第 3 章　　データ解析コンテスト練習

【練習問題】国勢調査からの収入予測

https://signate.jp/competitions/107

3.1　概要

教育年数や職業等の国勢調査データから年収が$50,000 ドルを超えるかどうかを予測するモデルを作成していただきます。
国勢調査は、すべての人を対象として実施され、国の最も重要かつ基本的な統計調査です。調査から得られる情報は，基礎データとして幅広い用途に利用され、国民生活に役立てられています。
今回は国勢調査から年収の予測（$50,000 を超えるか否か）に挑戦していただきます。
本コンペを活用して、SIGNATE でのデータ解析・モデル構築を体験してください。

データ概要：

- 課題種別：分類
- データ種別：多変量
- 学習データサンプル数：16280
- 説明変数の数：14
- 欠損値：あり

評価方法：

- 精度評価は、評価関数「Accuracy」を使用します。
- 評価値は 0～1 の値をとり、精度が高いほど大きな値となります。

$$\text{Accuracy} = \frac{n\left(\{i|y_i = \hat{y}_i \ (i = 1, 2, ..., N)\}\right)}{N}$$

データ：データ説明とダウンロード
投稿：結果提出
ランキング：結果ランキング

3.2 2 回目の例に含まない点

下記の点はこのデータセットのために必須な処理、または結果向上のために効果がある処理

- 欠損値対応
- カテゴリ特徴量を one-hot encoding にする
- 特徴量エンジニアリング
- 交差検証
- ハイパーパラメータチューニング
- 他のアンサンブル方法

3.3 レポートと報告会

レポート（スライド）内容：

- スライド 3、4 枚ほど
- 必須: 構築した一番結果が良いモデル（方法）、実験設定
- 必須: 結果と考察（結果と結果ランキング、ランキングを取った時間を記入）
- 任意: データ処理、特徴量エンジニアリング、探索的データ解析
- 任意: 苦労した点・工夫した点、モデルと結果考察など

その他：

- 登録ニックネームは自由、ニックネームをレポートに記入してください
- 報告会前の提出 (締め切り、2 月 3 日 9:00 AM)

- 報告会：
- 一人の考えは限られている、他人の方法や経験を聞くことは効果的です
- データ解析演習 5 回目、一人約 3 分間
- ランキングトップ 3 の人に質疑応答で時間の延長の可能性がある

- Special Thanks
  TA: 三島　大進
  TA: 小宮山　亮太

- 誤字・脱字などを見つけたら

口頭でもメールでも構いませんので、李 (jyli@yamanashi.ac.jp)まで連絡して下さい。