Final Report: Cryptocurrency Liquidity Prediction

1. Introduction

This project aims to predict the 24-hour trading volume (liquidity) of cryptocurrencies using machine learning.
Accurate liquidity prediction helps traders and exchanges understand how easily a coin can be bought or sold.
The system uses real-world market data from CoinGecko and provides predictions through a web interface.

2. Dataset Description

- Source: CoinGecko market data
- Files: coin_gecko_2022-03-16.csv, coin_gecko_2022-03-17.csv
- Records after cleaning: 992
- Features:
    - price
    - 1h, 24h, 7d price changes
    - market_cap
    - 24h_volume (target)

Target Variable:
24h_volume – represents trading liquidity, and is scaled using StandardScaler.

3. Methodology

1.Preprocessing

- Removed missing values using dropna()
- Converted date to datetime format
- Normalized all numerical features (e.g. price, volume, market_cap) using StandardScaler to improve model performance

2. Feature Engineering

Two new features were created:

- price_change_score = 0.2 × 1h + 0.3 × 24h + 0.5 × 7d
  This gives a weighted estimate of short-term price volatility.
- volume_to_marketcap = 24h_volume / market_cap
  Measures liquidity efficiency.

3. Model Training

- Used XGBoost Regressor (known for speed and accuracy in tabular data)
- Hyperparameters tuned via GridSearchCV
- Train-test split: 80% training, 20% testing

4. Model Evaluation

Evaluated using:

- R^2 Score: 0.92
- RMSE: 0.35
- MAE: 0.036

These metrics indicate excellent predictive performance with low error.

5. Model Saving

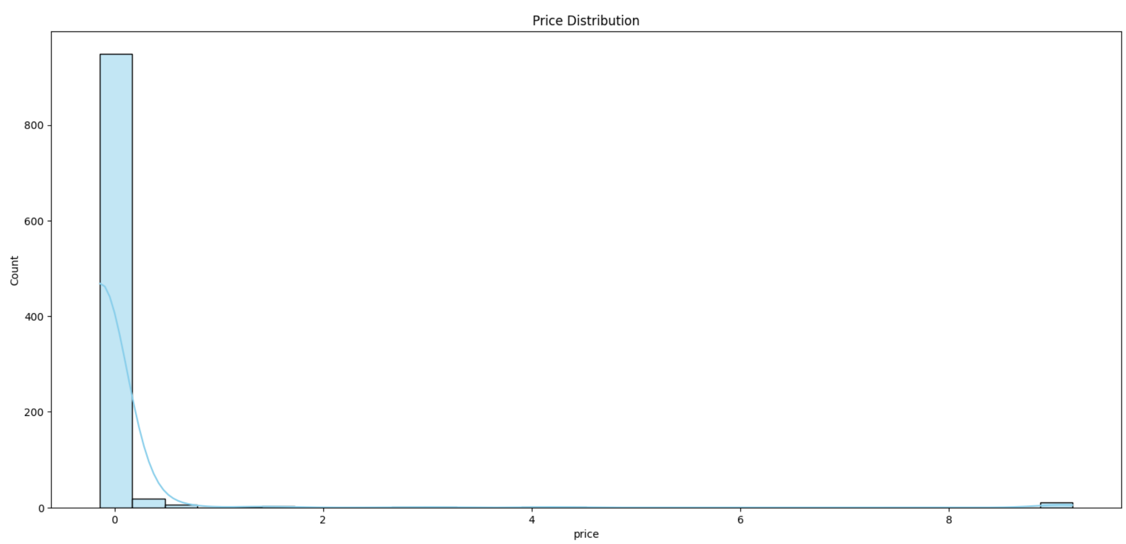
- Trained model saved as xgb_model.pkl
- Scaler saved as volume_scaler.pkl (for reverse transformation)

6. Deployment

- A Streamlit-based web app (app.py) accepts user inputs and shows predictions instantly

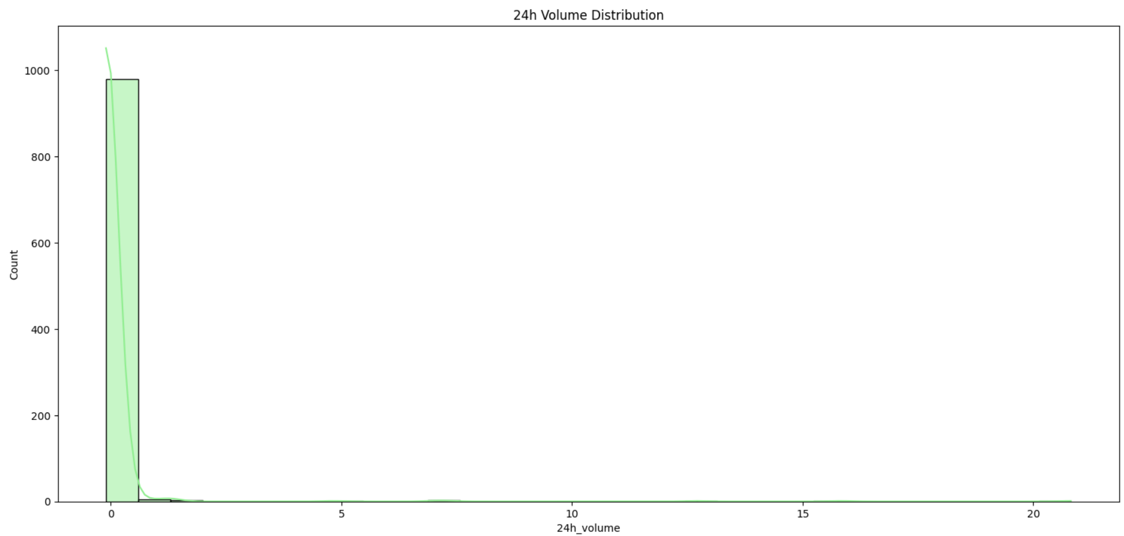
4. Visual Summary

1. Price Distribution



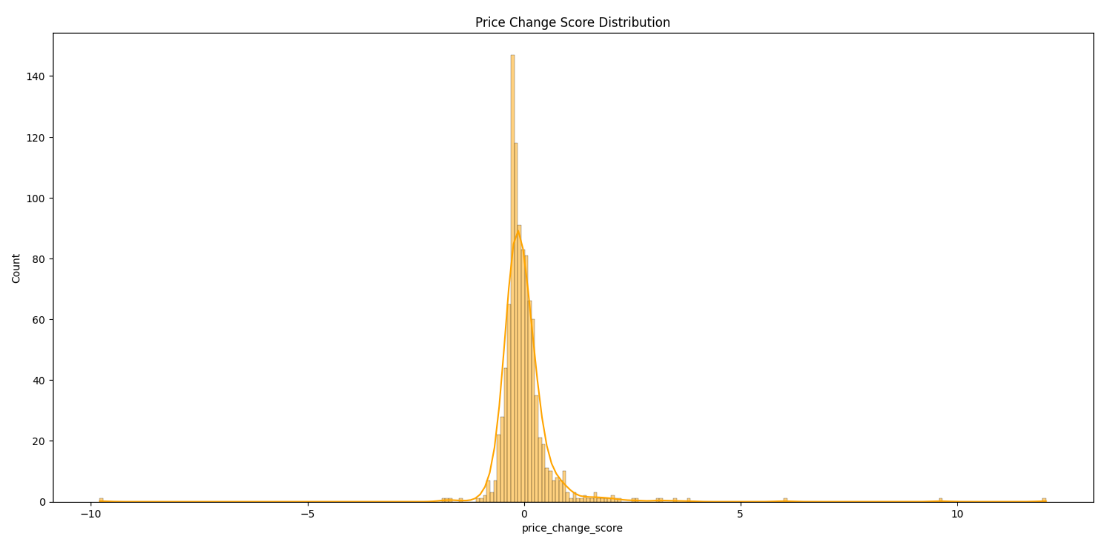
- Right-skewed distribution
- Most cryptocurrencies have low prices; a few (e.g. Bitcoin) are extreme outliers

2. 24h Volume Distribution



- Strongly skewed right
- Confirms that most coins are illiquid, and only a few dominate trade volume

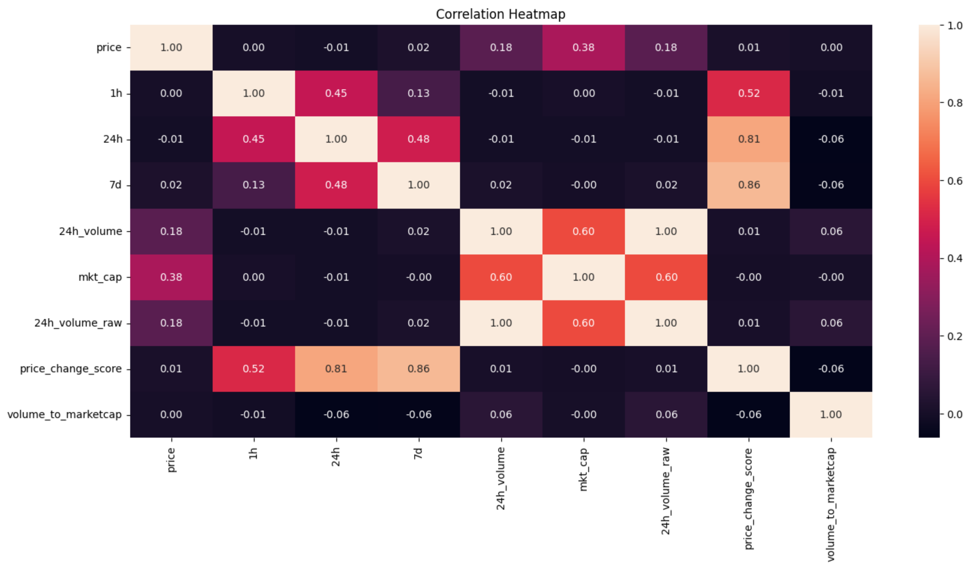
3.Price Change Score Distribution



- Bell-shaped, centered around 0

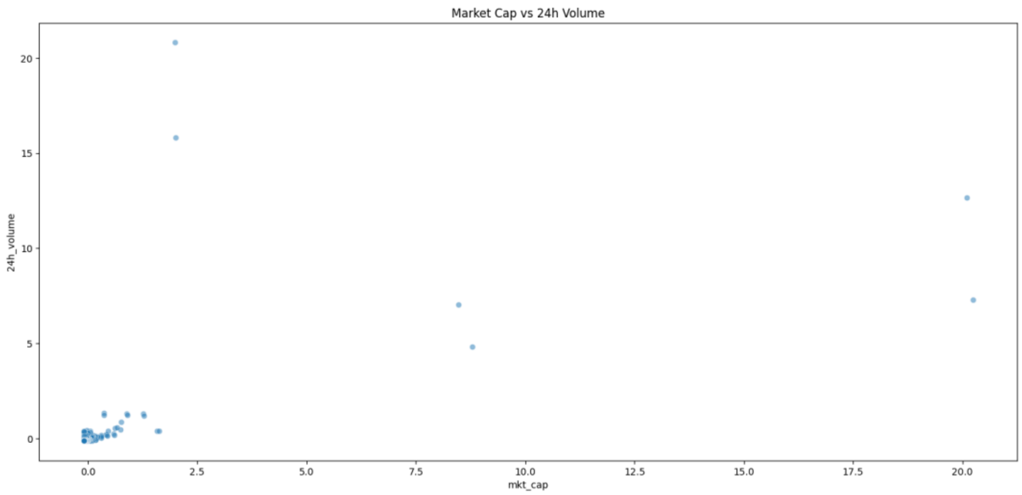
- Shows most coins had stable prices; some experienced high volatility

4.Correlation Heatmap



- price_change_score highly correlates with 24h and 7d changes
- volume_to_marketcap is largely uncorrelated — valuable as a unique feature
- No multicollinearity

5.Market Cap vs 24h Volume



- Weak positive trend
- Some high-market-cap coins have surprisingly low liquidity, making this a non-linear relationship

5. Key Insights

- Most cryptocurrencies have low liquidity, while a small set account for high volume
- Feature engineering significantly improves model performance
- volume_to_marketcap and price_change_score are powerful predictors
- XGBoost performed better than baseline regression models (like Linear Regression or Decision Trees)

6. Deployment

A fully functional Streamlit app was built:

- User inputs 4 parameters:
    - Price
    - Price Change Score
    - Volume-to-Market Cap Ratio
    - Market Cap
- The app uses the trained model to predict liquidity in real-time
- Output: Predicted 24-hour trading volume (scaled)

7. Model Performance Summary

Metric	Value	Description
R^2 Score	0.92	92% of the variance is explained
RMSE	0.35	Low average error
MAE	0.036	Very low average absolute error

8. Final Pipeline Flow

CSV Data → Preprocessing (drop NA, scale features)→ Feature Engineering (score, ratio)→ EDA (visualizations) → XGBoost Model Training (with tuning) → Model Evaluation (RMSE, R^2, MAE) → Model Saving → User Input via Streamlit → Prediction → Output

9. Conclusion

This project successfully built a machine learning model to predict cryptocurrency liquidity using real data.

The system is:

- Accurate ($R^2 \approx 0.92$)
- Scalable (modular code, saved model)
- Interactive (Streamlit web app)

Next steps could include:

- Integration with live APIs (e.g. CoinGecko API)
- Time-series models for forecasting
- Handling more granular data (e.g. hourly liquidity trends)