Risk-Aware Trader Classification Using Machine Learning

M. Sairam Karthik
B.Tech Computer Science, SRM University - AP

July 2025

Abstract

This report presents a machine learning pipeline designed to classify traders based on their risk profile using trade-level data. The solution integrates data preprocessing, exploratory data analysis, feature engineering, and model development using XGBoost, followed by prediction generation and output formatting.

1 Problem Statement

The objective of this assignment is to classify traders based on a risk factor using their trade-level data. Given multiple features like trade count, PnL, volume, and price, the task is to engineer meaningful features and build a classifier to predict the 'target label' associated with each trader. The model should focus on robust generalization and handle data noise or imbalance where necessary.

2 Dataset Overview

The dataset is composed of:

- **processed_trader_data.csv** - Trader-level aggregated information.
- **historical_data.csv** - Individual trade-level transaction data.

Key columns include:

- `trader_id`, `trade_count`, `pnl`, `volume`, `price`, `instrument`, `buy_sell_flag`, `target_label`

3 Approach and Methodology

3.1 Data Preprocessing

- Merged trade-level and trader-level datasets on `trader_id`.
- Removed outliers and missing values if any.
- Verified class balance and unique trader IDs.

3.2 Exploratory Data Analysis (EDA)

- Analyzed distribution of `pnl`, `volume`, and `price`.
- Observed almost zero or negative correlation between `trade_count`, `pnl`, `volume`, and `price`.
- Used boxplots, scatter plots, and heatmaps to visualize relationships and variance.

3.3 Feature Engineering

- Generated aggregate features per trader:
  - Mean and standard deviation of `price`, `volume`, and `pnl`.
  - Total `trade_count`, total `pnl`, and trade diversity (unique instruments).
- Encoded categorical variables (e.g. `instrument`, `buy_sell_flag`).

3.4 Modeling with XGBoost

- Used XGBoost classifier due to its ability to handle feature interactions, class imbalance, and non-linearity.
- Performed hyperparameter tuning using GridSearchCV.
- Evaluated model using F1-score, precision, recall, and confusion matrix.

3.5 Output and Submission

- Predicted labels for unseen traders in the test set.
- Formatted output as a CSV: `trader_predictions.csv` containing `trader_id` and `target_label`.

4 Results

- Achieved a balanced classification across target labels despite weak correlation between original features.
- XGBoost provided stability and generalization with fewer assumptions on data distribution.

5 Directory Structure

```
ds_M.SairamKarthik/

 csv_files/
    fear_greed_index.csv
    historical_data.csv
    processed_trader_data.csv
    trader_predictions.csv

 outputs/
    png files

 requirements.txt
 README.md
 report.pdf
```

6 Conclusion

The assignment successfully demonstrates a scalable and modular ML pipeline for risk-aware trader classification. The project applies sound preprocessing, feature engineering, and ensemble learning techniques to predict trader risk labels effectively.

Appendix

GitHub Repository: `https://github.com/Karthik0000007/web3_trading_insights`