CAIC TROPHY

TECH GC 2025

Mixed Signals - A Financial Prediction Challenge

Machine Learning in Finance

Introduction & Motivation

This competition involves building a model using real-world data derived from production systems, providing insight into the challenges of modeling financial markets. The challenge highlights several key difficulties, including fat-tailed distributions, non-stationary time series, and sudden shifts in market behavior.

Problem Statement

When approaching modeling problems in modern financial markets, several complexities arise. Prices of financial instruments may not always reflect all available information rationally. Additionally, distributions can exhibit fat tails, time series can be non-stationary, and data may fail to satisfy many underlying statistical assumptions.

Furthermore, financial markets involve numerous individuals and institutions that continuously adapt to technological advancements, societal changes, and geopolitical events. These factors contribute to the complexity of modeling such systems.

The challenge requires participants to build a predictive model using real-world financial data. The dataset consists of various features and target variables related to markets where automated trading strategies operate. Some features and target variables have been anonymised and lightly obfuscated while preserving the core problem characteristics.

Your task is to use the feature variables to predict some of the target variables (namely target-1 and target-2).

Solution Deliverables

- This is a coding challenge, so your solution must contain code.
- You are expected to perform exploratory data analysis (EDA), statistical tests, correlation analysis and other relevant analyses to determine the best features for each target variable.
- For predictions:
  - You will be provided with the first target variable value for each day.
  - Your code should avoid looking ahead into the future while making predictions.
- Submission requirements (to be released shortly) will include:
- Implementing a predict function.
- The function should accept a Pandas DataFrame similar to the sample dataset and populate the responder column.

- Code must be written in Python.
- You must provide a requirements file.
- You need to set seeds for reproducibility.
- Time limits for training and predictions will be announced soon.

Evaluation Parameters

Submissions are evaluated using a scoring function based on the sample-weighted zero-mean R-squared score of a specific response variable. The formula is given by:

The R-Squared score will be calculated for each of the two responders and we will be taking a weighted sum of both the R-Squared scores.

$$R^2 = 1 - \frac{\sum w_i (y_i - \hat{y}_i)^2}{\sum w_i y_i^2}$$

Your data analysis will also be judged on the basis of the experiments and the statistical tests you perform. There will be penalties if your code runtime exceeds the limits.

Weightage for different components:

- Rank based on weighted R-squared (75%) - Points will be awarded following a percentile based system.
- EDA, Statistical Analysis etc. (25%)
- Penalty for exceeding time limits

Data Description

Find the sample training dataset at the following link:

- date_id and time_id - Integer values that are ordinally sorted, providing a chronological structure to the data, although the actual time intervals between time_id values may vary.
- symbol_id - Identifies a unique financial instrument.
- weight(w_i) - The weighting used for calculating the scoring function.
- feature_{0...78} - Anonymised market data.
- target{0...8} - Anonymized responders clipped between -5 and 5. The target_1 and target_2 field is what you are trying to predict.

Team Size

A team of maximum 5 members will be allowed to register