Jun 26 Update

reporter: Jerry

Here I cleared up our discussion and shaped my understanding. Please let me know your opinion on this report.

Based on the two files you shared with me today, we discussed mainly on the train data filtering(i.e. using race_conditions in racecard) and the features we should take into account when modeling (categories in the PDF file about figure-based trainer profiles).

Thanks for your sharing, with the deepening of understanding, the model work flow seems more clear and structured. But there are still some questions/judgements need your affirm before I can move on the strategy. Please kindly check the red and **bold** words below.

Objective

Using the information we already have (from the future racecard, and the history data) to predict the performance

Information we have

- historical data
- July 2 races information
  - runner
  - jockey
  - trainer
  - race conditions

QUESTION

- How can I get the future racecard? Is there any structured data instead of this pdf form? That would be much easier rather than checking by hand.
- Can we get the surface information on the target future race? I can't find it in your race card.

Training set

Two steps

- Filter the data in the same situation as July 2.
    - horse
        - age
        - winless
            - Maiden
            - never won since n
        - sex
        - claiming
    - surface
        - turf
        - dirt
- Continue to filter the data of each participant on July 2, such as
    - trainer:
    - jockey
    - runner

Here we are interested in the trainer's profile

QUESTION

- What's your opinion on most import conditions to filter the data? Since we don't have to match exactly the same. Based on our discussion, list as the following. Is that right?
    - Maiden or not
    - Claim or not
    - Surface type

That is why you want me to just use race_type==M

Data Analysis

Upon now, we have each trainer's historical data (profile) in such race condition. Choose features to train the prediction model.

- features:

- anything valuable to impact performance.
- **Need add the HDW PSR rating into consideration**

Though I cannot understand the categories in tg_ftp.pdf, I find another naive trainer's profile

http://www.equibase.com/profiles/Results.cfm?type=People&searchType=T&eID=271889&rbt=TB

Same as tg_ftp.pdf, they both discussed the profiles of trainer in respective of the statistics which is essentially different our method using time-series prediction.

I think our analysis which considers the dynamic of time-series is better. But statistics for the past performance is definitely a great idea other than some indicator (I am afraid they cannot be easily learned by our model since the {1,0} set cannot elucidate notion of distance by its value). We can transfer our last 10 races data into some meaningful statistics instead, which is equivalent to apply a time window to time-series, similar to what we do in equity index and implied volatility.