Jリーグの観客動員数予測

https://signate.jp/competitions/27

- **お題**
  - 2014年後半のJ1及びJ2における試合の観客動員数を予測いただきます
- **データ概要**
  - 学習：2012～2014年シーズン前半の試合及びそれに関連するデータ
  - 予測：2014年シーズン後半の試合及びそれに関連するデータ
- **問題のポイント**
  - 今回は回帰問題です（数値を当てる問題）
  - データが一部分割されている為、データ結合が必要です
  - 文字列を分割して変数を作る等、データ加工も必要です
  - 目的変数の性質把握や予測に何が寄与するかの仮説設計が重要です

基本データ

- **モデル学習用データ（train.csv）**
  - J1,J2の2012～2014年のシーズン前半（7/31）
  - y（目的変数）となる各対戦カードの観客動員数も含む
- **モデル検証用データ（test.csv）**
  - J1,J2の2014年シーズン後半（8/2～11/23）
  - yは含まれていない
  - 今回の問題ではこのデータ中の対戦カードのみを予測
- **スタジアムデータ（stadium.csv）**
  - 各スタジアムの所在地と収容人数（※結構重要です）
- **試合詳細データ（condition.csv）**
  - 各対戦カードのスコアやスタメン、天候等
- **応募用サンプルファイル（sample_submit.csv）**
  - 対戦カードid，そのidの観客動員数の予測値

一部データに
抜けがあります

e.g)
第１試合　マリノスvsアントラーズ
第２試合　マリノスvsエスパルス
第４試合　マリノスvsガンバ

SIGNATE

追加データ

- **モデル学習用試合追加データ（train_add.csv）**
  - 2012年86件、2013年93件、2014年53件の追加対戦データ（全232件）
  - モデルの学習に使っていただいて構いません

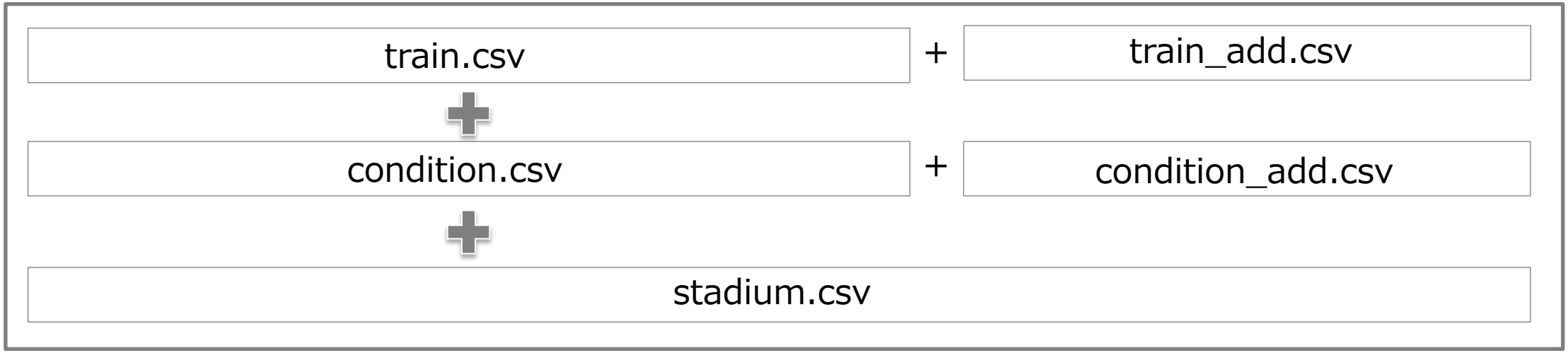
- **2014年度後半試合追加データ（2014_add.csv）**
  - 2014年後半の38件の対戦データ
  - あくまで補足データとなります
  - 得点が含まれているので、モデル検証用データと組み合わせることで、その対戦時にチームが現在何位なのかを割り出すことができます
    - 順位等を説明変数として加えたい時に利用してください

- **試合詳細追加データ（condition_add.csv）**
  - condition.csvの中で不足している試合の詳細データ（全270件）

2014年度後半試合追加データはあくまで補足データです。
応募時にはこのデータの対戦idの予測値は追加しないでください。

結局どうすれば良いのか？

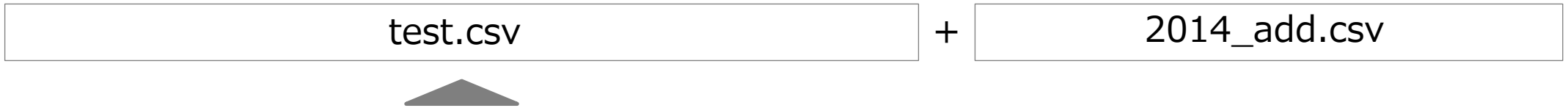
- **学習期間用の全データを作るには…**



- **2014年度後半試合追加データは何に使うの？**

  -> 順位を表す説明変数を作りたい時に利用します

  ※追加の試合データはコンペでの予測対象ではないので注意してください



コンペ応募時はこちらの対戦idの予測結果のみを応募
（sample_submit.csvの対戦idを予測）