Report on Exploratory and Predictive Study of Trade-Sentiment Data

Introduction

The uploaded notebook focuses on combining trading data with market sentiment information to uncover insights and build predictive models. Two main datasets were used:

1. **Historical trading data** (historical_data.csv), containing information on trades such as timestamps, execution prices, trade sizes, and profit/loss.
2. **Fear and Greed index data** (fear_greed_index.csv), providing daily sentiment levels to represent overall market psychology.

The primary aim was to merge these datasets, perform cleaning and transformation for analysis, explore patterns in trading behavior relative to sentiment, and finally, construct predictive models for both classification (sentiment prediction) and regression (execution price prediction).

Dataset Preparation, Cleaning and Transformation

Loading and Parsing:

- Trading data and sentiment index files were read from CSVs.
- **Date and time parsing** was performed: the trading timestamp (Timestamp IST) and sentiment date were converted to datetime objects to support merging and time-series analysis.

Cleaning:

- Redundant and irrelevant columns were dropped, including identifiers like key_0, Trade ID, Timestamp IST, and unused descriptors such as Coin, Account, and Direction.
- After processing, the datasets were merged on the day field, producing a combined dataset with **14,149 rows and 17 columns**.

Transformations and Feature Engineering:

- Rolling statistics were created to capture short-term patterns:
    - 7-day rolling mean and rolling standard deviation (volatility) of execution price.
    - 7-day rolling averages for Execution Price and Size USD.
- Forecasting proxies were added by scaling rolling averages with sentiment index values.
- Derived features included Price_Change (day-to-day price difference) and Rolling_Price_Mean.
- Encodings were applied:

    - Sentiment levels were label-encoded for classification tasks.
    - Categorical variables such as Side and Crossed were encoded for regression modeling.
- Interaction features combined categorical encodings with sentiment values, e.g., Side_Sentiment and Size_Sentiment.
- A cumulative size metric (Cumulative_Size_USD) was listed among regression features, indicating cumulative exposure over time.

The processed dataset was exported as a CSV for downstream use.

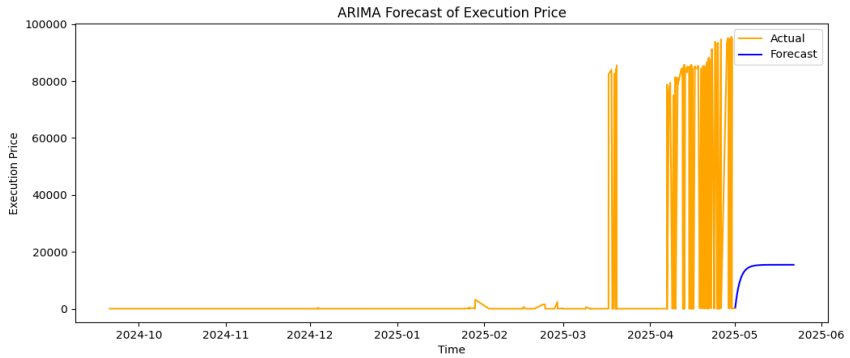
Exploratory Analysis and Insights

Sentiment Distribution:

A bar chart revealed the frequency of various sentiment categories. Results indicated an imbalance ie.., certain sentiment classes (notably "neutral/greed") were highly dominant, while others had very few samples (some as low as 3 entries).

Time-based Trends:

- **Weekly trend analysis** showed variations in trade sizes and volumes across different calendar weeks.
- **Daily trend plots** highlighted fluctuations in trading activity by date.



- **Hourly trends** illustrated intraday seasonality, identifying the hours with the highest transaction volumes.

Rolling Volatility:

The 7-day rolling volatility of execution price highlighted periods of high uncertainty, allowing comparison between calmer and more volatile trading phases.

Statistical Tests:

An **Augmented Dickey–Fuller (ADF) test** on execution prices reported:

- ADF Statistic = –7.81
- p-value ≈ 7.2e–12

This strong result indicated that the price series could be treated as stationary, validating the use of ARIMA for time-series modeling.

Risk – Reward Analysis:

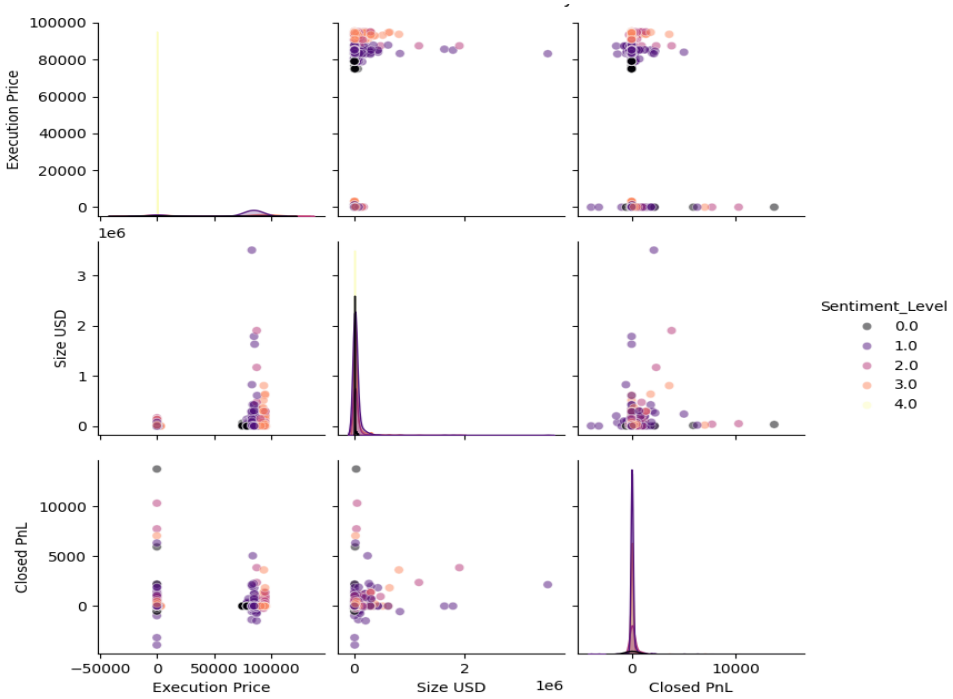
Risk-reward ratios were computed per sentiment level, by comparing the mean profit of winning trades to the average loss of losing trades. This provided a measure of how sentiment aligned with payoff efficiency.

Predictive Modeling

Random Forest Classifier - Predicting Sentiment:

A Random Forest model was built to predict sentiment categories using only three features: Execution Price, Size USD, and Closed PnL.

- **Performance:**
  - Overall accuracy on the held-out test set (30%) was **97%**.
  - Precision and recall were very high for dominant sentiment classes (near 98%).



  - However, extreme imbalance was observed: minority classes, such as one with only 3 samples, achieved **0 precision/recall**.

- **Feature Importance:**
  The model identified Size USD and Execution Price as the most influential predictors for sentiment levels.

This model demonstrates strong performance on frequent classes but is biased toward majority categories

XGBoost Regressor — Predicting Execution Price:

An XGBoost regression model was developed with a richer feature set, including:

- Encoded categorical features (Side_enc, Crossed_enc)
- Sentiment variables (Sentiment_Level, Side_Sentiment, Size_Sentiment)
- Rolling features (Rolling_Price_Mean)
- Cumulative and raw trade metrics (Size USD, Size Tokens, Cumulative_Size_USD, Price_Change)
- **Evaluation Metrics:**
    - Mean Squared Error (MSE): **2,033,374**
    - R^2 Score: **0.9990**

The very high R^2 suggests near-perfect fit, though this could also indicate overfitting.

- **Feature Importance:**
  The most influential variables included rolling averages and cumulative trade size, showing that both short-term dynamics and accumulated volume drive execution price prediction.

Conclusion

The notebook successfully integrated trading data with sentiment indicators to produce a robust analytical pipeline:

1. **Data Integration & Cleaning:** Two heterogeneous datasets were merged after rigorous cleaning and timestamp alignment.
2. **Feature Engineering:** Rolling statistics, cumulative exposure, sentiment encodings, and interaction terms enriched the dataset for modeling.
3. **Exploratory Analysis:** Distribution plots, volatility measures, and risk–reward ratios revealed strong patterns in trading behavior and market sentiment.
4. **Predictive Modeling:**
    - The **Random Forest classifier** achieved high accuracy for sentiment prediction but suffered from class imbalance.
    - The **XGBoost regressor** reported near-perfect R^2 for price prediction, highlighting the predictive strength of engineered features, though possibly overfit.

Overall, the notebook demonstrates how combining sentiment data with trade-level detail can yield high-performing predictive models.