BIG DATA ANALYSIS AND MODELING OF BINANCE TRADING DATA USING SPARK

Group 5
- Victor Dumaslan, Dongmei Han, Niranjan Rao, Akshit Tyagi, Chao Zheng

Course: DATA-228 Sec 24 & 72

Date: May 10, 2025

INTRODUCTION

WHY THIS PROJECT MATTERS

- **The Problem:**
  - Cryptocurrency markets are fast, volatile, and data-rich — but difficult to analyze at scale.
- **Opportunity:**
  - Historical trading data offers deep insights into market behavior, but requires powerful tools to process.
- **Our Goal:**
  - Build a scalable pipeline to explore crypto market dynamics using real historical data and open-source tools.
- **What We Do:**
  - From data storage and processing to ML modeling and interactive dashboards.

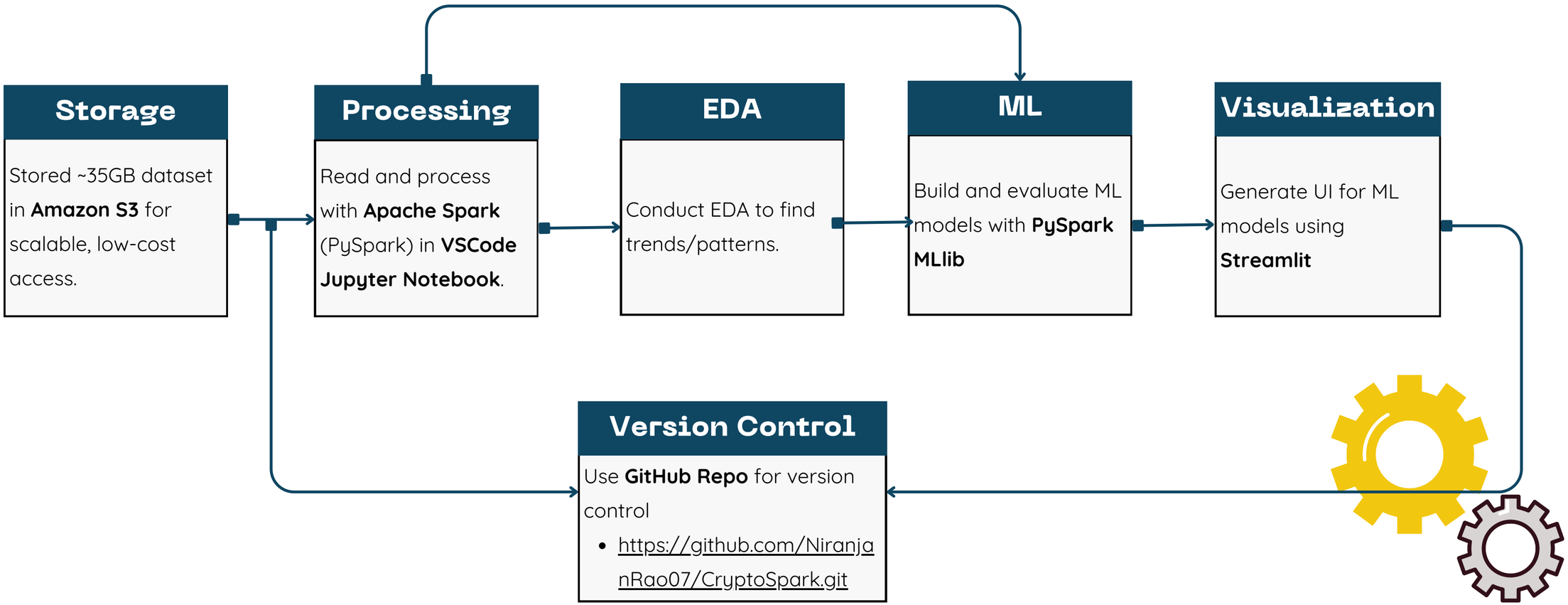
DATASET OVERVIEW

```
root
 |-- open: float (nullable = true)
 |-- high: float (nullable = true)
 |-- low: float (nullable = true)
 |-- close: float (nullable = true)
 |-- volume: float (nullable = true)
 |-- quote_asset_volume: float (nullable = true)
 |-- number_of_trades: integer (nullable = true)
 |-- taker_buy_base_asset_volume: float (nullable = true)
 |-- taker_buy_quote_asset_volume: float (nullable = true)
 |-- open_time: timestamp_ntz (nullable = true)
 |-- symbol: string (nullable = false)
 |-- date: date (nullable = true)
```

- **Source:**
  - Kaggle dataset (*https://www.kaggle.com/datasets/jorijnsmit/binance-full-history*)
- **Size:**
  - ~35GB across 1,000 cryptocurrency trading pairs.
- **Storage Format:**
  - Partitioned Parquet files stored on Amazon S3.
- **Granularity:**
  - 1-minute candlestick data from July 2017 to Nov 2022.

```
+-------------------+-------------------+
|         start_date|           end_date|
+-------------------+-------------------+
|2017-07-14 04:00:00|2022-11-17 22:22:00|
+-------------------+-------------------+
```

METHODOLOGY & TOOLS



KEY INSIGHTS FROM

EDA

How and Why We Chose LUNA-USDT for Analysis

- Over **1.5 billion** rows, **1000** trading pairs
- Focus on **active markets** with significant trading behavior

summary	open	high	low
count	1507350137	1507350137	1507350137
mean	2668804.5989923603	2668927.6850018147	2668663.7415869674
stddev	3.2131753575421417E10	3.213175357760682E10	3.213175357291572E10
min	1.0E-8	1.0E-8	1.0E-8
max	4.76802774E14	4.76802774E14	4.76802774E14

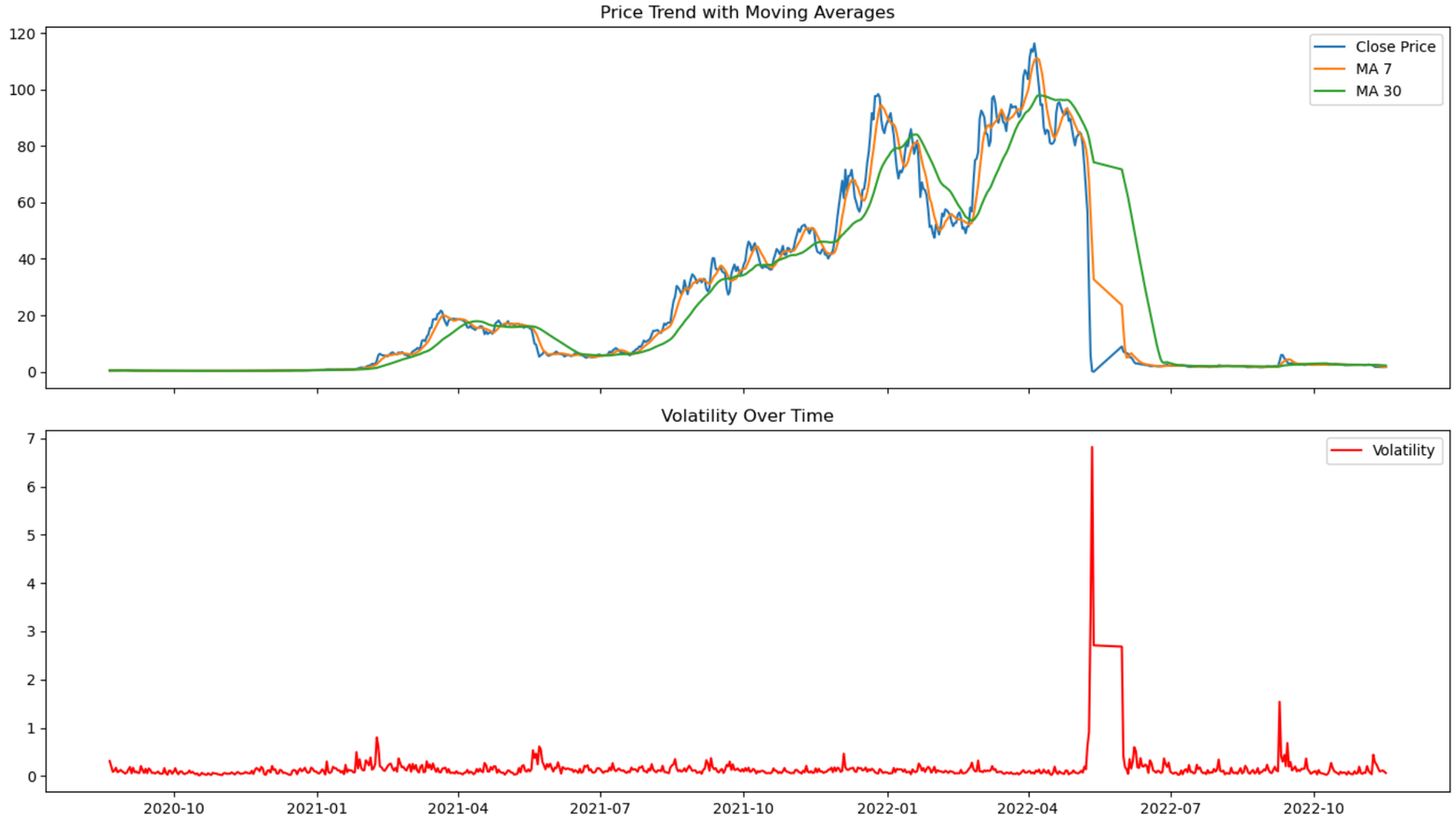
- **Step 1:** Aggregate to daily level (group by symbol and date, compute avg/sum)

- **Step 2:** Filter top 3 symbols by average daily volume and price
  - Volume > 1,000,000
  - Price > $10
  - Exclude illiquid or irrelevant tokens

symbol	avg_volume	avg_price
LUNA-BUSD	1.043141477942703...	23.692329510184745
LUNA-USDT	5.352539523832988E8	24.127279731496017
SUSHIUP-USDT	6.818181314479403E7	16.152223319382767

- **Step 3:** **LUNA-USDT** selected for its **high volume** and **price stability.**

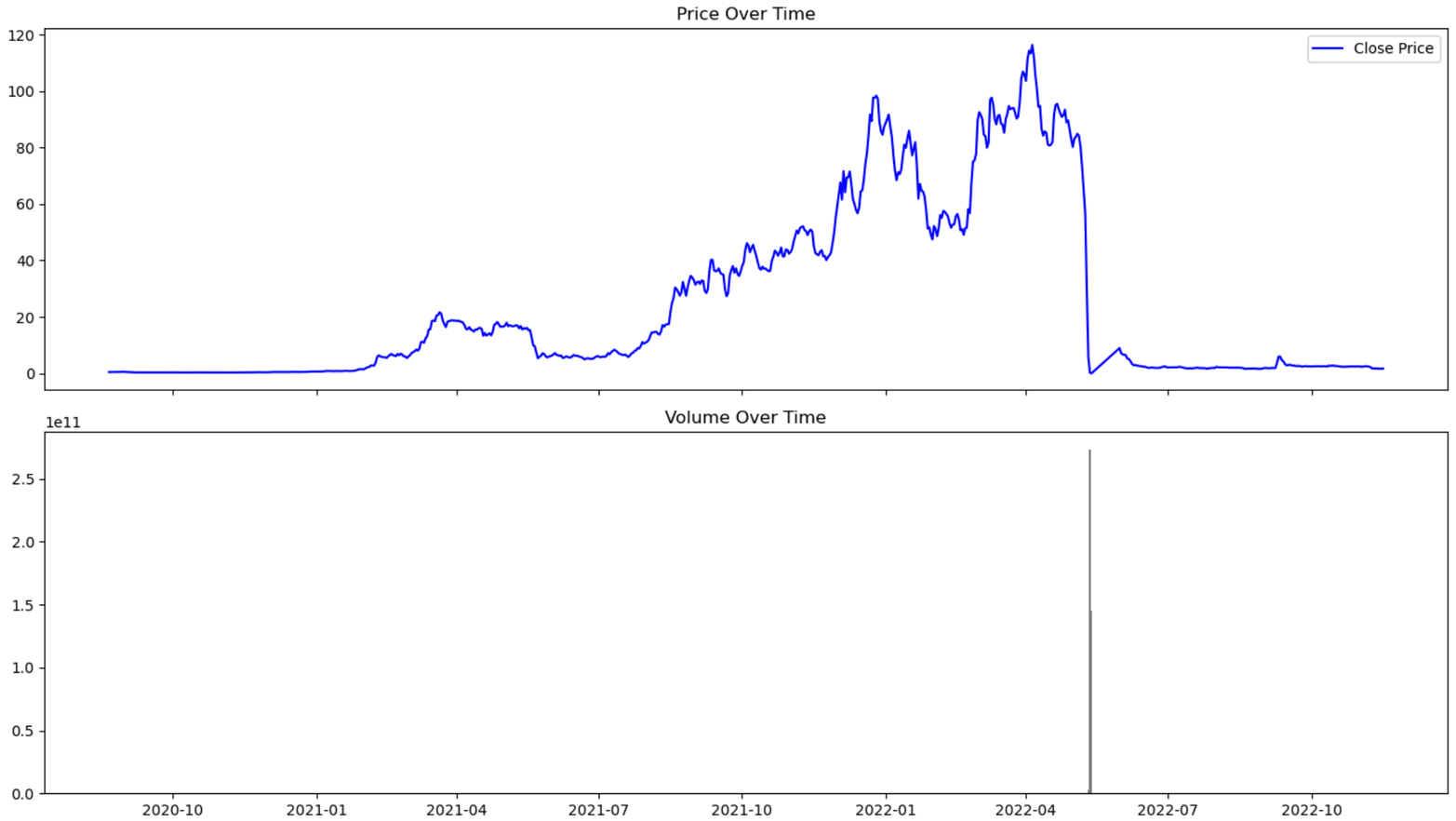
Price & Volatility Analysis of LUNA-USDT



KEY INSIGHTS FROM EDA

- Strong upward trends in 2021 followed by **a dramatic crash in mid 2022**.
- **Moving averages** confirm rising momentum before the crash and total loss of trend after.
- **Volatility spike** marks market panic during collapse.
- Suggests LUNA-USDT experienced a **boom-bust cycle** with little post-crash recovery.

Volume Patterns & Price Behavior of LUNA-USDT

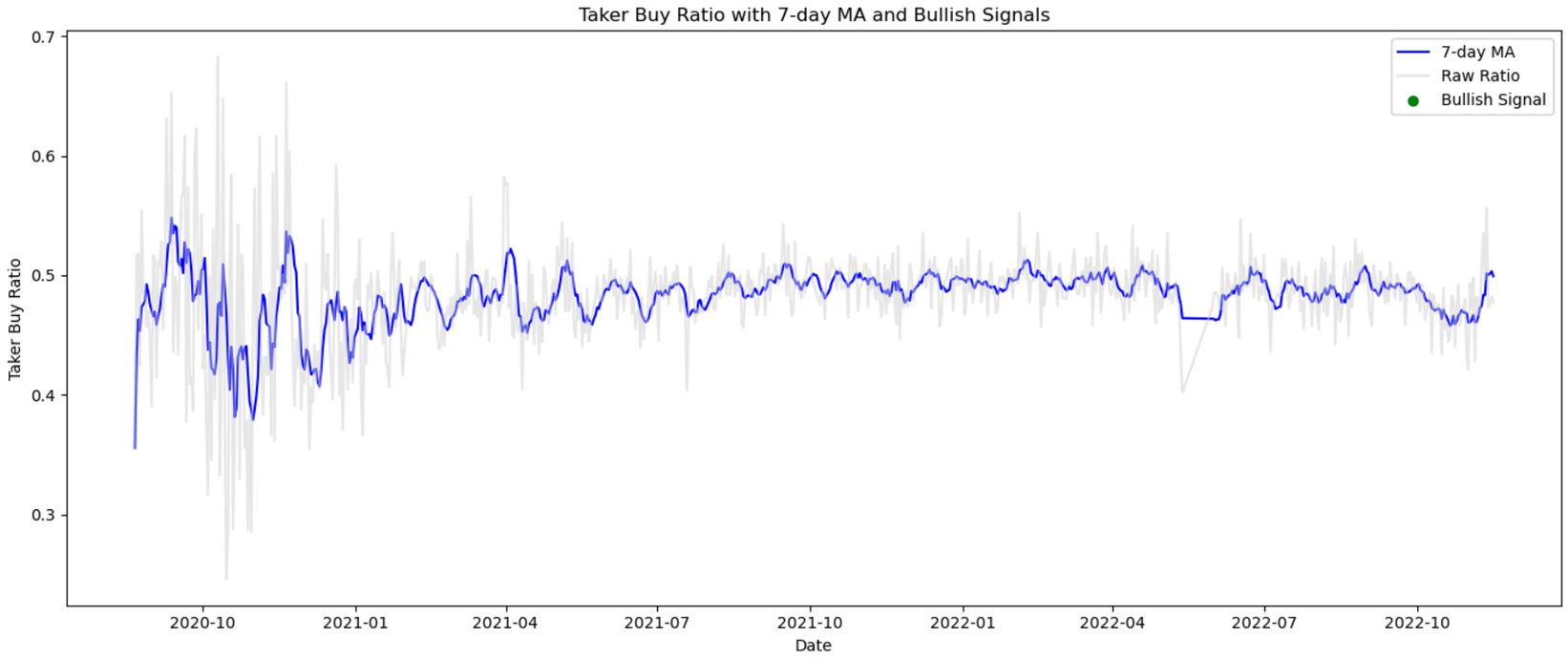


KEY INSIGHTS FROM **EDA**

- **High volume days often align with price drops**, suggesting panic selling or sell-side pressure.
- No strong volume buildup before price increases → not a typical bullish volume profile.

KEY INSIGHTS FROM EDA

Taker Buy Ratio & Market Sentiment of LUNA-USDT



- No bullish markers → **Buy pressure didn't align with price increases**
- 7-day average shows **sporadic buyer activity,** not sustained momentum
- Suggests **false signals** or **weak follow-through** from buyers

PROCESSING & CLEANING DATA

- Dropped unnecessary columns: quote_asset_volume, number_of_trades.
- Removed rows with missing or invalid values.
- Ensured consistency and reliability in core trading metrics.

FEATURE ENGINEERING

We engineered new features to enhance the dataset:

- Daily Return - Captures day-to-day price movement
- Volatility - Measures intraday price range
- Moving averages (7 day & 30 day)
- Cumulative Return

In [7]:
```
# Add additional features, like daily return and volatility
sample_df = sample_df.withColumn("daily_return", (col("close") - col("open")) / col("open")) \
       .withColumn("volatility", (col("high") - col("low")) / col("open"))
```

In [8]:
```
sample_df = sample_df.repartition("symbol")

# Cumulative return
window_spec = Window.partitionBy("symbol").orderBy("date") \
    .rowsBetween(Window.unboundedPreceding, Window.currentRow)

# Moving Averages (7-day and 30-day)
window_7 = Window.partitionBy("symbol").orderBy("date").rowsBetween(-6, 0)
window_30 = Window.partitionBy("symbol").orderBy("date").rowsBetween(-29, 0)

# Calculate cumulative return and moving averages
sample_df = sample_df.withColumn("ma_7", avg("close").over(window_7)) \
    .withColumn("ma_30", avg("close").over(window_30)) \
    .withColumn("cumulative_return", exp(_sum(log(1 + col("daily_return"))).over(window_spec)) - 1)
```

In [9]:
```
# Drop rows with nulls in critical columns
sample_df = sample_df.dropna(subset=["ma_7", "ma_30", "cumulative_return", "daily_return", "volatility"])
```

ML MODEL & EVALUATION

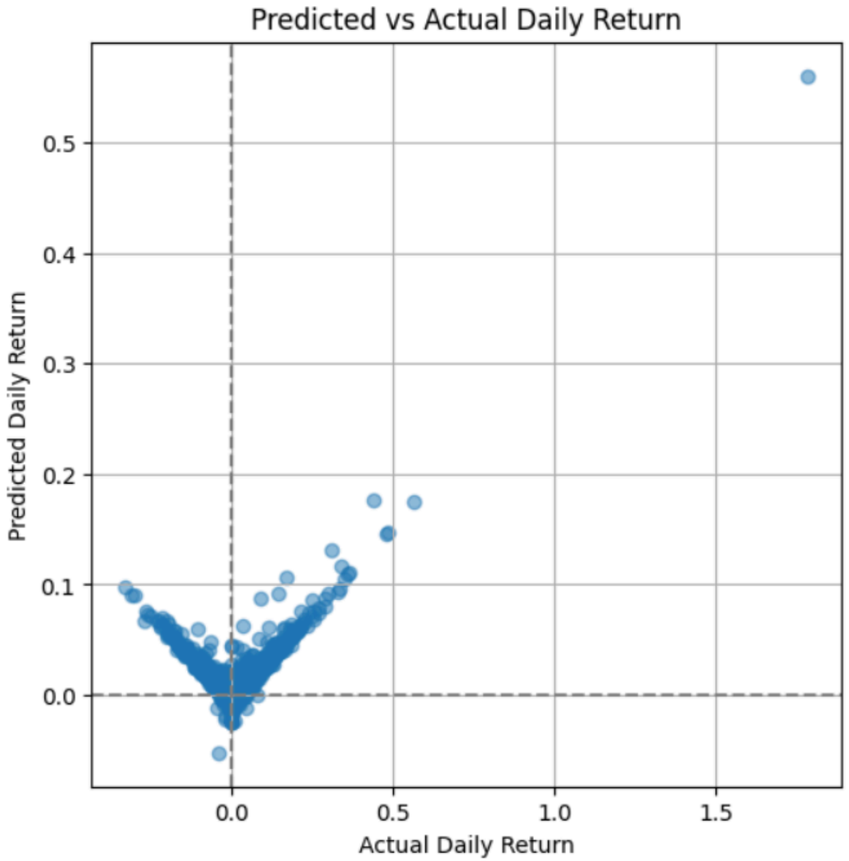
- Goal: Predict or classify crypto asset behavior using historical minute-level data
- **Models used:**
  - Linear Regression
  - Logistic Regression
  - XGBoost
- **Features used:**
  - Price stats: Open, High, Low, Close, Volume
  - Engineered: Daily Return, Volatility, MA(7), MA(30), Cumulative Return

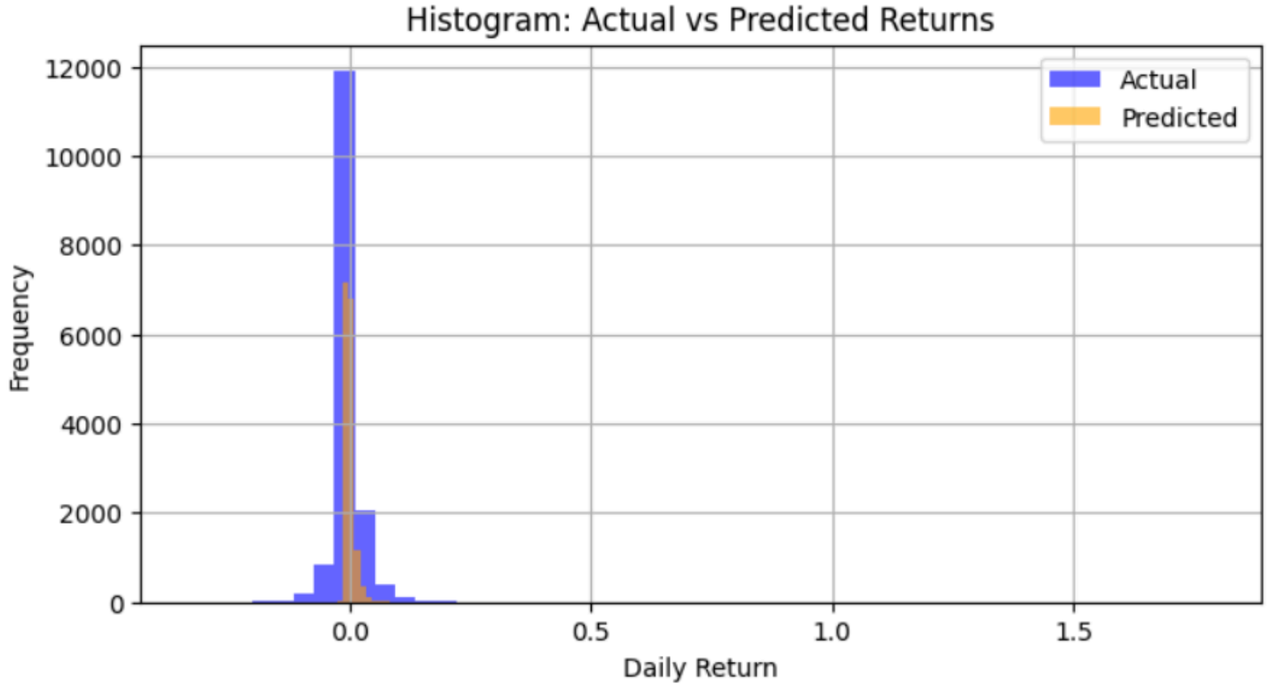
Linear Regression Model

Evaluation Metrics

RMSE: 0.0339

R2 Score: 0.0922 → *Very Low*





Conclusion: Linear regression fails to model non-linear and complex market behavior.

- Poor fit
- Histogram has wider distribution than predictions
- Scatter plot shows prediction bias toward zero

Logistic Regression Model

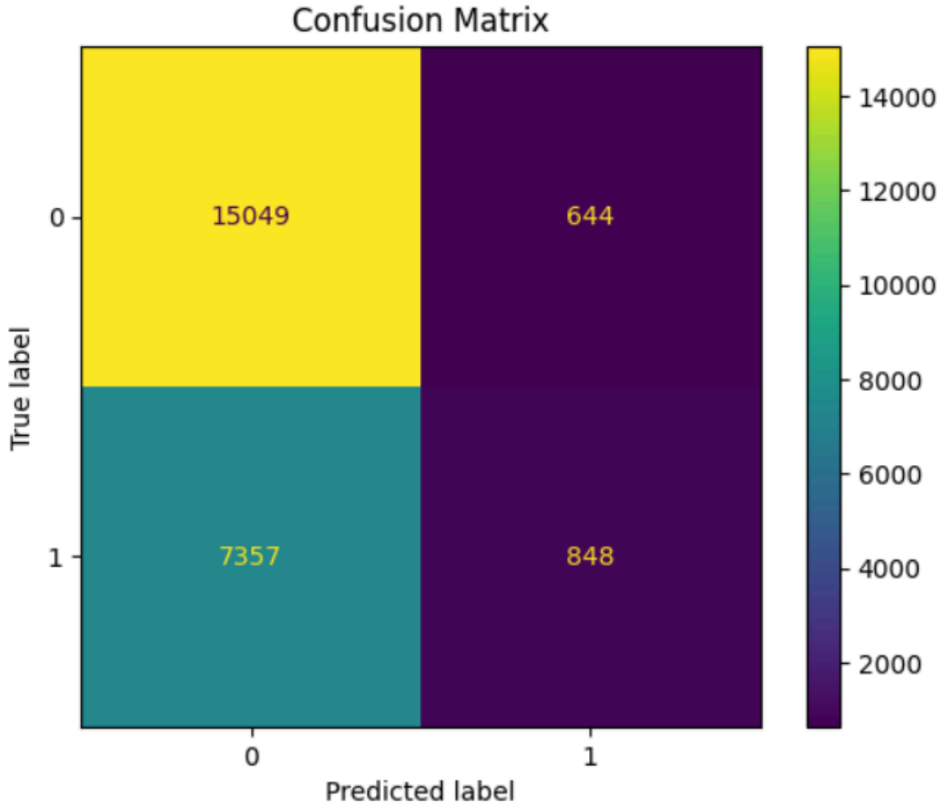
Evaluation Metrics

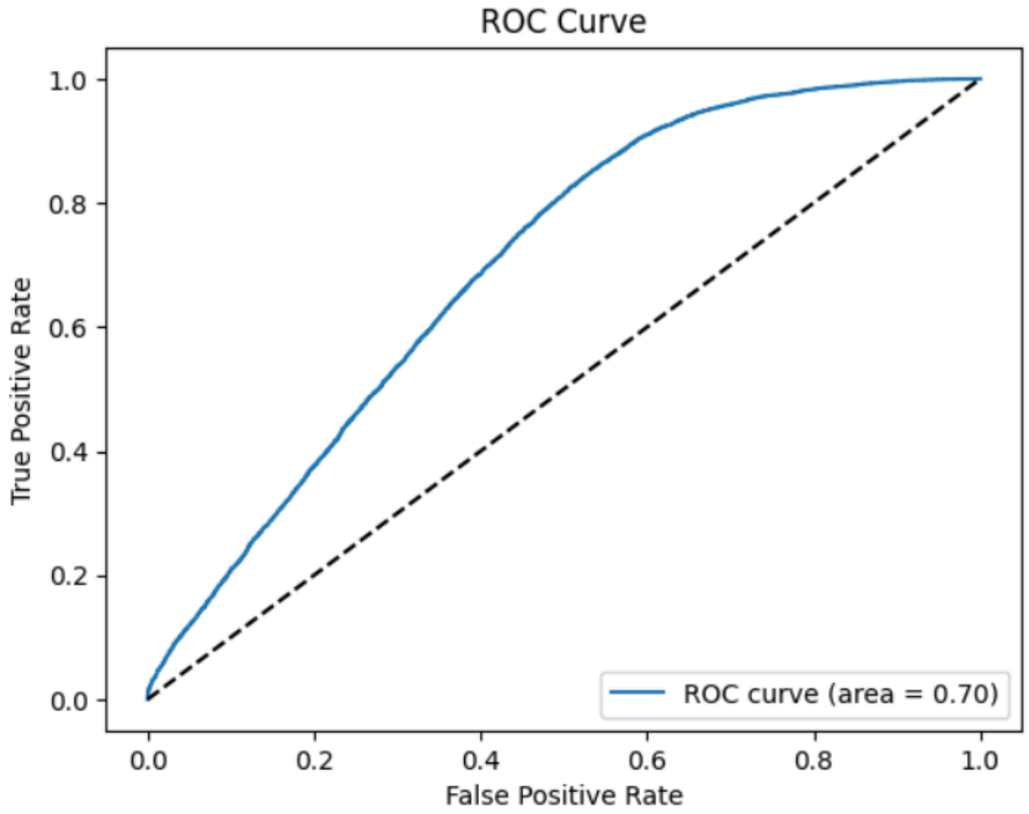
Accuracy: 0.6671 **ROC-AUC:** 0.6984

Precision: 0.6353

Recall: 0.6671

F1 Score: 0.5805





Conclusion: Logistic Regression is a reasonable baseline model, but it still struggles

- Class 0 (negative return)
- Class 1 (positive return)
- ROC Curve indicates moderate classification performance

ML MODEL & EVALUATION

XGBoost Model

- Dealing with large-scale crypto time series data
- XGBoost Model
- Features Used
  - Price stats: Open, High, Low, Close, Volume
  - Engineered: Daily Return, Volatility, MA(7), MA(30)
  - Shifted the target column to create labels(lead)

ML MODEL & EVALUATION

XGBoost Model

demo link

Limitations:

- May overfit without proper regularization
- Sensitive to outliers and extreme volatility.
- Difficult to interpret compared to simpler models.

DEMO

FUTURE WORKS

- Address Model Overfitting & Misleading Predictions – In future iterations, we aim to enhance model generalization by experimenting with regularization, advance assemble tuning to minimize overfitting & improve reliability.
- Improve Interoperability – Integrate SHAP or LIME to make model predictions more transparent and explainable
- Deploy Real-time prediction pipeline – Extend the batch pipeline into a streaming solution allowing predictions on live crypto feeds using Spark Structured Streaming
- Launch on a multi-node Spark cluster (Databricks, EMR, or Kubernetes) to speed up EDA, feature engineering, and retraining.

CONCLUSION

- Performed scalable EDA on binance data.
- Executed data preprocessing and structured feature engineering for enhanced analysis.
- Trained and evaluated linear regression, logistic regression and XGBoost .
- Developed a fully functional UI to predict the daily return of a cryptocurrency using technical indicators.

THANK YOU