📘 Final Report – Cryptocurrency Volatility Prediction

Executive Summary

This project presents a robust machine learning pipeline for forecasting next-day cryptocurrency volatility using engineered features and an XGBoost model. The solution includes a reproducible data workflow, model training with time-series validation, and an interactive Streamlit application for real-time prediction and risk regime classification. The pipeline aligns tightly with insights derived from exploratory data analysis and is designed for modular deployment and future extensibility.

Problem Statement

Cryptocurrency markets are characterized by high volatility and regime-dependent behavior, posing challenges for traders and risk managers. This project aims to predict next-day volatility using historical price and liquidity data, enabling proactive decision-making and scenario testing.

Methodology

Data and Features

- **Inputs:** Daily OHLCV and market cap data across multiple cryptocurrencies.
- **Engineered Features:**
  - Log returns
  - Rolling volatility (7d, 14d)
  - Garman–Klass volatility
  - High–low spread
  - Turnover ratio
  - Calendar features (day of week, month, weekend flag)
- **Target Variable:** Next-day Garman–Klass volatility (`target_vol`)

Modeling

- **Algorithm:** XGBoost regressor within a pipeline.
- **Validation Strategy:** Time-series split (train/validation/test).
- **Performance Metrics:**

Split	RMSE	MAE	R^2
Train	0.00314	0.00199	0.9911
Validation	0.01019	0.00633	0.8790
Test	0.00881	0.00485	0.8965

Artifacts and Deployment

- **Model Artifact:** Saved as `xgb_volatility.joblib`
- **App Interface:** Streamlit application with two modes:
  - **Auto Mode:** Uses latest available features for selected cryptocurrency.
  - **Manual Mode:** Accepts user-defined feature inputs for scenario testing.
- **Risk Regime Classification:** Based on 80th percentile of historical GK volatility.

Results

Quantitative Performance

- High R^2 scores across splits indicate strong predictive power.
- Low RMSE and MAE values confirm model accuracy and stability.

Qualitative Behavior

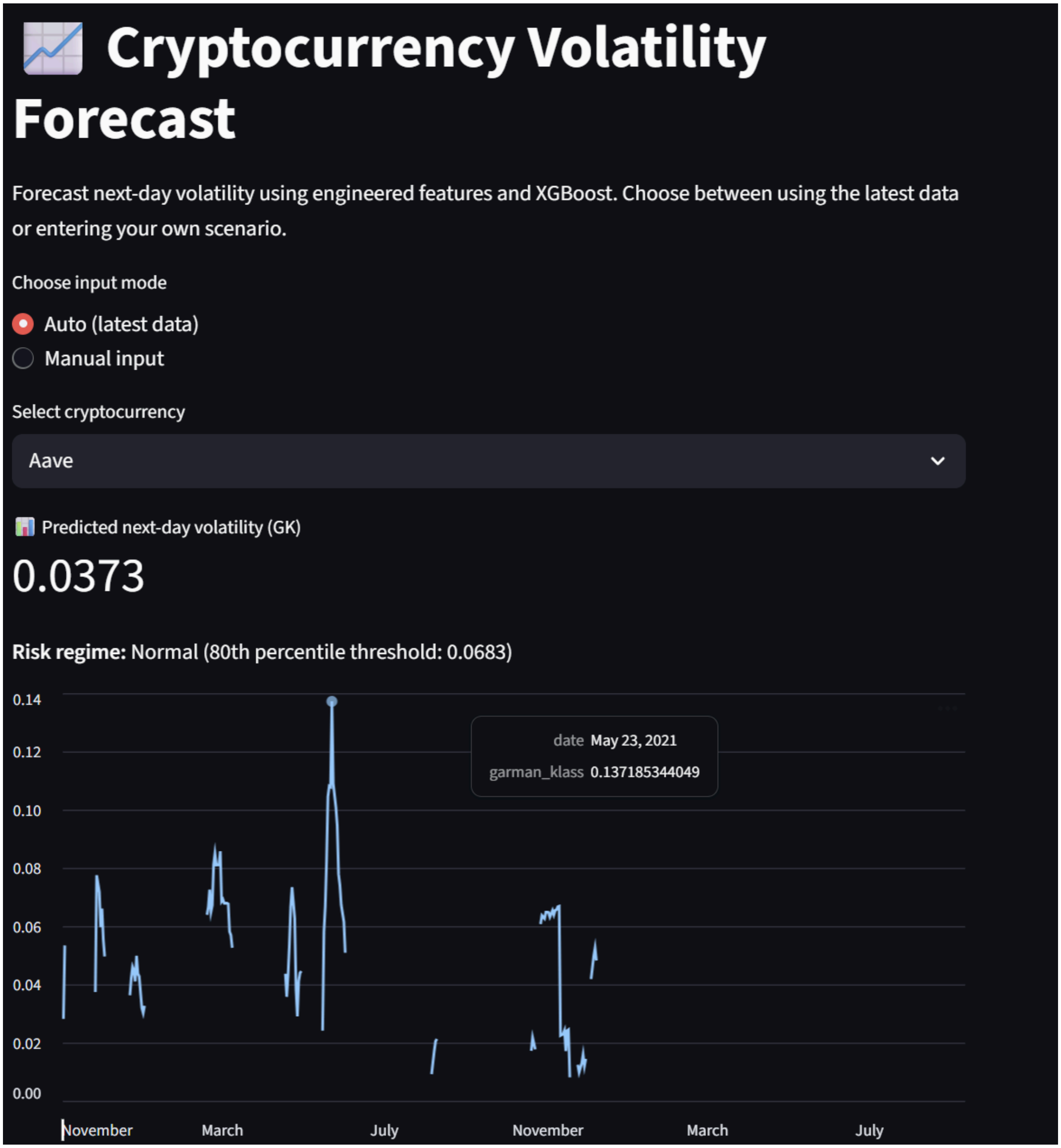
- Predictions align with known volatility regimes.
- Manual mode enables flexible scenario analysis (e.g., higher spreads → higher predicted volatility).

Visual Outputs

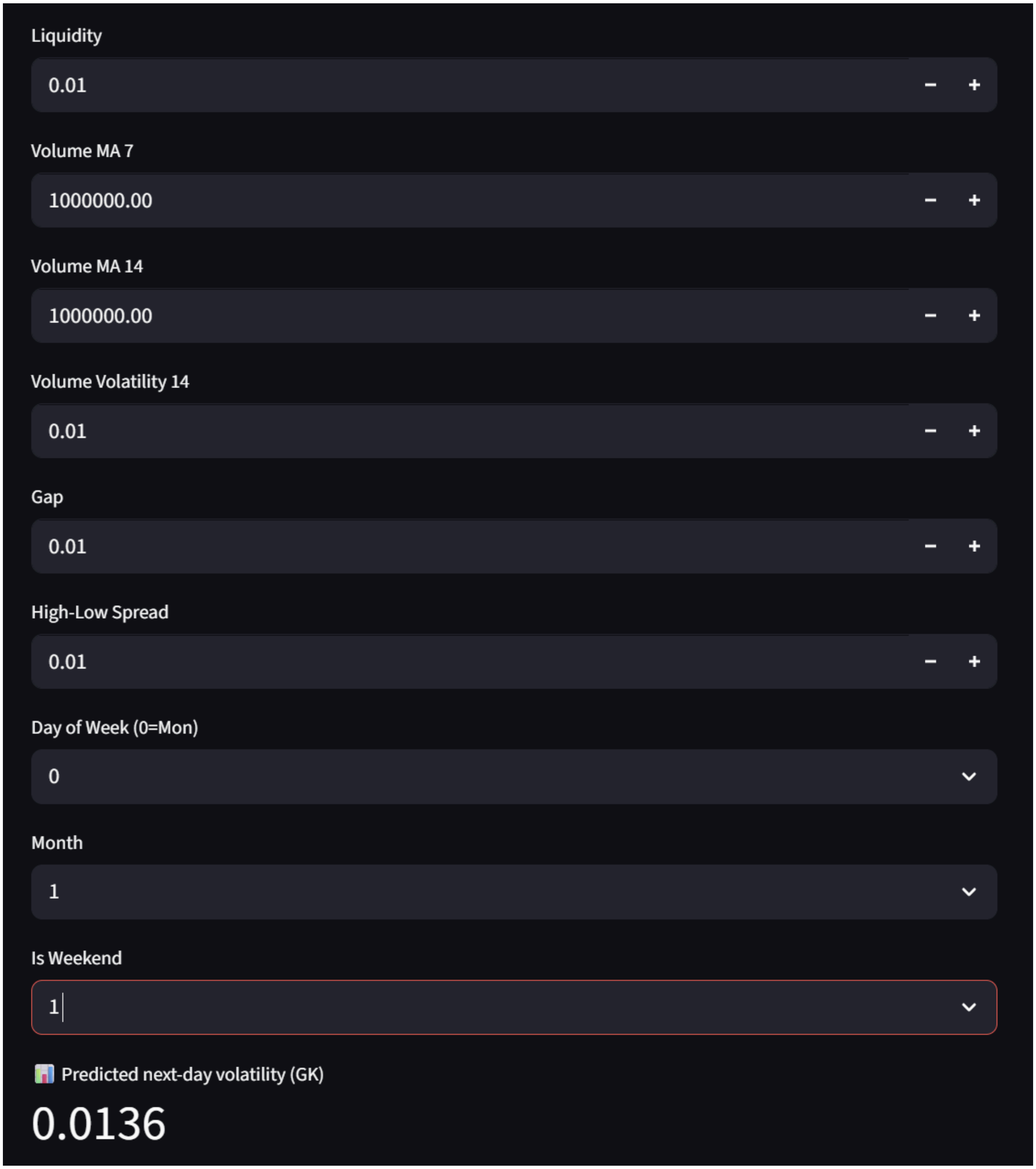
- Historical GK volatility line chart with predicted overlay.
- Feature importance highlights GK, vol_14d, and hl_spread as key drivers.

Deployment and User Experience

- **Streamlit App Features:**
  - **Modes:**
    - **Auto:** Uses latest features per crypto; guards against empty histories.



- **Manual:** User-defined feature inputs for scenario analysis.



- Real-time prediction of next-day GK volatility.
- Risk regime classification: “High” vs “Normal”.
- Graceful handling of missing data and empty histories.
- Caching for efficient performance.

- **User Interface Highlights:**

    - Dropdown selection for cryptocurrency.
    - Interactive sliders and inputs for manual mode.
    - Visual feedback via metrics and line charts.

Limitations

- **Volatility Proxy:** GK is a simplified daily estimator; intraday realized volatility could improve fidelity.
- **Data Gaps:** Sparse histories for some assets may affect threshold stability.
- **Regime Shifts:** Volatility behavior changes over time; periodic retraining is recommended.

Future Work

- **Feature Expansion:** Incorporate ATR, liquidity depth, realized volatility, and regime indicators.
- **Modeling Enhancements:** Explore monotonic constraints, quantile regression, and ensemble methods.
- **Evaluation:** Implement rolling-window backtesting and dynamic threshold calibration.
- **Deployment Extensions:** Batch scoring, REST API, and dashboard for multi-asset comparison.