📝 Cryptocurrency Market Cap Analysis & Prediction Pipeline

Overview

This project builds a data pipeline to analyze and predict cryptocurrency market behavior using historical CoinGecko data. The pipeline includes data cleaning, feature engineering, modeling, and evaluation focused on the 24h_mkt_cap_ratio.

Objectives

- Integrate and preprocess CoinGecko market data.
- Handle outliers and missing values.
- Engineer domain-relevant features.
- Train and evaluate a regression model.

Data Sources

- coin_gecko_2022-03-16.csv
- coin_gecko_2022-03-17.csv

Pipeline Stages

1. Data Ingestion

- Load and merge datasets using pandas and numpy.

2. Data Cleaning

- Drop symbol, coin, and date.
- Convert all columns to float.

3. Outlier Detection

- IQR-based clipping applied per numeric column.

4. Missing Value Imputation

- Median imputation on 1h, 24h, 7d, and 24h_volume.

5. Feature Engineering

- variability_score: Std dev of 1h, 24h, 7d.
- 24h_mkt_cap_ratio: 24h change / market cap.
- coin_number: Derived from market cap and price.

6. Exploratory Data Analysis

- Histograms and KDE plots to visualize distributions.

7. Data Preparation

- Drop leakage-prone columns.
- Split into features (X) and target (y).
- Apply standard scaling and train-test split (70-30).

8. Model Training

- Regression model (e.g., RandomForestRegressor) trained on X_train.

9. Model Evaluation

- Predictions on X_test.
- Performance assessed using standard regression metrics (e.g., MAE, RMSE).

Technologies Used

- Python (Pandas, NumPy, Seaborn, Scikit-learn)

Outputs

- Cleaned dataset
- Engineered features
- Trained model
- Predictions for 24h_mkt_cap_ratio

Future Enhancements

- Real-time data streaming from APIs
- Ensemble modeling
- Hyperparameter tuning

- Longer time series with temporal features

Conclusion

This pipeline provides a solid base for further cryptocurrency modeling, capable of powering decision-support tools or trading strategies using machine learning.
