DataMix

データサイエンティスト育成コース
ベーシックステップ

統計学 PBL
「リテール銀行のテレマーケティング」

株式会社データミックス

講義中のルール

- 不明な点がありましたら、その場で質問してください。
- 話を遮っていただいて構いません。

今回のプロジェクト

プロジェクト　ケース概要

あなたは銀行のデジタル部門のデータサイエンティストチームに所属しています。

本部門では、「**定期預金の口座数**」がKPIとなっています。
そのため、口座数獲得のために、定期的に**架電**によるダイレクトキャンペーンを行っています。

しかし、現状では、なかなか口座獲得数が上がっていないことがマーケティングチームの懸念事項となっています…

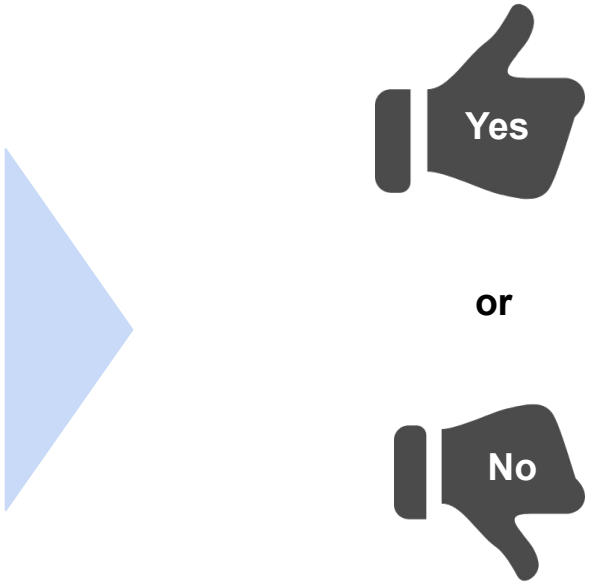
そこで、次回のキャンペーンの成果向上に向けたデータ分析を行う必要があります。あなたがお願いされている仕事は次の2つです。

プロジェクト　今回の課題

1. マーケティングチームのトークスクリプト作成のインプットとして、過去のテレマーケティングデータを使った**ターゲットのユーザー像（ペルソナ）**を浮かび上がらせることが求められています。
2. 今後のテレマーケティングの収益(売上 - 費用)を最大化させるため、**予測モデルを用いたアタックリスト**を出力するアルゴリズムの作成を求められています。
   なお、テレマーケティングを実施するにあたって、1人の顧客に架電するコストは一律500円かかります。
   一方、1件獲得したときの平均LTV（≒ 売上）は、
   一律5,000円です。

（注意: 上記1,2は必ずしも同じモデルである必要がありません。）

口座開設までのフロー



既存の顧客リストを使って、コールセンターから顧客に架電する。
ただし、定期預金口座開設に関する返事(“Yes”または”No”) がない場合は、同じ顧客に対して、複数回架電する。

結果は”yes”または”no”の二値

課題2において、作成していただきたい関数

アタックリストを出力する関数

- 引数: dataset（trainデータの列と全く同じ並び順）
- 出力: 全ユーザーに関して、電話をした/しないほうが良い人を1/0にしたフラグのリスト

次回講義の冒頭に、**テストデータ**をCSVファイルとして、お渡しします。その後1時間で期待収益を計算していただきます。

用意していただきたいアウトプット

1. 以下の内容を含んだ、プレゼンテーション（~20分）
   - 使用したデータについて
   - 分析アプローチ
   - 分析結果１：ターゲットとすべきペルソナ
   - 分析結果２：予測精度、期待される収益

2. （任意）個人ごとにマークダウンなどで記載したレポートとコードをGithubにアップしてください（プログラムに参加することが外部にわかると問題になる方は実施しなくて結構です）。

各カラムの定義

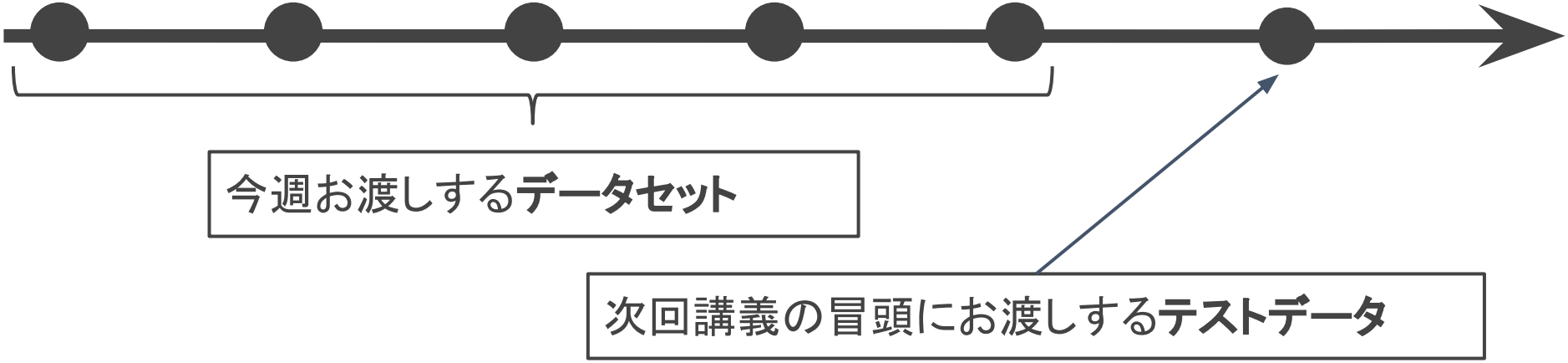
カラム名	定義
age	年齢
job	職業
marital	婚姻状況
default	クレジットの支払遅延
education	最終学歴
housing	不動産ローンの有無
loan	個人ローンの有無
contact	連絡デバイス
day_of_week	最終連絡曜日
duration	通話時間（秒）
campaign	キャンペーン期間中の接触回数

カラム名	定義
pdays	前回の接触からの経過日数
poutcome	以前のキャンペーン結果
previous	以前のキャンペーンの接触回数
emp.var.rate	employment variation rate (※詳細なし)
cons.price.idx	消費者物価指数
cons.conf.idx	消費者信頼感指数
euribor3m	3ヶ月ユーリボー指標金利
nr.employed	四半期ごとの就業者数
y	テレマーケティングの結果

データについて

今週お渡しするデータは過去のテレマーケティングデータです。

次回講義の冒頭でお渡しする**テストデータ**は、直近で実施されたテレマーケティングデータです。



※注意
次回お渡しするデータは直近実施されたテレマーケティングデータなので、経済指標は特定の値に定まっています。

データのカラムや定義を確認

以下のデータの説明ページを見て、データのカラムや定義の確認をしましょう。

データの説明は以下を参照
https://archive.ics.uci.edu/ml/datasets/bank+marketing

※ データはURLからダウロードせず、お渡しするデータを使って下さい。

注意したい論点

1. 次回のキャンペーン開始時（架電時）に使えない説明変数が存在していないか？
2. どの変数を使用すれば、ユーザー像を語ることができそうか？

【参考】不均衡データへの対処テクニック

サンプリングによる対応と目的関数の重みを調整するアプローチが基本

Oversampling / Undersampling テクニック

元データ

成約
成約
成約
失注
失注
失注
失注
失注
失注
失注
失注
失注
失注

学習データ

成約を50%/ 失注を50%

成約
成約
失注
失注

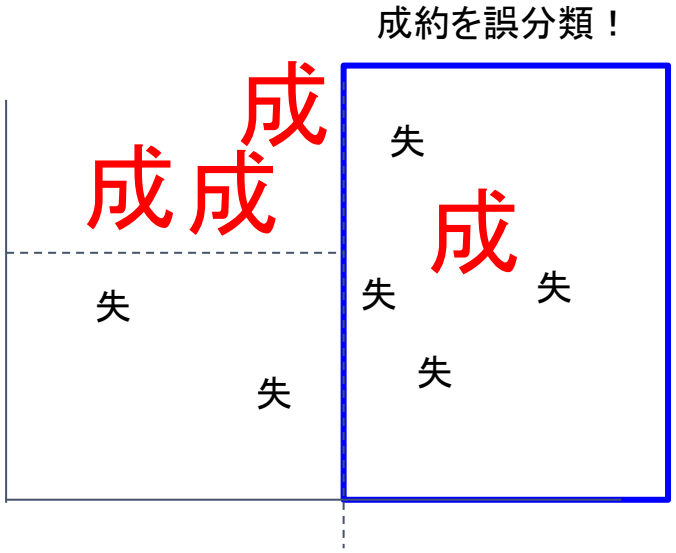
テストデータ

元の割合を維持

成約
失注
失注
失注

目的関数の重みを調整

誤分類した場合のペナルティを調整する



不均衡データへの対処テクニックの例

Rで行う不均衡データへの対処テクニック

```
randomForest(y~.,
             data = xxx,
             sampsize = c(xxx, xxx),
             ntree = xxx)
```

各決定木作成におけるデータを調整します。

e.g.
`sampsize = c(500, 500)`
それぞれの決定木を作成するときに、"yes"から500件、"no"から500件サンプリングします。

※パッケージrandomForestにも重み付けはありますが、下記の通り、改善段階のようです。
Wishlist (formerly TODO):
** Implement the new scheme of handling classwt in classification.*
【引用： https://github.com/cran/randomForest】

pythonで行う不均衡データへの対処テクニック

```
RandomForestClassifier(…,
                       class_weight=xxx,
                       …)
```

"yes"もしくは"no"に対して重み付けを行います。

e.g.
*class_weight="balanced"*とした場合は以下の計算が実行されます。

```
n_samples / (n_classes * np.bincount(y))
```

つまりは、あるクラスへの重みは、学習データにおける出現頻度に反比例して調整されます。
言い換えると学習データにおける出現頻度が多いと重みは小さくなり、少ないと重みは大きくなります。

【引用
：https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html 】

Rでデータのハンドリングをしてみましょう

Rstudioを開いてください。

使用するファイルは、

- Macの方は「bank_marketing_utf8.R」
- Windowsの方は「bank_marketing_sjis.R」

です。