Junior Data Scientist – Trader Behaviour Insights

Name: Neel Patel

Internship: Data Science Internship

Company: Primetrade.ai

Date: 25/09/2025

Email id: patelneel0742@gmail.com

Table of Contents

1. Abstract

- This project analyses cryptocurrency trading data to understand trade patterns, profitability, and associated factors.
- The work includes data cleaning, visualization, exploratory data analysis (EDA), and development of insights through statistical and ML techniques.

2. Objective

- Analyse how trading behaviour (profitability, risk, volume, leverage) aligns or diverges from overall market sentiment (fear vs greed).
- Identify hidden trends or signals that could influence smarter trading strategies.

3. Introduction

- To identify profitable trade patterns.
- To evaluate the impact of fees and execution price on profitability.
- To visualize trends across timestamps.

4. Dataset Description

The datasets were provided as part of the internship assignment by the hiring team. They were accessed via Google Drive links shared in the task brief.

For this project the two datasets that were used are:

4.1) Bitcoin Market Sentiment Dataset(fear_greed_index.csv)

4.2) Historical Trader Data from Hyperliquid (historical_data.csv)

4.1. Fear & Greed Index Dataset (fear_greed_index.csv)

- Source: Provided as part of the internship resources (Bitcoin Market Sentiment data).
- Size: 2000+ rows × 4 columns.
- Columns:
    i. timestamp → Unix timestamp of the data record.
    ii. value → Numeric index value representing market sentiment.
    iii. classification → Sentiment category (e.g. Extreme fear, Fear, Neutral, Greed, Extreme Greed).
    iv. date → Human-readable date format.
- Data Quality:
    1. No missing values across columns.

2. No duplicates detected.

- Preprocessing:
    1. Converted timestamp to datetime format for time-series analysis.
    2. Ensured categorical values in classification were standardized.

4.2. Historical Trader Data (historical_data.csv)

- Source: Provided by internship as trade-level transactional data.
- Size: 2,00,000+ rows × 17 columns
- Columns:
    i. Account → Trader's account identifier.
    ii. Coin → Cryptocurrency traded.
    iii. Execution Price → Price at which the trade was executed.
    iv. Size Tokens → Trade size in tokens.
    v. Size USD → Trade size in USD value.
    vi. Side → Buy/Sell indicator.
    vii. Timestamp IST → Trade execution time in IST.
    viii. Start Position → Trader's position at start of trade.
    ix. Direction → Direction of the trade.
    x. Closed PnL → Profit or Loss from trade closure.
    xi. Transaction Hash → Unique transaction identifier.
    xii. Order ID → Order identifier.
    xiii. Crossed → Indicates if trade was crossed.
    xiv. Fee → Transaction fee charged.
    xv. Trade ID → Trade identifier.
    xvi. Timestamp → Trade execution time.

Data Quality:

1. No missing values (all 211,224 rows have complete entries).
2. No obvious duplicates at column level.

Preprocessing:

1. Converted Timestamp and Timestamp IST to proper datetime objects for analysis.
2. Created derived feature is_profitable (based on Closed PnL > 0).

3. Standardized column naming conventions (e.g., unified timestamp formats).

5. Methodology

This project followed a structured data science workflow consisting of four major steps:

5.1 Data Cleaning:

- Converted timestamps (timestamp, Timestamp IST, Timestamp) into proper datetime format.
- Verified no missing values (both datasets had complete entries).
- Removed duplicates based on transaction hash and order ID.
- Ensured consistent datatypes across numeric and categorical variables.

5.2 Exploratory Data Analysis:

1) Sentiment vs. Profitability:

- Average Closed PnL was highest during *Extreme Greed* sentiment (~64.1) and lowest during *Neutral* (~32.5).
- Profitability ratio: trades during *Extreme Greed* had ~46% profitable vs. 54% loss-making.

2) Visual Insights:

- Line charts showed Fear & Greed Index fluctuates heavily, aligning with trading spikes.
- Bar charts revealed imbalance in profitable vs. non-profitable trades across sentiment classes.
- Scatterplots indicated larger USD trade sizes contributed disproportionately to profit variability.

3) Daily Aggregation:

- Correlation between daily PnL and sentiment value was weak (same day ≈ 0.027, lagged sentiment ≈ 0.004).
- Suggests sentiment alone is not a strong predictor of profitability.

5.3 Feature Engineering:

- Created is_profitable flag (1 if Closed PnL > 0, else 0).
- Derived Side_flag (Buy = 1, Sell = 0).
- Calculated Fee % impact = Fee / Size USD * 100 for transaction cost analysis.

5.4 Insights and Modeling:

(a) Unsupervised Clustering

- Applied KMeans clustering (k=3) on features: Closed PnL, Size USD, value, Side_flag.
- Cluster characteristics:
    - Cluster 0: Low PnL (~22.9), small trades (~5.2k USD).
    - Cluster 1: Moderate PnL (~47.7), mid-size trades (~4.8k USD).
    - Cluster 2: Very high PnL (~16,769), large trades (~679k USD).
- Insight: High profits were concentrated among fewer large trades.

(b) Supervised Machine Learning

Two models were applied to predict is_profitable:

i. Logistic Regression

- Accuracy: 63%
- ROC-AUC: 0.64
- Moderate performance, struggled with imbalanced classes.

ii. Random Forest Classifier

- Accuracy: 80%
- ROC-AUC: 0.89
- Significantly better predictive performance.
- Feature Importances:
    - Fee (38.3%)
    - Size USD (35.8%)
    - Sentiment value (18.6%)
    - Side_flag (7.2%)
- Insight: Transaction fees and trade size were the most important factors driving profitability, with sentiment playing a secondary role.

6. Result

6.1 Exploratory Data Analysis (EDA) Insights:

- Most traded coin**:** The dataset shows that the majority of trades were concentrated in **HYPE, @107, and BTC**, which together accounted for nearly 60% of all trades.
- Profitability vs. Sentiment:
    - Average Closed PnL was highest during Extreme Greed (≈ 64) and lowest under Neutral (≈ 32).
    - Profitability ratio shows that “Fear” and “Extreme Fear” phases still yielded ~41% profitable trades, while “Neutral” conditions had the lowest win rate (~36%).

- Direction impact: BUY/SELL side had different effects on profitability, captured through Side_flag. The feature importance analysis (Random Forest) confirmed that fees and trade size had the strongest impact on profitability.
- Time-based patterns**:** Trades displayed clear peaks during 19:00–21:00 IST, coinciding with overlaps of European and U.S. trading sessions. Secondary activity spikes were observed during 01:00–04:00 IST, reflecting after-hours trading or arbitrage opportunities. In contrast, activity was lowest during mid-day (11:00–12:00 IST), suggesting reduced market engagement in local hours.
- Clustering analysis:
    - KMeans identified 3 trade clusters:
        - Cluster 0 → Small trades with low average PnL.
        - Cluster 1 → Medium trades with moderate PnL.
        - Cluster 2 → High-value trades (large USD size) driving extreme PnL swings.

6.2 Modeling Results:

- Logistic Regression:
    - Accuracy ≈ 63%, ROC-AUC ≈ 0.64.
    - Model captured some signal but struggled to generalize well.
- Random Forest:
    - Accuracy ≈ 80%, ROC-AUC ≈ 0.89.
    - Significantly better than Logistic Regression, confirming non-linear relationships matter.
    - Feature importance ranking:
        1. Fee (38%)
        2. Size USD (36%)
        3. Sentiment Value (19%)
        4. Trade Direction (7%)

This suggests fees and trade sizes are the strongest determinants of profitability, overshadowing even market sentiment.