データサイエンティスト育成
ベーシックステップ

R による統計モデリング実践課題

2017/09/02

Masayuki Mizuno

Agenda

- 1. 目的
- 2. アプローチ
- 3. <Step 1> 予測モデルの作成
  - a.学習データの準備
  - b.説明変数の準備
  - c.ロジスティック回帰分析による予測モデルの作成
- 4. <Step 2> ターゲットとすべきユーザー像を設定
  - a.予測モデルの解釈
  - b.期待される収益
  - c.今後のアクション
- Appendix 1.予測モデル作成プロセス

1. 目的

- 次回キャンペーンの ROI を最大化するために、ターゲットとすべきユーザー像を設定する。

2. アプローチ

- <Step 1> 予測モデルの作成
  - 過去の実績データをもとに、予測モデルを作成
    - a.学習データの準備
    - b.説明変数の準備
    - c.ロジスティック回帰分析による予測モデルの作成

- <Step 2> ターゲットとすべきユーザー像を設定
  - 予測モデルの解釈をもとに、ターゲットとすべきユーザー像を設定
    - a.予測モデルの解釈
    - b.期待される収益

3. <Step 1> 予測モデルの作成
- a.学習データの準備（1/3）

- 過去のキャンペーンにおける実績データ
  - 2008 年から 2010 年の 37068 件
    - このうち 7 割 --- 学習データ
    - このうち 3 割 --- 検証データ
  - Input variables:
    - # bank client data:
      - 1 - age (numeric)
      - 2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown’)
      - 3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
      - 4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown’)

3. <Step 1> 予測モデルの作成
- a.学習に使用したデータ（2/3）

- 5 - default: has credit in default? (categorical: 'no','yes','unknown’)
- 6 - housing: has housing loan? (categorical: 'no','yes','unknown’)
- 7 - loan: has personal loan? (categorical: 'no','yes','unknown')
  # related with the last contact of the current campaign:
- 8 - contact: contact communication type (categorical: 'cellular','telephone’)
- 9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec’)
- 10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri’)
- 11 - duration: last contact duration, in seconds (numeric).

- # other attributes:
  - 12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
  - 13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

3. <Step 1> 予測モデルの作成
- a.学習に使用したデータ（3/3）

- 14 - previous: number of contacts performed before this campaign and for this client (numeric)
- 15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success’)

- # social and economic context attributes
  - 16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
  - 17 - cons.price.idx: consumer price index - monthly indicator (numeric)
  - 18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
  - 19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
  - 20 - nr.employed: number of employees - quarterly indicator (numeric)

- Output variable (desired target):
  - 21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

3. <Step 1> 予測モデルの作成
- b.説明変数の準備

- 除外する説明変数
  - duration
    - This attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
  - campaign
    - 当該のキャンペーン中に何回電話したかを示すものであり、次回のキャンペーン時には事前には分からない。
- 追加する説明変数
  - noloan
    - ローンを持っている場合には yes。それ以外（持っていない or 不明）の場合には no。

3. <Step 1> 予測モデルの作成
- c.ロジスティック回帰による予測モデルの作成

- モデル作成の詳細な手順は Appendix 参照
- 採用した説明変数
  - job
  - marital
  - default
  - contact
  - month
  - day_of_week
  - poutcome
  - cons.conf.idx
  - nr.employed
  - noloan

4. <Step 2> ターゲットとすべきユーザー像を設定
- a.予測モデルの解釈

- 選択された説明変数
  - job+marital+default+contact+month+day_of_week+poutcome+cons.conf.idx+nr.employed+noloan
- オッズ比の比較によるターゲット像の設定
  - job
    - unknown と比較して、retired が約 2 倍、student が約 1.6 倍
  - default
    - no は yes と比較して、約 5200 倍
  - contact
    - cellular は telephone と比較して、約 1.4 倍
  - poutcome
    - success は failure と比較して、約 158 倍
  - month
    - 最低月 may と比較して、mar が約 8 倍、dec が約 6 倍、oct が約 5 倍

4. <Step 2> ターゲットとすべきユーザー像を設定
- b.期待される収益

- 3 割の検証データを使用
  - yes : 849 件／no : 10,272 件／合計 11,121 件
    - yes の割合 : 7.63% ⇒ ランダムにアタックして成約する確率
- 検証データを予測モデルに適用した結果
  - 成約確率が 0.196 より大きいユーザーをターゲット
  - 全体としての成約率と期待収益
    - 成約率
      - 0.38
    - 期待収益
      - 168,000 円

4. <Step 2> ターゲットとすべきユーザー像を設定
- c.今後のアクション

- 既存ユーザー
  - 予測モデルを活用

- 新規ユーザー
  - 予測モデルへの入力値がプロファイル済みであれば、予測モデルを活用
  - プロファイルがない場合には、オッズ比の比較によって設定されたターゲット像を参考にして優先度を決める

Appendix 1. 予測モデル作成プロセス

- [01] .-duration-campaign でロジスティック回帰
  - 影響を持つ回帰係数
    - job
    - default
    - contact
    - day_of_week
    - poutcome
  - AIC = 12233
- [02] 01 の結果に対して step を実行
  - 選択された説明変数
    - marital+default+contact+month+day_of_week+poutcome+emp.var.rate+cons.price.idx+noloan
    - job が選択されなかった
  - AIC = 12210

Appendix 1.予測モデル作成プロセス

- [03] 02 の結果に対して car::vif を実行
  - VIF 値が高い説明変数
    - emp.var.rate
    - cons.price.idx
- [04] .-duration-campaign-emp.var.rate-cons.price.idx でロジスティック回帰
  - 影響を持つ回帰係数
    - job
    - default
    - contact
    - day_of_week
    - month
    - poutcome
    - nr.employed

Appendix 1.予測モデル作成プロセス

- AIC = 12231

- [05] 04 の結果に対して step を実行
  - 選択された説明変数
    - marital+default+contact+month+day_of_week+poutcome+cons.conf.idx+nr.employed+noloan
    - job が選択されなかった
  - AIC = 12210
  - 期待収益 = 160,550 円（しきい値=0.197）
  - 適合率 = 0.37

Appendix 1.予測モデル作成プロセス

- [06] 05 に job を追加
  - 選択された説明変数
    - job+marital+default+contact+month+day_of_week+poutcome+cons.conf.idx+nr.employed+noloan
  - AIC = 12211
  - 期待収益 = 168,000 円（しきい値=0.196）
  - 適合率 = 0.38
  - 以下、オッズ比

(Intercept)	jobadmin.	jobblue-collar	jobentrepreneur	jobhousemaid	jobmanagement
6.358923e+29	1.355394e+00	1.199955e+00	1.405601e+00	1.135284e+00	1.279851e+00
jobretired	jobself-employed	jobservices	jobstudent	jobtechnician	jobunemployed
1.983710e+00	1.171161e+00	1.295085e+00	1.632395e+00	1.301084e+00	1.452752e+00
maritalmarried	maritalsingle	maritaldivorced	defaultno	defaultyes	contactcellular
8.119835e-01	9.239079e-01	8.067746e-01	1.307689e+00	2.498818e-04	1.387143e+00
monthdec	monthnov	monthapr	monthmar	monthjul	monthaug
6.427424e+00	1.619836e+00	2.120105e+00	8.238642e+00	2.940679e+00	1.880362e+00
monthoct	monthjun	day_of_weektue	day_of_weekwed	day_of_weekthu	day_of_weekfri
5.424432e+01	2.830605e+00	1.176972e+00	1.253795e+00	1.270239e+00	1.117979e+00
poutcomefailure	poutcomesuccess	cons.conf.idx	nr.employed	noloanyes	
5.819081e-01	3.683838e+00	1.039822e+00	9.863970e-01	1.089818e+00	