Trade and Sentiment Data Analysis Report

Overview

This notebook analyzes historical trade data and cryptocurrency market sentiment to identify potential relationships between sentiment and trading outcomes.

Data Loading and Preparation

The analysis utilizes two datasets:

- historical_data.csv: Contains detailed information on individual trades, including execution price, size, direction, and profit/loss.
- fear_greed_index.csv: Contains the Fear & Greed Index values and classifications over time.

Data loading was performed using pandas, and initial inspection included displaying the head and descriptive statistics of both dataframes (tradedf and sentidf). Missing values were identified in tradedf, and several columns deemed irrelevant for the analysis ('Account', 'Transaction Hash', 'Trade ID', 'Order ID') were dropped.

Timestamps in both dataframes were converted to datetime objects and a 'Date' column was created to facilitate merging.

Data Merging

The two datasets were merged based on the 'Date' column, creating a merged_df containing trade data linked to the corresponding daily sentiment.

Outlier Treatment

An IQR-based method was applied to remove outliers from key numerical columns ('Execution Price', 'Size Tokens', 'Size USD', 'Start Position', 'Closed PnL', 'Fee', 'value') in the merged dataframe. This step aimed to improve the robustness of subsequent analysis and modeling.

Data Filtering

Coins with fewer than 300 trades after outlier removal were dropped to focus on cryptocurrencies with sufficient data for meaningful analysis. This resulted in keeping only two coins: BTC and ETH.

Exploratory Data Analysis (EDA)

EDA included:

- **Distribution of Closed PnL:** A histogram showed the distribution of profit and loss after outlier removal.
- **Closed PnL by Coin:** A boxplot and calculation of the average Closed PnL per coin revealed that BTC had a slightly higher average profit compared to ETH.
- **Average Features by Classification:** Bar plots visualized the average 'Closed PnL', 'Execution Price', 'Start Position', 'Size Tokens', and 'Size USD' across different sentiment classifications.
- **Number of Trades per Classification by Coin:** A countplot illustrated the distribution of trades within each sentiment classification for BTC and ETH.

Feature Engineering and Preparation for Modeling

- The 'classification' column was custom-encoded numerically based on the sentiment order (Extreme Fear to Extreme Greed).
- The 'Coin' column was encoded numerically using category codes.
- The 'timestamp' was converted to a numerical representation.
- Irrelevant columns ('Coin', 'timestamp', 'classification') were dropped from the dataframe used for modeling (df_filtered).
- A log transformation (log_pnl) was applied to the target variable ('Closed PnL') to address skewness.

Predictive Modeling

- The data was split into training (80%) and testing (20%) sets.
- Two regression models were trained to predict the transformed Closed PnL:
    - **Random Forest Regressor:** Achieved an R^2 score of 0.1734 and an RMSE of 0.1300. A scatter plot showed the actual versus predicted values.
    - **XGBoost Regressor:** Achieved a slightly better performance with an R^2 score of 0.1970 and an RMSE of 0.1281.

Conclusion

The analysis provides insights into the relationship between market sentiment and trading outcomes for BTC and ETH. While the predictive models (Random Forest and XGBoost) showed a modest ability to predict logged Closed PnL (R^2 around 0.17-0.20), suggesting that the selected features have some predictive power, there is still significant room for improvement. Further feature engineering, exploring different models, and hyperparameter tuning could potentially enhance the prediction accuracy. The visualizations highlight how trading activity and outcomes vary across different sentiment classifications and between the two analyzed coins.

Final info after cleaning amd EDA:

This report summarizes the descriptive statistics of the df_filtered DataFrame, which contains the trade data after removing outliers and filtering for coins with a sufficient number of trades (BTC and ETH).

Key Observations from df_filtered.describe():

- **Number of Records:** The filtered dataset contains **8,082** trade records. This is significantly less than the original merged dataset (27,932), indicating the impact of outlier removal and filtering.
- **Numerical Features:**
    - **Execution Price:** The mean execution price is now around **76,468**, with a standard deviation of **31,801**. The range is from 1823.9 to 108410. Compared to the merged_df, the mean is higher and the standard deviation is lower, suggesting

that lower-priced coin trades were largely removed, and the spread is reduced after outlier treatment.

- **Size Tokens:** The mean is very low (**0.142**) and the standard deviation is also low (**0.428**), with a maximum of **5.64**. This indicates that trades with extremely large token sizes were effectively removed by the outlier treatment.
- **Size USD:** The mean is around **2502**, with a standard deviation of **3393** and a maximum of **19118.76**. Similar to 'Size Tokens', the outlier removal significantly reduced the scale and variability of this feature.
- **Start Position:** The mean is close to zero (**11.01**), with a standard deviation of **38.41**. The range is much tighter (-120.35 to 147.68) compared to the original data, showing the impact of outlier removal on this feature as well.
- **Closed PnL:** The mean Closed PnL is now very close to zero (**0.2757**), with a standard deviation of **0.857**. The range is from -2.42 to 4.69. This indicates that while the average PnL is slightly positive, the distribution is centered around zero after filtering. The extreme profit/loss values were removed.
- **Fee:** The mean fee is about **0.588**, with a standard deviation of **0.866** and a maximum of **3.17**. This is a much smaller range and lower values compared to the original data, confirming the removal of trades with exceptionally high fees.
- **value (Fear & Greed Index):** The average sentiment value is about **51.91**, similar to the merged_df. The range is from 15 to 94, indicating that the sentiment data itself did not have extreme outliers that were removed by the filtering process applied to the trade data.
- **timestamp_numeric:** The values are all 1.0, which is likely due to the conversion process or the nature of the timestamp data after filtering. This column may not be providing meaningful temporal variation in its current form.

- **Encoded Categorical Features:**
  - **classification_encoded:** The mean is around **2.115**, with a standard deviation of **1.166**. The range from 0 to 4 confirms that all classification categories are still present in the filtered data.
  - **coin_encoded:** The mean is about **0.152**, with a standard deviation of **0.359**. The range from 0 to 1 confirms that only the two dominant coins (BTC and ETH) remain after filtering.

Summary of Changes from merged_df to df_filtered:

The outlier removal and filtering steps significantly impacted the numerical features, drastically reducing their range, mean, and standard deviation. This suggests that the original dataset contained extreme values that were removed, resulting in a cleaner dataset for analysis and modeling, primarily focused on the two most frequently traded coins (BTC and ETH). The sentiment 'value' distribution remained relatively unchanged. The timestamp_numeric column's values being consistently 1.0 needs further investigation to understand its utility as a feature.