Overview of Two Kaggle Competitions in Financial Market

Jing Guo

Strats Associate, Goldman Sachs

PhD in Financial Engineering, Columbia University

January 28, 2018

Outline

Overview

- Kaggle Algorithm Trading Challenge was organized by the Capital Markets Cooperative Research Center (CMCRC - www.cmcrc.com) in Macquarie University, Sydney, Australia.
- The timeline is from Nov 11, 2011 to Jan 8, 2012.
- Winning prize is cash $8,00 and consideration for entry to the CMCRC PhD program.
- The winning team goes to Ildefons Magrans de Abril and Masashi Sugiyama de Abril and Sugiyama (2013), two engineer researchers in Tokyo Institute of Technology, Japan.

Introduction

- The competition was to predict the short term response following large liquidity shocks.
- A liquidity shock is defined as any trade that changes the best bid or ask price.
- Liquidity shocks occur when a large (series of small) trade consumes all available volume of best offer.
- This kind of model can be used to optimize execution strategies of large transactions.
- Following a liquidity shock the spread may be temporarily widened, and/or result in permanent price shifts.

Training & Testing Dataset

- The training dataset consists of 754,018 samples of trade and quote data observations before and after a liquidity shock.
- There are 102 different securities of the London Stock Exchange (LSE) included.
- The liquidity shock takes place at time interval 51, i.e., time interval 1-50 are pre-liquidity and time interval 52-100 are post liquidity.
- The test dataset consists of 50000 samples similar to the training dataset but without the post-liquidity shock observations.

Training & Testing Dataset (Cont'd)

- Further variables
  - security id (*security_id*)
  - indicator of buyer or seller (*initiator*)
  - the volume-weighted average price causing the liquidity shock (*trade_vwap*)
  - the total size of the trade causing the liquidity shock (*trade_volume*)
- The test dataset consists of 50,000 samples similar to the training dataset but without post-liquidity-shock observations (i.e. time interval 51-100).
- The goodness of fit is measured by root-mean-square error (RMSE).

$$\textbf{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Outline

Overview

- Kaggle Two Sigma Financial Modeling Challenge was organized by *Two Sigma Investments*, who has been applying technology and systematic strategies to financial trading since 2001.
- The timeline is from Feb 22, 2017 to Mar 1, 2017.
- The winning prize is cash $25,000.
- Top 7 teams are awarded with cash prizes.
- It is a very noisy data set with high result variance. Only one team of top ten on public leaderboard made it to the top ten on private leaderboard.

Introduction

- The competition was to predict price returns y_t using only feature information **at time** t X_t.
- It is a code competition. Instead of giving estimations given test dataset, as in most Kaggle competitions, this competition asks participants to submit explicit prediction functions without test set, i.e. instead of to give $\hat{y}_{test}|X_{test}$, participants should submit $\hat{f}$ without X_{test}, and will be evaluated based on $\hat{f}(X_{test})$.
- Although the organizer does not provide the economic meanings of explanatory and response variables, the training data set is huge with $1 \sim 2$ million lines, indicating it is probably high frequency data.
- There is time constrain to model processing, preventing participants from using time consuming models.

Training & Testing Dataset

- The training dataset consists of 1,710,756 samples of X (financial features) and y (probably price returns of a variety of assets).
- There are 126 financial features, *derived* feature 0-4, *fundamental* feature 0-63 and *technical* feature 0-44. Actual physical meaning of each feature is not given. However, we will see in the following lectures how a smart participant analyzed and derived the physical meanings of some features.
- Timestamps, ranging from $0 \sim 1800$, are without actual meaning. If the unit is one second, then it is 30-minute data.
- There are 2000+ assets appear in the data set. The information of multiple assets can appear in one timestamp.

Training & Testing Dataset (Cont'd)

- Response variable $y_{i,t}$ is probably the return of asset i at time t. The mean of y across different timestamps and assets are roughly 0.
- The goodness of fit is measured by $R = sign(R^2)\sqrt{|R^2|}$, where

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}.$$

  Here R^2 could be negative.
- Because of 1-hour computation time constrain, fancy machine learning models are not possible. Most posted high ranking solutions use regularized linear regression as main predictor.

References I

de Abril, I. M. and Sugiyama, M. (2013). Winning the kaggle algorithmic trading challenge with the composition of many models and feature engineering, *IEICE Transactions on Information and Systems* (3): 742–745.