**: [-1.785, -13.782, 0.526, 0.235, -1.676]
  - **Mean Cross-Validation R²**: -3.296

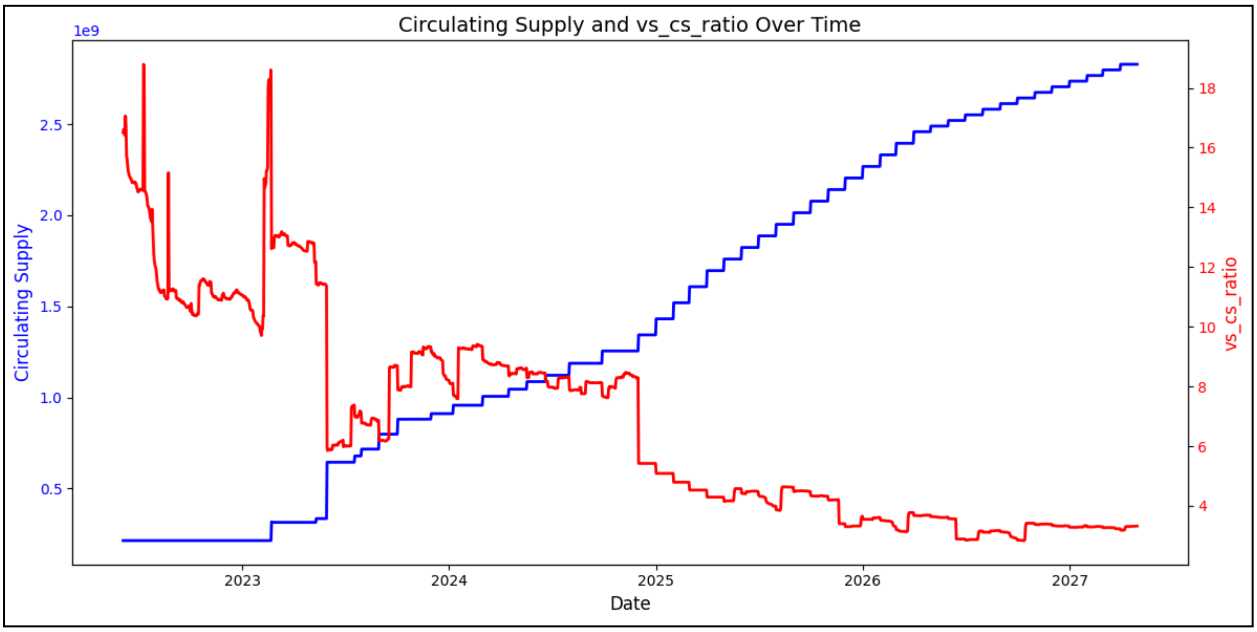


Prediction Results

1. Circulating Supply vs. Votable Supply



2. Circulating Supply vs. VS_CS_Ratio



Approach 2: Random Forest Regressor with OP Price

Objective

To incorporate the OP token price as an additional feature, hypothesizing that market dynamics may have a significant correlation with votable supply trends.

Dataset

- Circulating Supply and its derived features.
- Airdrops data and its derived features.
- OP Price.
- Votable Supply.

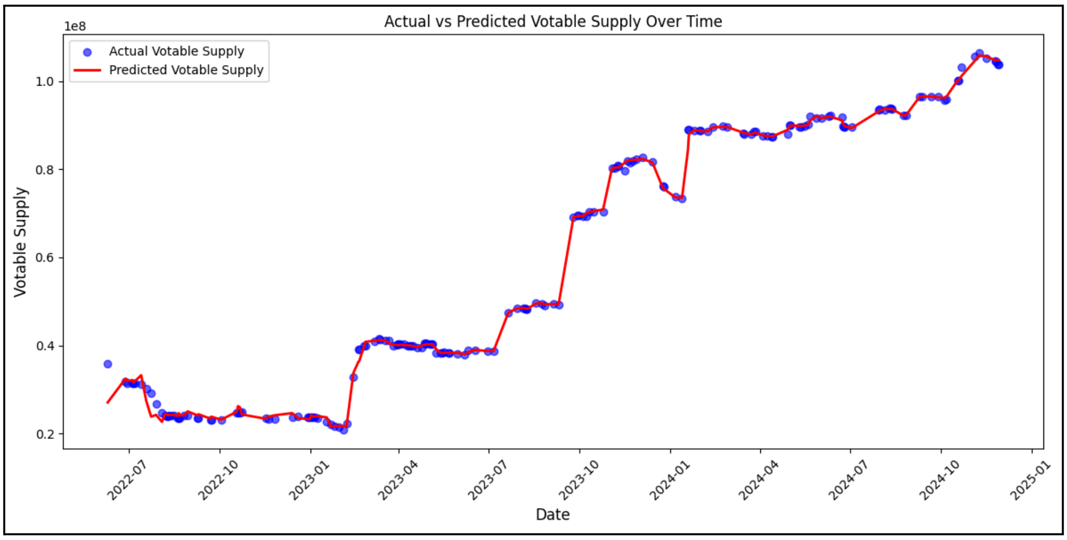
Features Used

['Circulating Supply', 'OP_Price', 'airdrop_tokens', 'airdrop_number', 'airdrop_targets_delegators', 'airdrop_lag', 'circulating_lag_1', 'circulating_lag_7', 'circulating_lag_30', 'cs_rolling_mean_7', 'cs_rolling_mean_30', 'cs_growth_rate_1', 'effective_airdrop_tokens', 'airdrop_effect']

Testing and Results

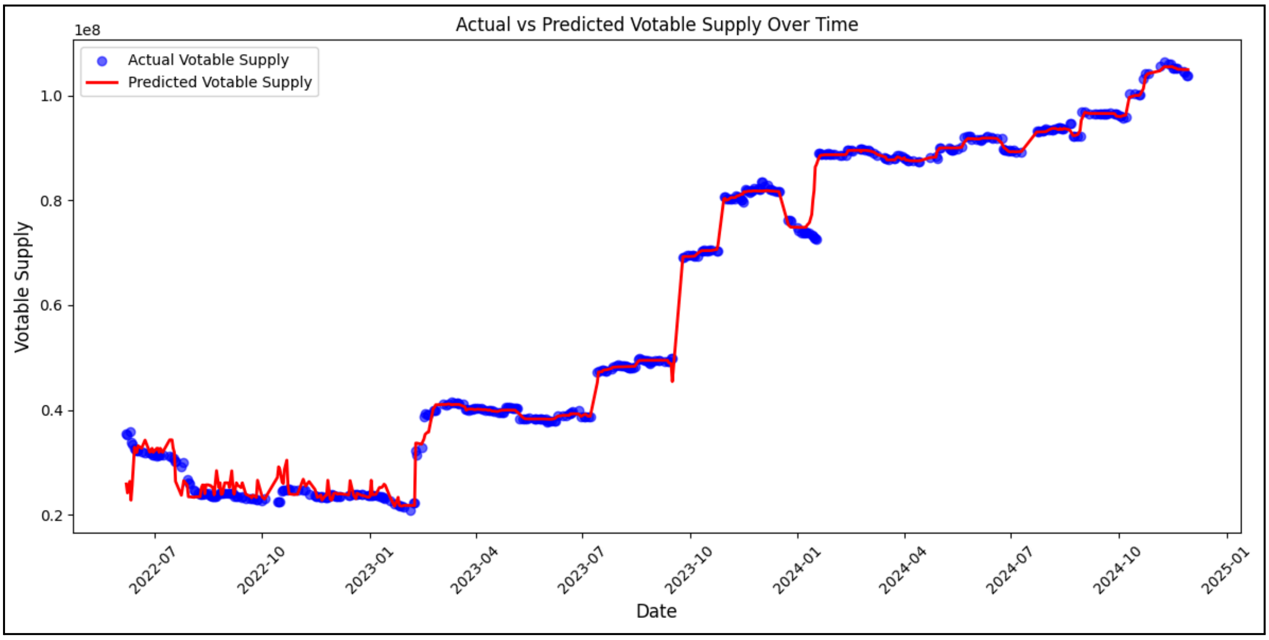
2.1 Random Split: Train 80% - Test 20%

- **Best Parameters**:
  {'max_depth': 10, 'min_samples_split': 10, 'n_estimators': 500}
- **Performance Metrics**:
  - **MAE**: 504,154.97
  - **MSE**: 1,122,128,344,538.37
  - **R² Score**: 0.9986
  - **Cross-Validation R² Scores**: [0.9953, 0.9961, 0.9918, 0.9967, 0.9950]
  - **Mean Cross-Validation R²**: 0.9950



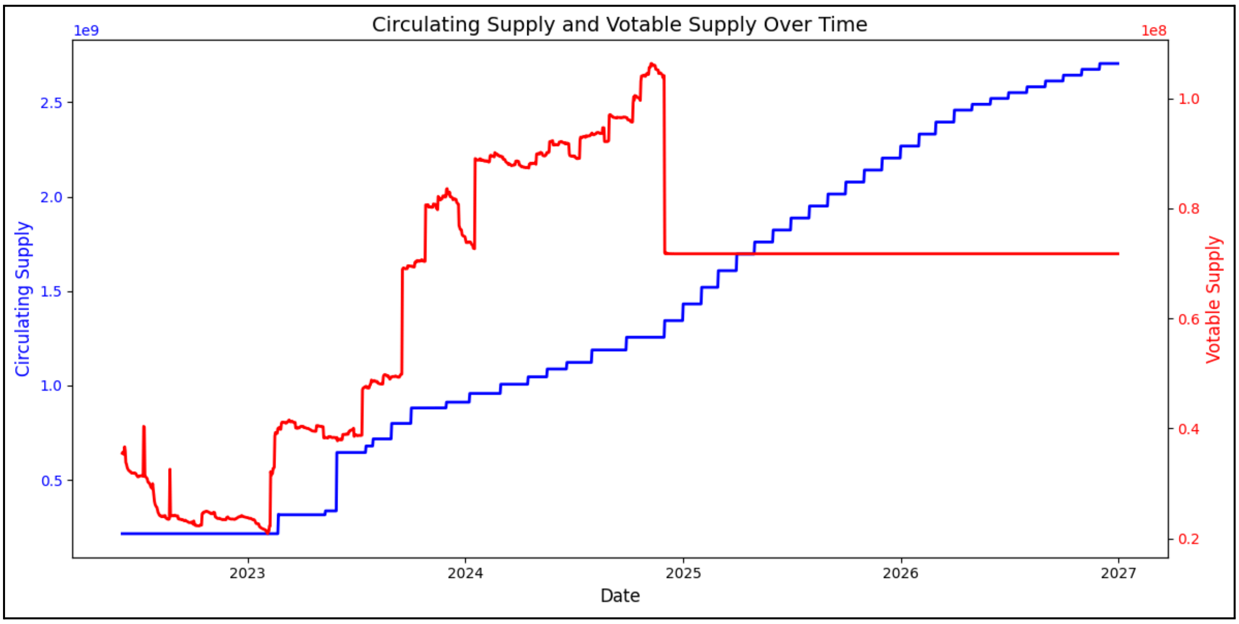
2.2 Random Split: Train 50% - Test 50%

- **Best Parameters**:
  {'max_depth': None, 'min_samples_split': 10, 'n_estimators': 200}
- **Performance Metrics**:
  - **MAE**: 874,673.27
  - **MSE**: 3,646,751,910,851.84
  - **R² Score**: 0.9955
  - **Cross-Validation R² Scores**: [0.9933, 0.9960, 0.9980, 0.9963, 0.9952]
  - **Mean Cross-Validation R²**: 0.9958

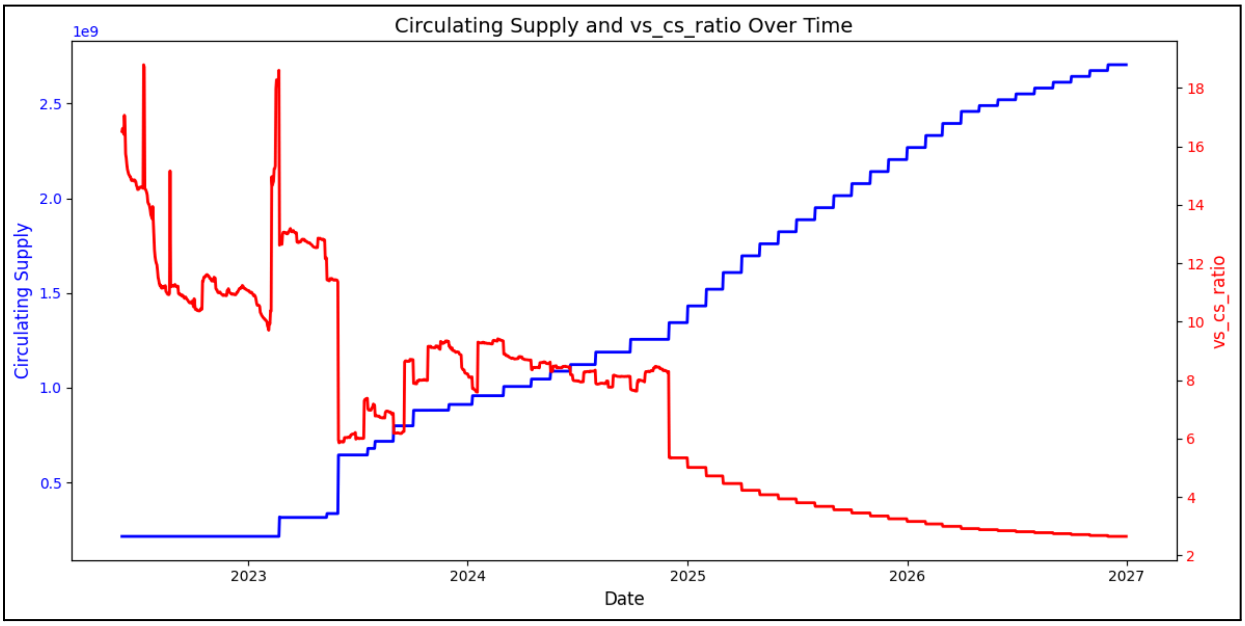


Prediction Results

1. Circulating Supply vs. Votable Supply



2. Circulating Supply vs. VS_CS_Ratio



Approach 3: Random Forest Regressor with Derived Features

Objective

The primary objective of this model was to assess the impact of derived features, specifically the ratio between Circulating Supply and Votable Supply (vs_cs_ratio), on the performance of predicting Votable Supply. Derived features were used to capture additional relationships and insights from the data.

Dataset

- Circulating Supply and its derived features.
- Airdrops data and its derived features.
- OP Price.
- Votable Supply and its derived features.

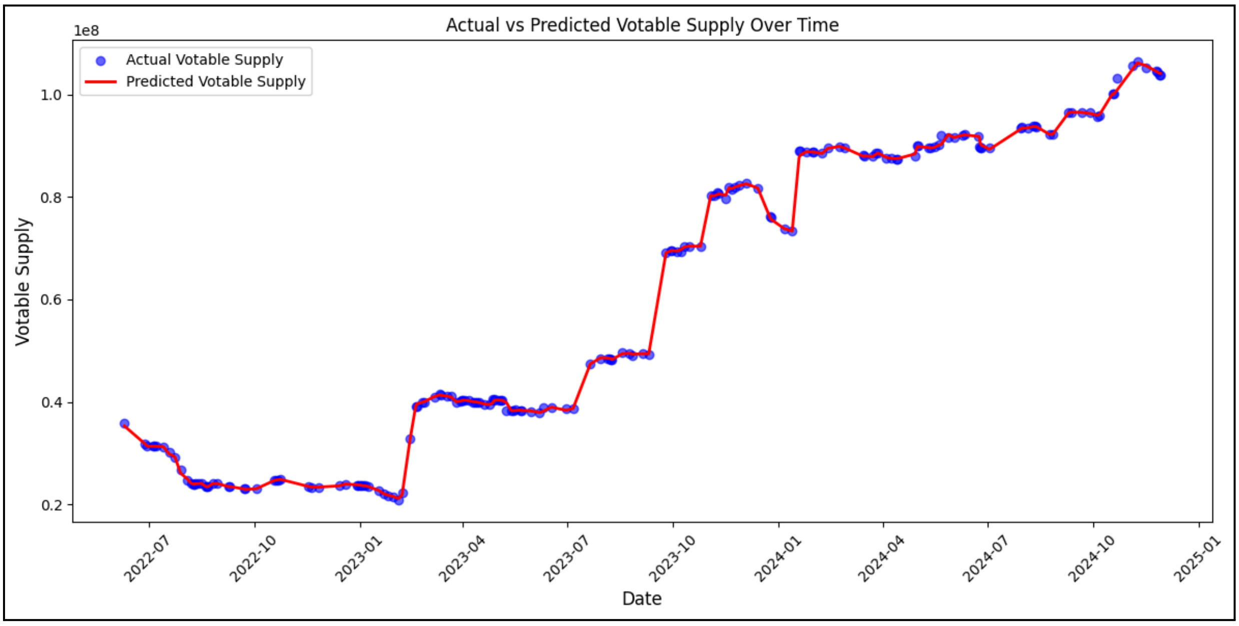
Features Used

['Circulating Supply', 'OP_Price', 'airdrop_tokens', 'airdrop_number', 'airdrop_targets_delegators', 'airdrop_lag', 'votable_lag_1', 'circulating_lag_1', 'votable_lag_7', 'circulating_lag_7', 'votable_lag_30', 'circulating_lag_30', 'vs_rolling_mean_7', 'cs_rolling_mean_7', 'vs_rolling_mean_30', 'cs_rolling_mean_30', 'vs_growth_rate_1', 'cs_growth_rate_1', 'effective_airdrop_tokens', 'airdrop_effect', 'vs_cs_ratio']

Testing and Results

3.1 Random Split: Train 80% - Test 20%

- **Best Parameters**:
  {'max_depth': 10, 'min_samples_split': 2, 'n_estimators': 100}
- **Performance Metrics**:
  - **MAE**: 164,895.64
  - **MSE**: 131,403,615,960.88
  - **R² Score**: 0.9998
  - **Cross-Validation R² Scores**: [0.9993, 0.9995, 0.9976, 0.9986, 0.9991]
  - **Mean Cross-Validation R²**: 0.9988


Actual vs Predicted Votable Supply Over Time
1e8
Actual Votable Supply
Predicted Votable Supply
Votable Supply
1.0
0.8
0.6
0.4
0.2
2022-07
2022-10
2023-01
2023-04
2023-07
2023-10
2024-01
2024-04
2024-07
2024-10
2025-01
Date

Approach 4: Polynomial Regression (degree = 2)

Objective

The objective of this approach is to leverage Polynomial Regression (degree = 2) to predict the future votable supply by capturing the nonlinear relationships between the features in the dataset.

Dataset:

- Circulating Supply and its derived features
- Airdrops data and its derived features
- OP_Price

Features Used

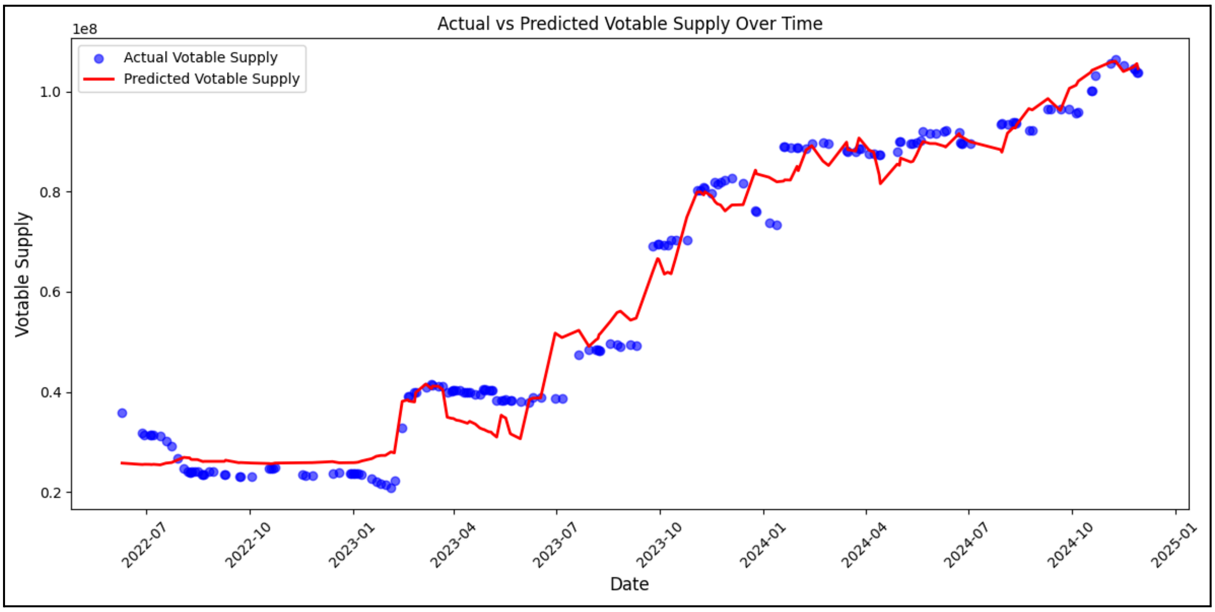
['Circulating Supply', 'OP_Price', 'airdrop_tokens', 'airdrop_number', 'airdrop_targets_delegators', 'airdrop_lag', 'circulating_lag_1', 'circulating_lag_7', 'circulating_lag_30', 'cs_rolling_mean_7', 'cs_rolling_mean_30', 'cs_growth_rate_1', 'effective_airdrop_tokens', 'airdrop_effect']

Testing and Results

4.1 Test Model on Random Split: Train-80% Test-20%

Testing Results:

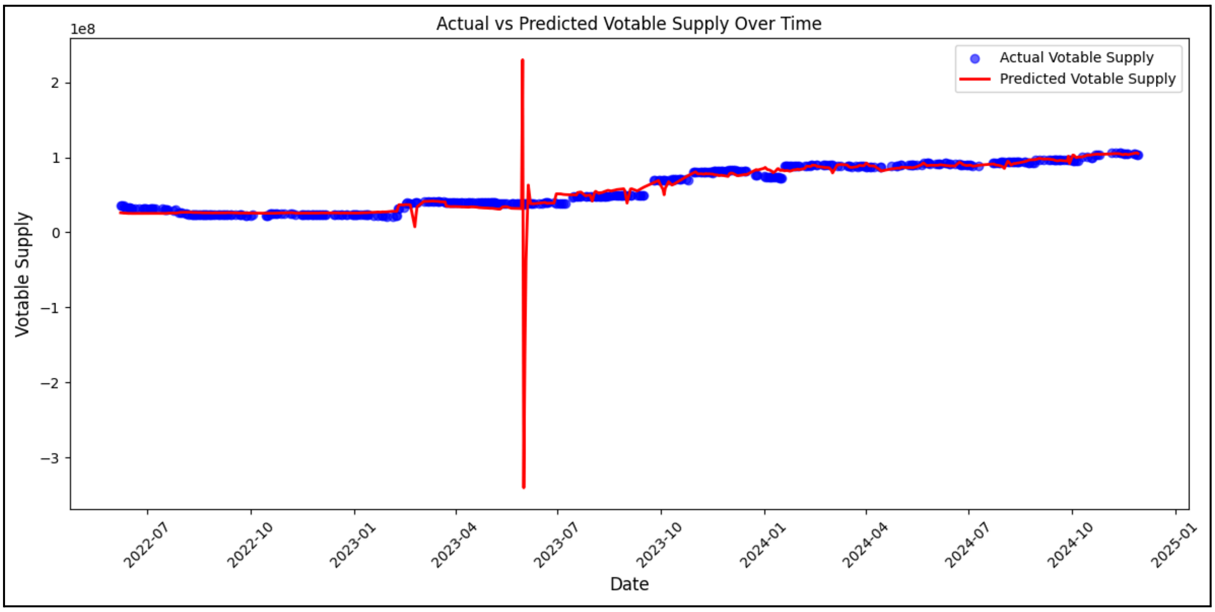
- MAE: 3587089.388592087
- MSE: 19469265646068.934
- R²: 0.9764902408577893



4.2 Test Model on Random Split: Train-50% Test-50%

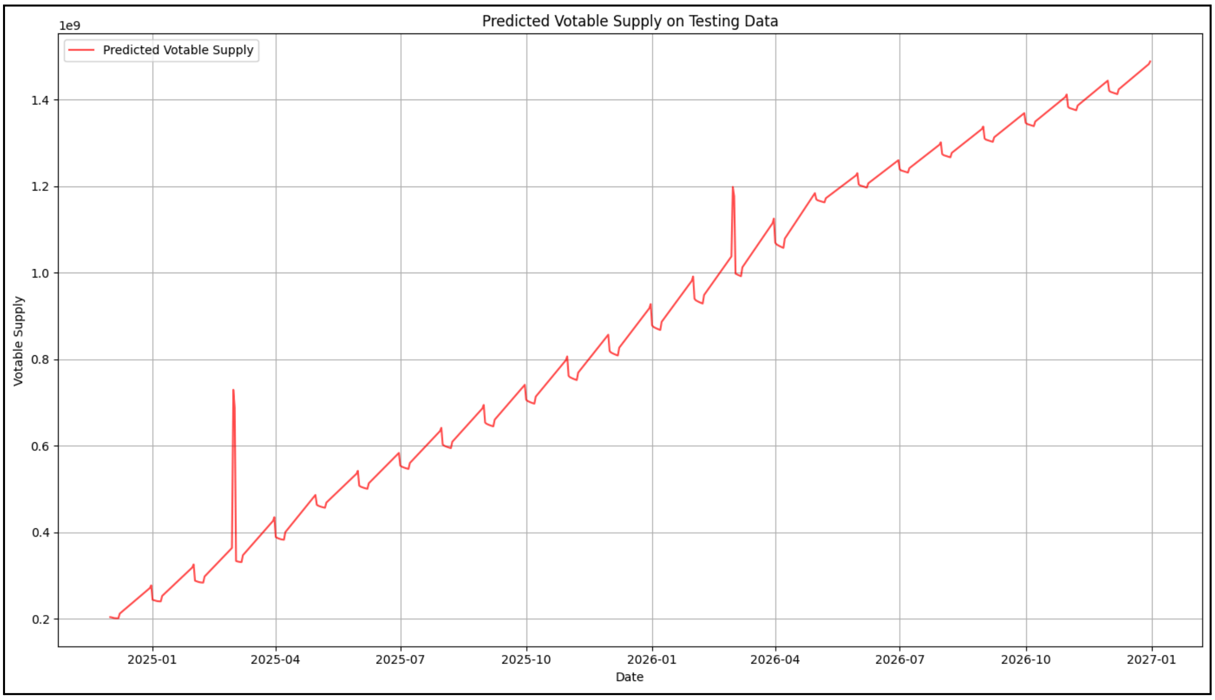
Testing Results:

- MAE: 5885396.45059682
- MSE: 524179355296876.5
- R²: 0.35663883136969154

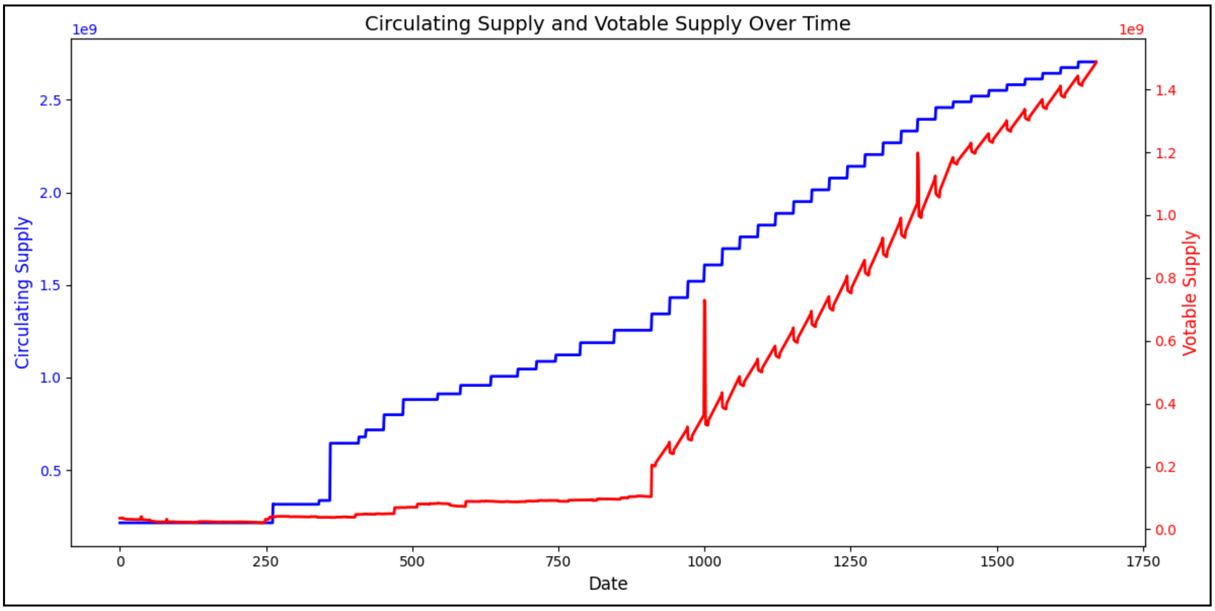


Prediction Results

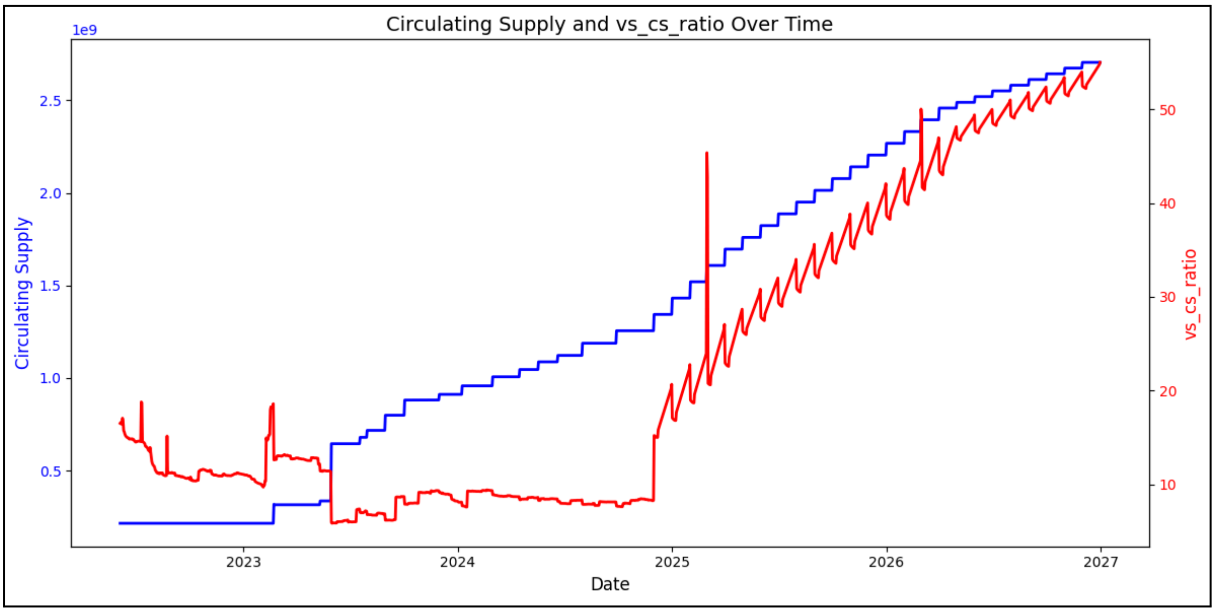
1. **Predicted Future Votable Supply**



2. Circulating Supply vs. Votable Supply



3. Circulating Supply vs. VS_CS_Ratio

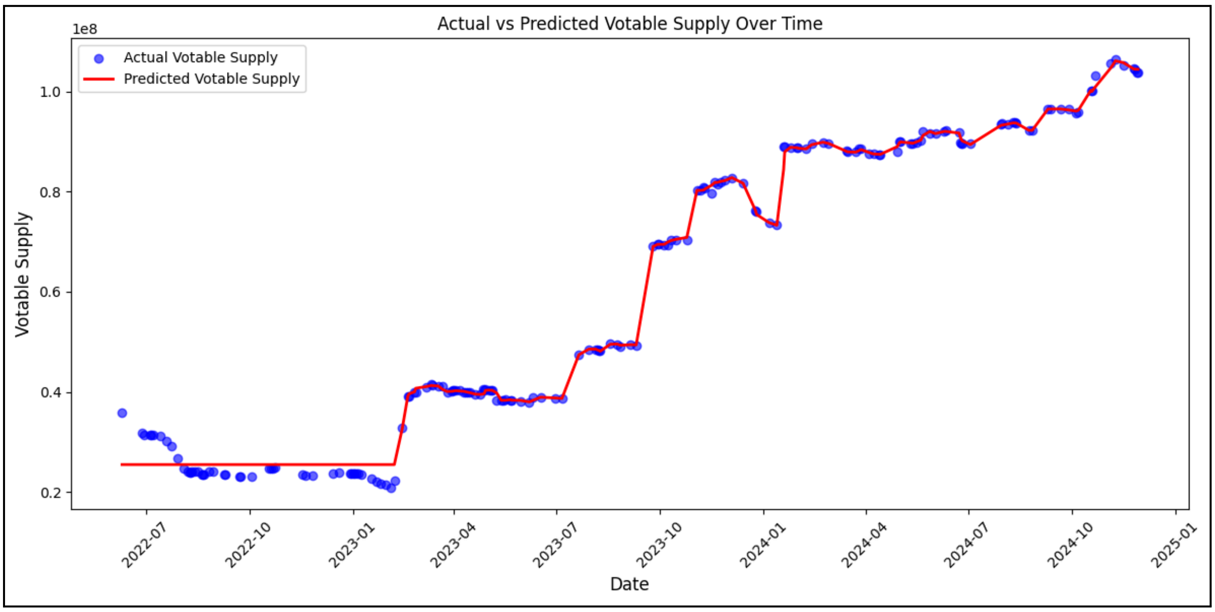


Approach 5: Random Forest Regressor Using Historical and Future Data Integration

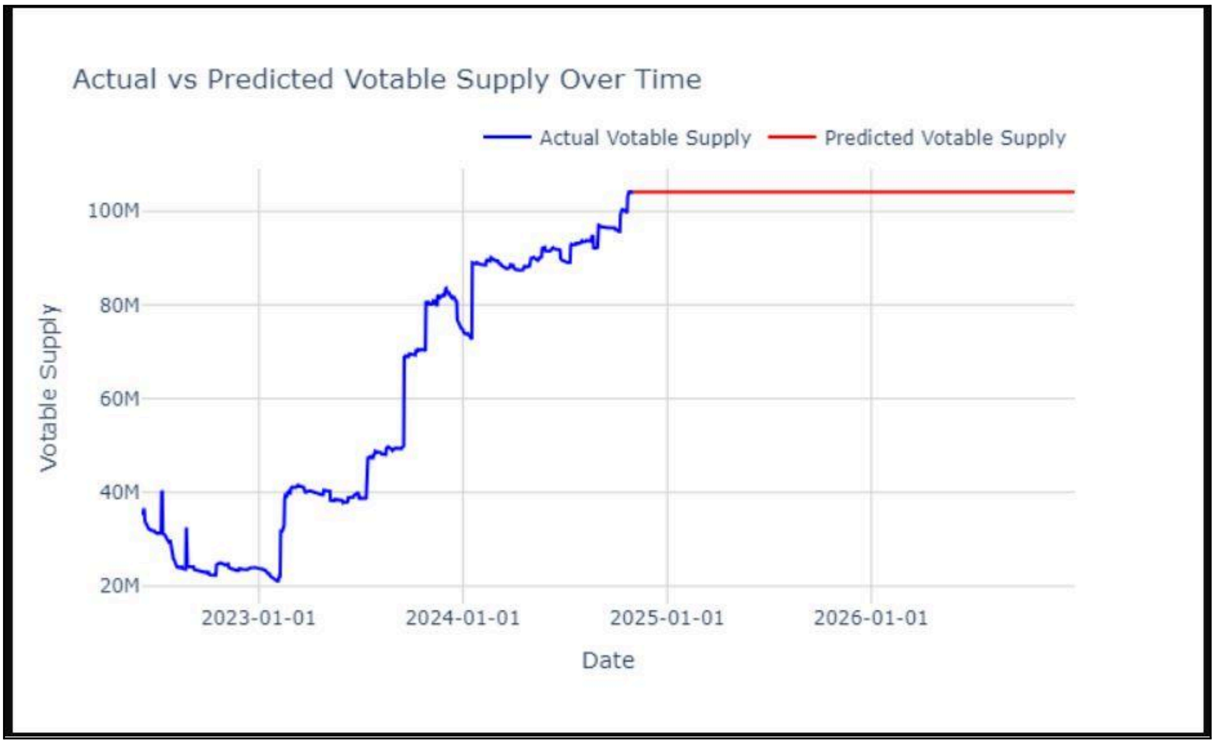
Testing and Results

5.1 Random Split: Train 80% - Test 20%

- **Dataset**
  - Circulating Supply and its derived features
  - Airdrop data and its derived features
- **Best Parameters:**
  - min_samples_split: 5
  - max_depth: 10
  - n_estimators: 500
- **Performance Metrics:**
  - **MAE**: 918252.9268157823
  - **MSE**: 3172994892650.1694
  - **R² Score**: 0.9961685074803666
  - **Cross-Validation R² Scores**: [0.99478075, 0.99309557, 0.99389371, 0.99444824, 0.99182163]
  - **Mean Cross-Validation R²**: 0.9936079808227015



Prediction Results



Approach 6: ARIMA Model

Walk-Forward Validation:

- An ARIMA(1,1,1) model is fitted iteratively for each time step in the test set.
- After forecasting each value, the prediction is compared with the actual observation, and the observation is added to the history to simulate a real-time forecasting process.

Dataset:

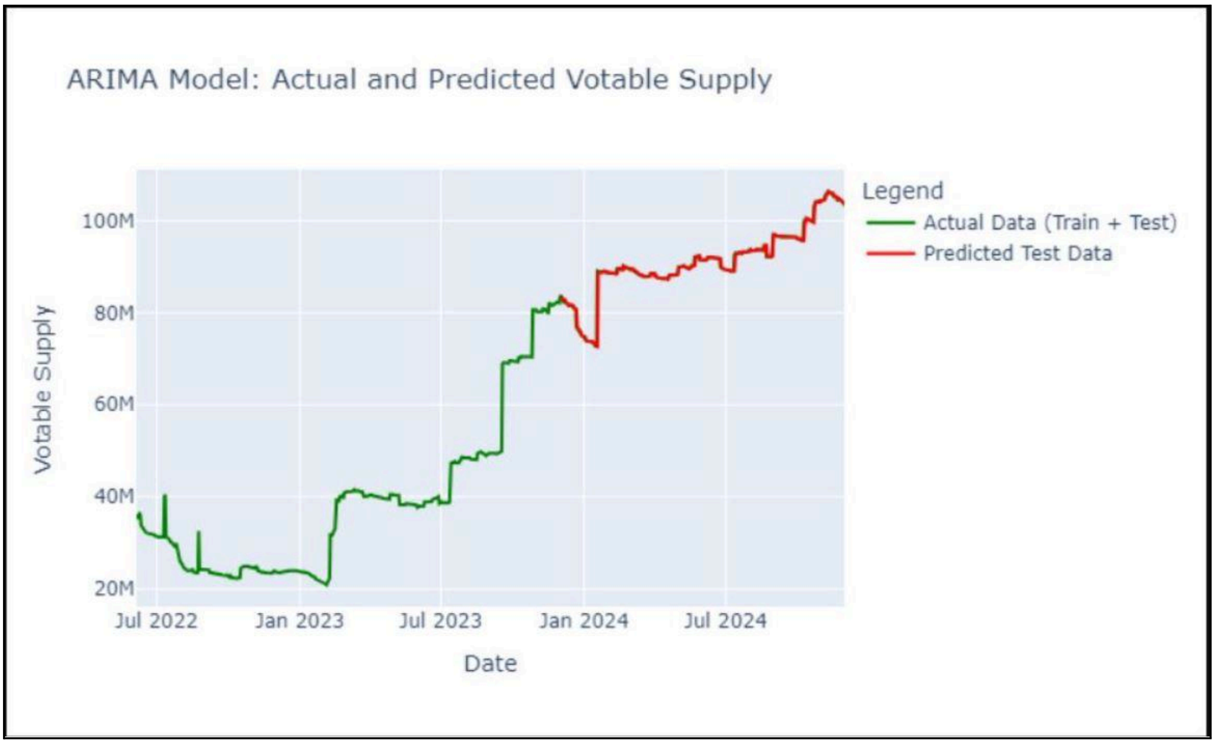
- Votable Supply

Testing and Results

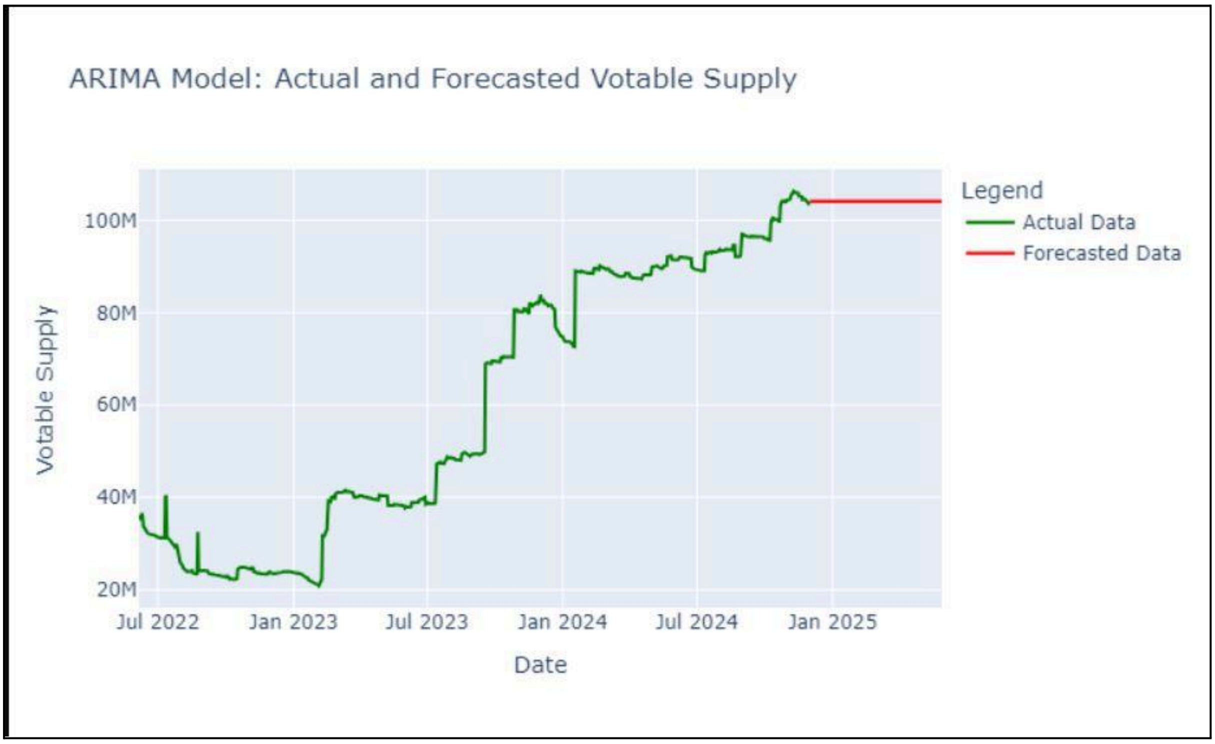
6.1 Test Rolling Forecast ARIMA Model (Train-60%, Test-40%, order=(1,1,1))

Testing Results:

- Test RMSE: 1017545.011
- Test R² Score: 0.981



Forecasting Results:-

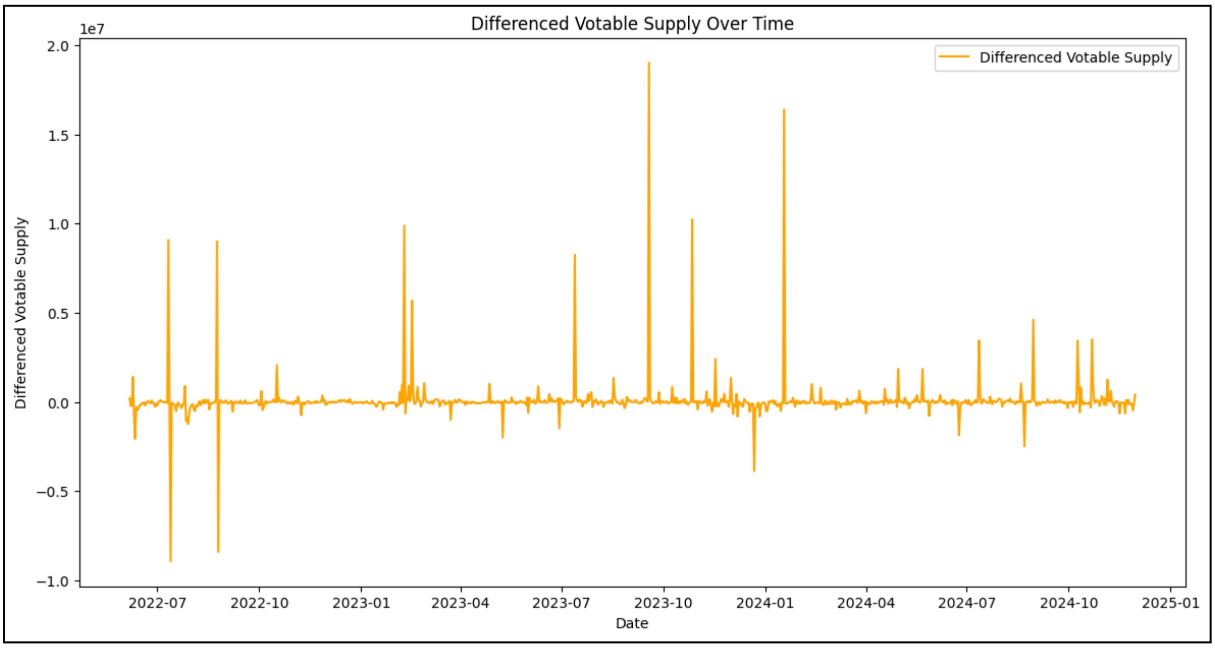


6.2 Forecast using ARIMA Model

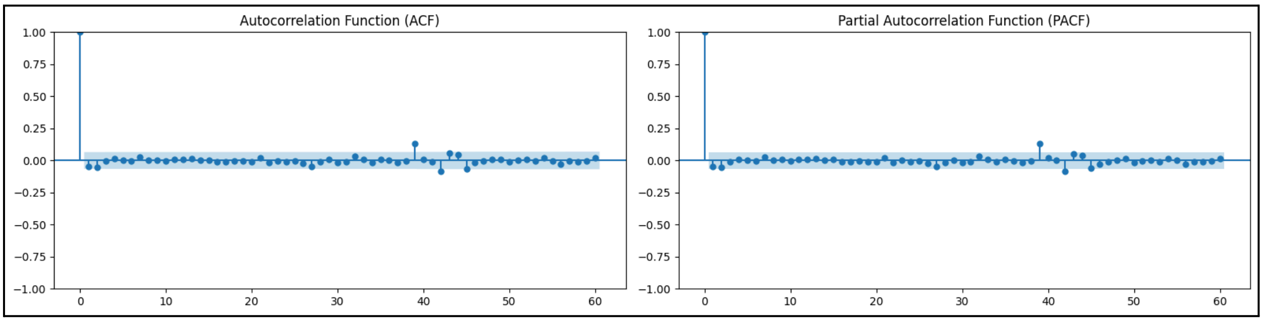
Statistical Results:

- **Stationarity Results**
  - ADF Statistic (Original): 0.0980
  - p-value (Original): 0.9659
  - Interpretation: The original series is Non-Stationary.
  - ADF Statistic (Differenced): -22.9562
  - p-value (Differenced): 0.0000
  - Interpretation: The differenced series is Stationary.

- **Differenced Votable Supply Graph**

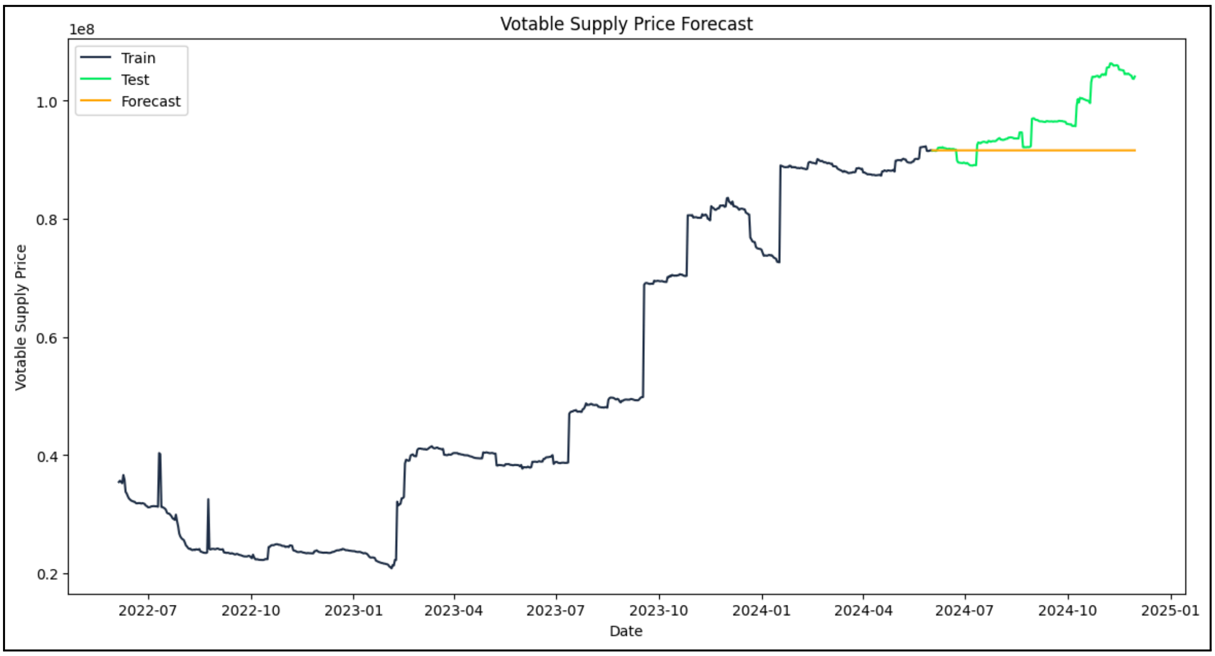


- **ACF and PACF plots**



Model Evaluation (Train-80%, Test-20%, order=(1,1,1))

- AIC: 22597.609560346034
- BIC: 22620.554192733704
- RMSE: 7073001.7025



Additional Findings:

- ARIMA produced constant predictions, which limited its effectiveness due to several factors:
  - Its reliance on stationarity.
  - Inability to incorporate additional influencing factors like Circulating Supply and OP Price.
  - Oversimplification of complex patterns.

Next Steps: To address ARIMA's limitations, we are now exploring advanced models like LSTM (Long Short-Term Memory), which are better suited for handling multivariate data and capturing more dynamic and accurate patterns.

Approach 7: LSTM Model (Univariate)

7.1 Dataset:

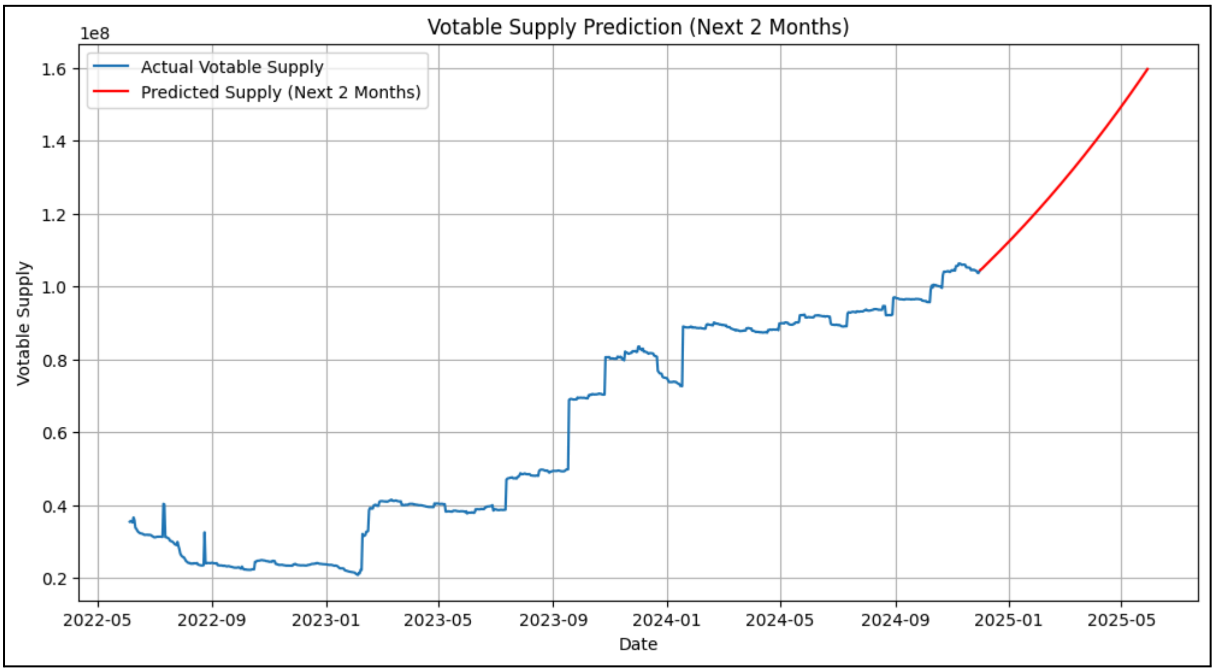
- Votable Supply (without removing event-related data).

Event-Related Data:

- **Airdrop Events:**
  - February 9, 2023: 11.7 million tokens distributed.
  - September 18, 2023: 19.4 million tokens distributed.
- **Other Events:**
  - February 16, 2023 (PDP): 5.68 million tokens.
  - July 13, 2023: 8.25 million tokens.
  - October 27, 2023 (ACC): 10.25 million tokens.
  - January 18, 2024 (a16z): 16.40 million tokens.

Prediction Results

- Forecasted data for the next 180 days.



7.2 Dataset:

- Votable Supply (after removing event-related data).

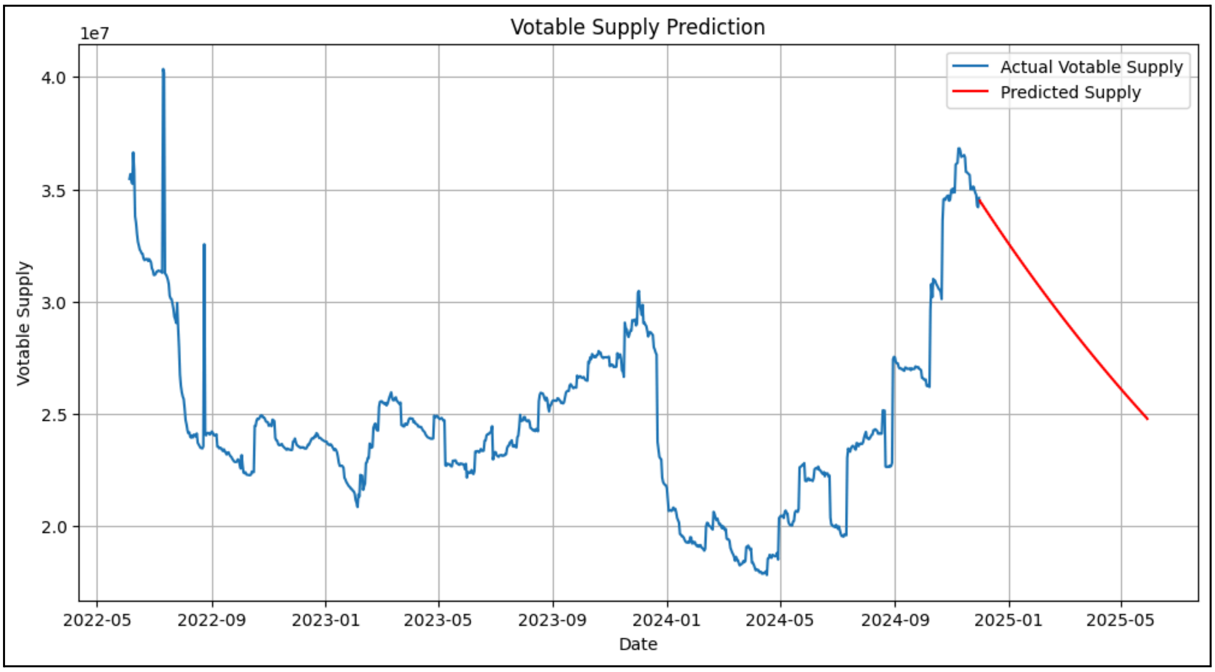
Event-Related Data:

- **Airdrop Events:**
  - February 9, 2023: 11.7 million tokens distributed.
  - September 18, 2023: 19.4 million tokens distributed.

- **Other Events:**
    - February 16, 2023 (PDP): 5.68 million tokens.
    - July 13, 2023: 8.25 million tokens.
    - October 27, 2023 (ACC): 10.25 million tokens.
    - January 18, 2024 (a16z): 16.40 million tokens.

Prediction Results

- Forecasted data for the next 180 days.



Reason for Not Using Univariate LSTM for FVS Prediction:

Univariate LSTM relies solely on past Votable Supply (VS) data to predict future Votable Supply (FVS). This limited approach fails to consider other influential factors, such as Circulating Supply (CS) and OP_Price, which are crucial for improving the accuracy and reliability of FVS predictions.

Approach 8: LSTM Model (Multivariate)

Dataset:

- Circulating Supply
- Votable Supply (after removing the events data)
- OP_Price

Event-Related Data Removed:

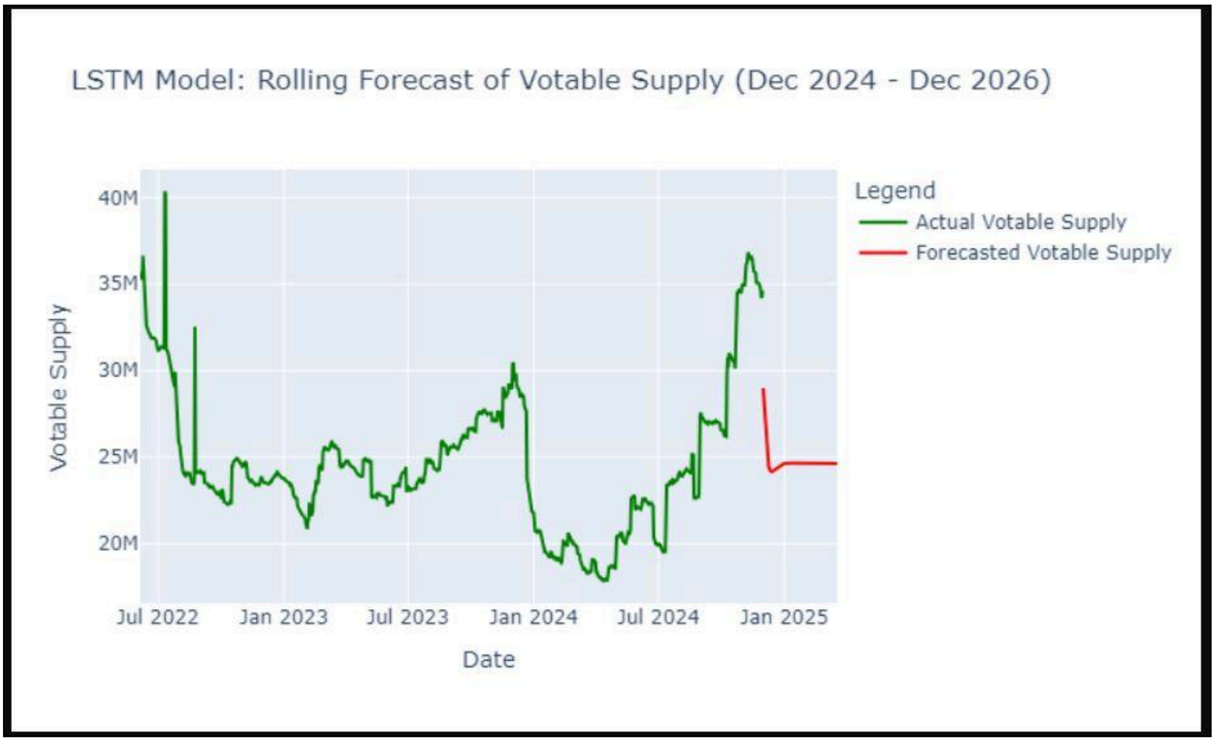
- **Airdrop Events:**
  - February 9, 2023: 11.7 million tokens.
  - September 18, 2023: 19.4 million tokens.

- **Other Events:**
  - February 16, 2023 (PDP): 5.68 million tokens.
  - July 13, 2023: 8.25 million tokens.
  - October 27, 2023 (ACC): 10.25 million tokens.
  - January 18, 2024 (a16z): 16.40 million tokens.

Prediction Results

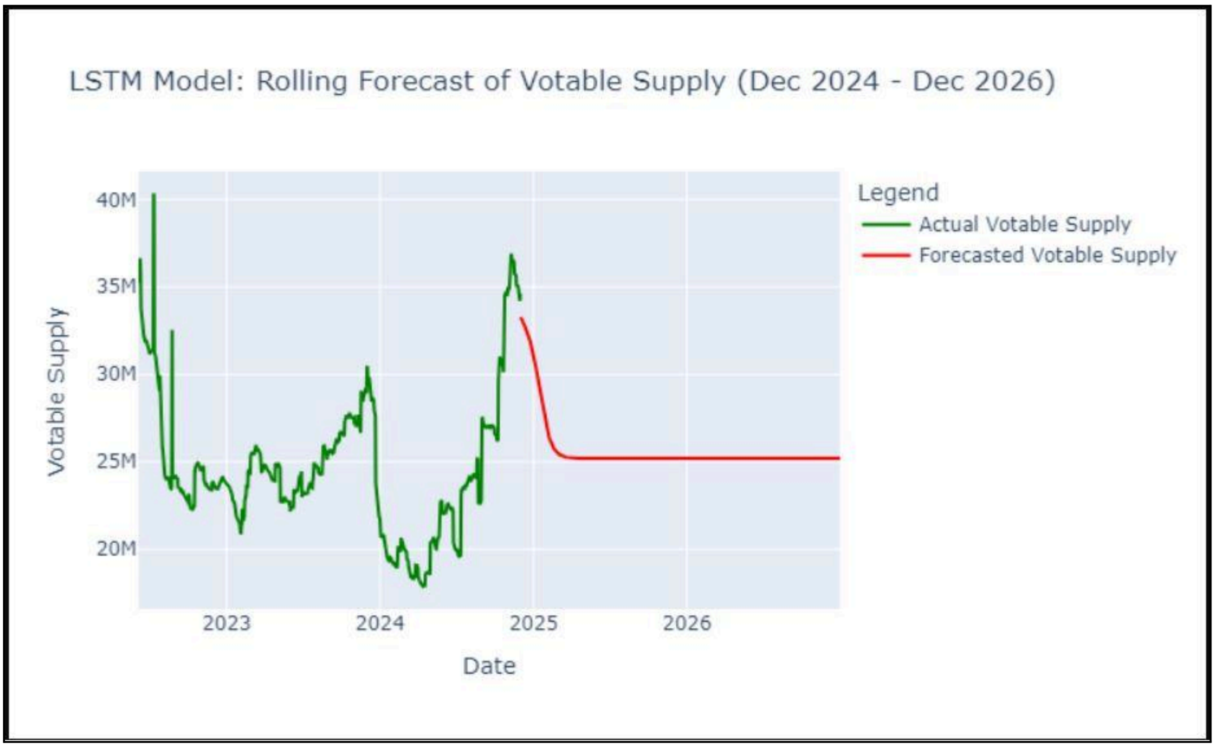
8.1 Rolling Window Approach (7-Day)

- Number of past days used for prediction: 30 days.
- Forecasted votable supply using a 7-day rolling window approach for training and predictions.



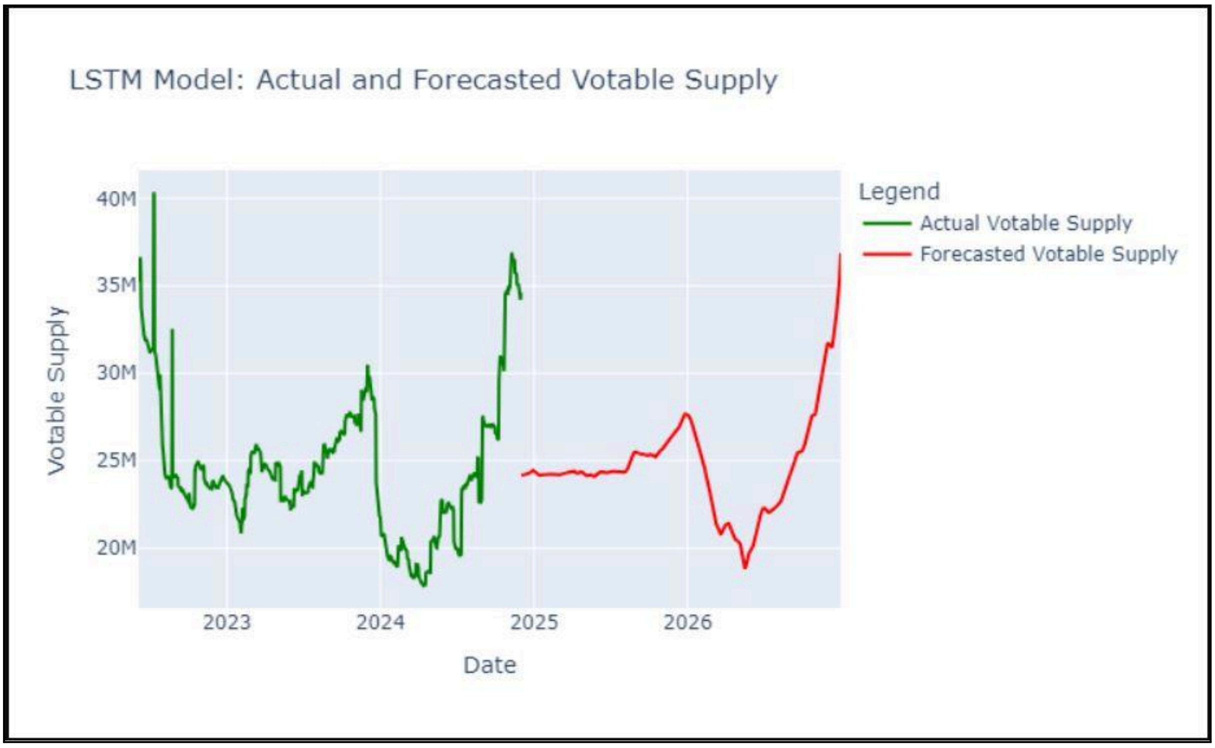
8.2 Rolling Window Approach (1-Day)

- Number of past days used for prediction: Entire dataset.
- Forecasted votable supply using a 1-day rolling window approach for training and predictions.



8.3 Forecasting the data for 2 years

- Number of past days used for prediction: 90 days
- Forecasted votable supply using past 90 days data for training and predicting the Votable Supply till December 2026.



Reason for Not Using Multivariate LSTM for FVS Prediction:

The Multivariate LSTM model, despite incorporating additional features such as Circulating Supply (CS) and OP_Price, shows limitations in its predictive performance. When forecasting Future Votable Supply (FVS) for a 2-year period, the model exhibits a sharp decrease in votable supply, which is inconsistent with expected trends. Additionally, when using the rolling mean approach, the predictions stabilize to constant values after a certain duration, indicating that the model fails to capture the long-term variations and dynamics in the data effectively.