Data Science Report: Trader Behavior vs Bitcoin Market Sentiment

Candidate: Raja Prabu Manivel
Date: September 2025

1. Objective

The objective of this analysis is to explore and understand the relationship between trader behavior and Bitcoin market sentiment using historical trading data and sentiment indicators. Specifically, we aim to analyze how trading behavior variables such as profitability (Closed PnL), trade volume (Size USD, Size Tokens), leverage, and trade direction align or diverge from overall market sentiment (Fear vs Greed). By identifying patterns and building predictive models, we can extract actionable insights to support smarter trading strategies.

2. Datasets

2.1 Bitcoin Market Sentiment Dataset - Columns: Date, Classification (Fear / Greed) - This dataset provides daily Bitcoin market sentiment ratings from February 2018 to May 2025. It categorizes the market into Fear, Neutral, and Greed phases, offering a high-level perspective of market psychology that can impact trader behavior.

2.2 Historical Trader Data (Hyperliquid) - Columns include: Account, Coin, Execution Price, Size Tokens, Size USD, Side, Timestamp IST, Start Position, Closed PnL, Order ID, Crossed, Fee, Trade ID, and other direction-based boolean flags. - This dataset contains over 200,000 historical trades across multiple coins from 2018 to 2025. It provides granular transaction-level insights into trader actions, profitability, and risk exposure.

3. Data Preprocessing

3.1 Merging Datasets - The two datasets were merged on the date column to align trader behavior with corresponding market sentiment. - Ensured that the date ranges match and filtered data to include relevant overlapping periods.

3.2 Handling Missing Values - Only a few rows contained missing values (classification and value). These rows were removed to ensure data integrity.

3.3 Outlier Handling - Numerical columns such as Closed PnL, Execution Price, Size Tokens, and Size USD were checked for outliers using the IQR method. - Outliers were imputed using the median to prevent skewing model performance while maintaining dataset size.

3.4 Feature Engineering - Categorical variables (Side, Dir_*) were one-hot encoded. - High-cardinality columns like Account and Coin were dropped for modeling due to overfitting risk. - Continuous numerical columns were standardized, while boolean features were kept unscaled.

3.5 Train/Test Split - The data was split into training and test sets (80/20) for model evaluation.

4. Exploratory Data Analysis (EDA)

4.1 Market Sentiment Distribution - Visualizations showed that Greed periods were more frequent than Extreme Fear, reflecting market optimism.

4.2 Trade Volume & Profitability - Boxplots and scatterplots revealed that traders tend to execute larger volume trades during Greed phases. - Closed PnL distribution showed some extreme profit/loss values, handled via median imputation.

4.3 Correlation Analysis - Heatmaps of numerical features indicated moderate correlation between trade size (Size USD) and profitability (Closed PnL). - Boolean trade direction flags were influential in predicting sentiment.

4.4 Confusion Matrix Visualization - Confusion matrices for Decision Tree and Random Forest models were plotted and saved to the outputs folder.

5. Modeling

5.1 Logistic Regression - Multiclass classification with market sentiment as target. - Performance was poor (train accuracy ~0.29), likely due to non-linear relationships and dominance of categorical features.

5.2 Decision Tree Classifier (DTC) - Initial depth of 5 yielded test accuracy ~0.82. - Adjusted depth to 8–10 improved test accuracy to ~0.91. - Captured non-linear patterns and interactions among features effectively.

5.3 Random Forest Classifier (RFC) - Hyperparameters: max_depth=15, n_estimators=200. - Achieved train accuracy ~0.93 and test accuracy ~0.91, demonstrating strong generalization. - RFC handled feature interactions better than DTC and reduced overfitting risk.

6. Feature Importance

- Order ID, Size USD, and direction flags like Dir_Sell, Dir_Open Long had the highest predictive value.
- Continuous features such as Execution Price and Closed PnL contributed moderately.
- Visualizations of feature importance were saved in the outputs folder.

7. Insights & Recommendations

- Traders are generally more profitable during Greed periods and exhibit higher risk-taking during Fear periods.
- Tree-based models (DTC, RFC) are suitable for predicting sentiment due to their ability to capture non-linear interactions.
- Logistic Regression is not recommended for this dataset due to poor performance.
- Future improvements: include temporal features, sequence modeling of trades, and encoding high-cardinality categorical features like Coin.

---

8. Limitations

- High-cardinality categorical features were dropped, which may omit coin-specific insights.
- Sentiment data is aggregated at daily frequency; intra-day sentiment effects are not captured.
- Market conditions change rapidly, so historical patterns may not always predict future behavior.

---

9. Final Conclusion

After extensive preprocessing, EDA, feature engineering, and modeling, the analysis demonstrates that trader behavior is strongly influenced by market sentiment. Tree-based models such as DTC and RFC achieve high predictive accuracy (~91%) and effectively capture non-linear relationships between features and sentiment. Logistic Regression, while interpretable, underperforms due to the complexity of interactions.

Recommendations: - Prefer Decision Tree and Random Forest models for predictive analysis. - Focus on key features like trade size, order direction flags, and fees for insights. - Explore temporal analysis and sequential trade modeling for enhanced predictive capabilities.

> Overall, this project provides actionable insights into trader behavior relative to market sentiment and highlights the importance of tree-based models for accurate classification and strategy optimization.

---

10. Submission Files

- notebook_1.ipynb (primary analysis)
- csv_files/merged_data.csv (preprocessed merged dataset)
- outputs/ (visualizations and plots)
- ds_report.pdf (this document)
- README.md (setup instructions and notes)

Colab Link:

https://drive.google.com/drive/folders/1xL19tJkYeLalaJcEdqRs103if_Oit_tH?usp=drive_link