**銀行データ　分析方法の検討**

* 目的
  + これまでビジネスデータサイエンスの教科書で学んできたことを実践し、さらに理解を深めるため、kaggle上のデータセットを用いてデータサイエンスをしてみる。
  + 裏アジェンダとしては、pythonのリハビリと、Rの取得。まずはpythonから初めてデータの中をつけて、次にRを用いてデータサイエンスをしてみる。
* 実施要領
  + まずは、データを眺めた後、分析の設計をしてみる。
  + その後操作が慣れているpythonでざっくりデータをいじった後、Rでビジネスデータサイエンスの教科書に記載されていることを実施してみる
* データ
  + 参照：<https://www.kaggle.com/volodymyrgavrysh/bank-marketing-campaigns-dataset>
  + 概要：ポルトガルの銀行のマーケティング（定額預金加入）の結果。電話でのコールで加入を促進
  + データカラム
    - age 56：年齢
    - job housemaid：職業
    - marital married：結婚有無
    - education basic.4y：学歴
    - default no：破産しているか
    - housing no：家あるか
    - loan no：ローンあるか
    - contact telephone：連絡先
    - month may：最終コンタクト月
    - day\_of\_week mon：最終コンタクト曜日
    - duration 261：最後のコールの長さ※目的のカラムと連動
    - campaign 1：キャンペーン期間中の有効連絡回数
    - pdays 999：前回のキャンペーンから最後に連絡してからの経過期間
    - previous 0：このキャンペーンの前にクライアントに対しての連絡数
    - poutcome nonexistent：前回のキャンペーンのアウトカム
    - emp.var.rate 1.1：雇用変動指数？
    - cons.price.idx 93.994：消費指数
    - cons.conf.idx -36.4：信頼感の関数
    - euribor3m 4.857：欧州の金利指標
    - nr.employed 5191：雇用者数
    - y no：アウトプット（キャンペーン契約したか）

①分類（第4章に該当）

* 想定目的：
  + キャンペーン中を仮定。顧客をスコアリングして電話をかける人を選ぶ
* アウトプット：
  + 顧客ごとのキャンペーンの成功、失敗の予測結果及びその正誤（confusion matrix）
  + どの変数が分類に寄与していたかの分析結果
* 予定されるプロセス：
  + 目的関数：
    - ROC
    - FPR?、FNR?→理由づけ必要
  + アルゴリズム（いずれもOOSで検証、CVしても良いが一旦ホールドアウトで）  
    <https://data-analysis-stats.jp/python/ホールドアウト検証と交差検証/>
    - Python:ロジスティック回帰、ランダムフォレスト・・・
    - R：ロジスティック回帰（ラッソ回帰使ってみる）、
  + 前処理：
    - Stringの値が入っているカラムをmaskingする
    - Outputと明確に関係性があるカラムを除く
    - ・・・

②主成分分析（因子分解、第7章に該当）

* アウトプット：
  + 顧客クラスターの作成（次回マーケティングのための施策検討のため）
  + どのクラスターがどんな特徴を持っているか確認
* プロセス：
  + アルゴリズム
    - Python：PCA
    - R：PCA、PCR
  + 前処理：分類と同等

③統制変数（第6章、第5章の実験も）

* アウトプット：
  + どの特徴量がキャンペーンの契約/拒否に影響を与えるかを確認