DataScientist Society

ロジスティック回帰モデル (Logistic Regression Model)

～ロジスティック回帰モデル、決定木モデルとの比較、モデル構築時の注意点～

今回の目標

前回は決定木モデルによってモデル構築を行ったが、
今回はロジスティック回帰モデルによるモデル構築を行う。

下記の5項目を今回の目標とする。

- ロジスティック回帰モデルについて理解する。
- ロジスティック回帰モデルと決定木モデルの違いを理解する。
- モデル構築を行う時の注意点について理解する。
- モデル構築
- Submit

今回の目標 (目次)

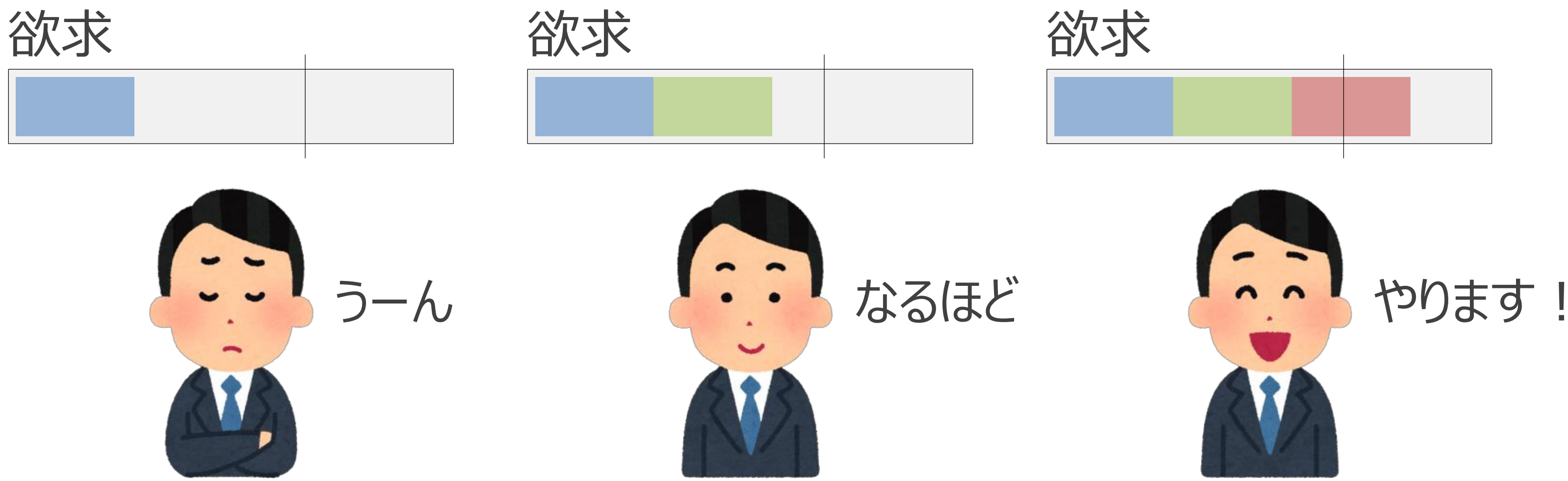
前回は決定木モデルによってモデル構築を行ったが、
今回はロジスティック回帰モデルによるモデル構築を行う。

下記の5項目を今回の目標とする。

- ロジスティック回帰モデルについて理解する。
- ロジスティック回帰モデルと決定木モデルの違いを理解する。
- モデル構築を行う時の注意点について理解する。
- モデル構築
- Submit

ロジスティック回帰モデルの仮説 ～銀行の顧客マーケティングを例に～

- 仮説1：顧客は、潜在的な「欲求」を持っている。
- 仮説2：欲求がある閾値を超えると、顧客は定期預金を申し込む。
- 仮説3：欲求は、説明変数の線形和で表される。



余剰資金を
定期預金口座へ
移しませんか？

30代で
定期預金始める人
多いですよ？

今申し込むと
1,000ポイント
もらえます！

ロジスティック回帰モデルの式

- 定期預金を申し込む確率のロジスティック回帰モデル
(残高と年齢から欲求が決まる場合の例)

$$申し込む確率 = 1/(1 + \exp(-欲求)) \quad \cdots \text{仮説1と仮説2に対応}$$
$$\boxed{欲求 = a_{残高} \times 残高 + a_{年齢} \times 年齢 + b} \quad \cdots \text{仮説3に対応}$$

自分でロジスティック回帰モデルを構築するときは、この式にもとづいて
「欲求を高める要因は何か？」を考え仮説を立てる。

- なお、上記式を変形すると以下のようになることから、
「対数オッズを説明変数の線形和で表したモデル」と解釈することもできる。

$$欲求 = \boxed{\log\left(\frac{申し込む確率}{申し込まない確率}\right) = a_{残高} \times 残高 + a_{年齢} \times 年齢 + b}$$

要因の仮説が正しいかを確かめるときは、この式にもとづいて
「対数オッズ(＝欲求)と説明変数(＝要因)が線形か？」を確認する。

(参考) ロジスティック関数を選ぶ利点

- exp(回帰係数) が調整済みオッズ比を表すため、各説明変数が目的変数へ与える影響を確かめられる。
- 年齢が上がると申し込み確率が上がるか確かめる例

$$
\begin{aligned}
&\text{申し込みオッズ} \\
&= \frac{\text{申し込む確率}}{\text{申し込まない確率}} \\
&= \exp\left(a_{\text{残高}} \times \text{残高} + a_{\text{年齢}} \times \text{年齢} + b\right) \\
&= \exp\left(a_{\text{残高}} \times \text{残高}\right) \exp\left(a_{\text{年齢}} \times \text{年齢}\right) \exp(b)
\end{aligned}
$$

$$
\begin{aligned}
&\text{年齢差の申し込みオッズ比} \\
&= \frac{\text{31歳の申し込みオッズ}}{\text{30歳の申し込みオッズ}} \\
&= \frac{\exp\left(a_{\text{残高}} \times \text{残高}\right) \exp\left(a_{\text{年齢}} \times \text{31歳}\right) \exp(b)}{\exp\left(a_{\text{残高}} \times \text{残高}\right) \exp\left(a_{\text{年齢}} \times \text{30歳}\right) \exp(b)} = \boxed{\exp\left(a_{\text{年齢}}\right)}
\end{aligned}
$$

年齢以外の影響を除いた調整済みオッズ比。

年齢が1歳上がると申し込み確率が……

- 上がる：オッズ比>1
- 変わらない：オッズ比≒1
- 下がる：オッズ比<1

ロジスティック回帰モデルの構築手順

1. 欲求を高める要因の仮説を立てる。
   例）ライフステージに応じて定期預金の必要性が変わるのでは？
2. 要因を変数に落とし込む。
   例）ライフステージは年齢と連動するため、年齢を変数にしよう。
3. 変数と対数オッズ（＝欲求）が線形になっていることを確認する。
   例）横軸を年齢、縦軸を対数オッズにして散布図を描き、
   点が直線状に並ぶか確認する。
4. 線形になっていない場合は、1に戻る。
   例）散布図が50代を境にV字を描いており、線形になっていない。
   支出のピークである50代に向けて資産形成するから、
   その前後で定期預金の必要性が下がるのかもしれない。

→ **後ほど、別資料で詳しく説明**

今回の目標 (目次)

前回は決定木モデルによってモデル構築を行ったが、
今回はロジスティック回帰モデルによるモデル構築を行う。

下記の5項目を今回の目標とする。

- ロジスティック回帰モデルについて理解する。
- ロジスティック回帰モデルと決定木モデルの違いを理解する。
- モデル構築を行う時の注意点について理解する。
- モデル構築
- Submit

ロジスティック回帰モデルと決定木モデルの比較 (1/3)

説明変数の種類

- カテゴリ変数を用いる場合には、ダミー変数化が必要。
- ダミー変数とは？
  - カテゴリ毎に 0, 1 のフラグに変換したもの。
  - カテゴリ数－1 のダミー変数を作れば表せる (残り1つは、すべて0で表現)。
  - カテゴリの数が多いと過学習してしまう。
  - R：caretパッケージを使用すると便利。
  - Python：Pandasのget_dummies関数を使用すると便利。

月
1月
2月
3月
4月
5月
6月
7月
8月
9月
10月
11月
12月

ダミー変数化

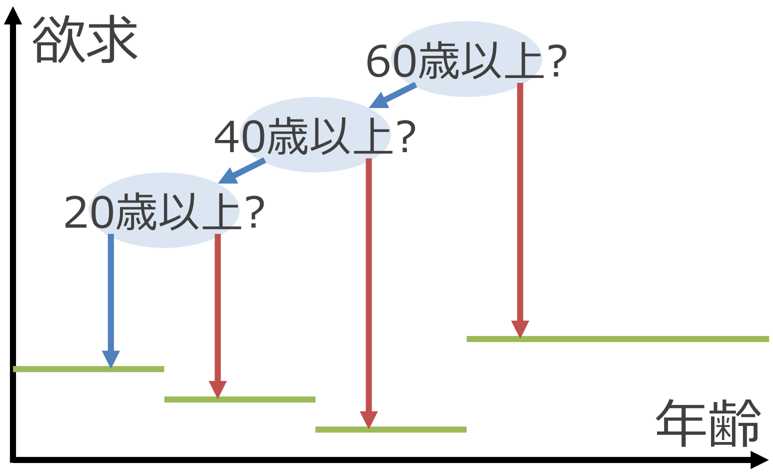
month_1	month_2	month_3	month_4	month_5	month_6	month_7	month_8	month_9	month_10	month_11
1	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0

ロジスティック回帰モデルと決定木モデルの比較 (2/3)

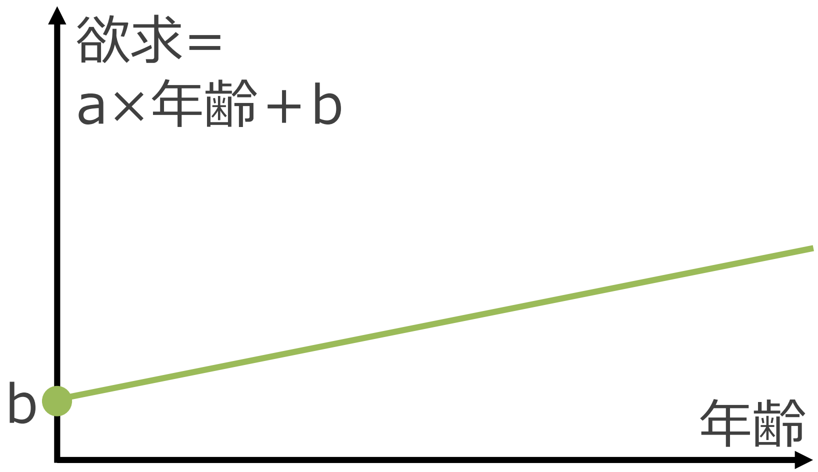
連続性、単調性

- 連続性：決定木モデルだと出力値はノードごとの値となるが、ロジスティック回帰モデルだと連続値となる。
- 単調性：式の形から、説明変数の値が増加すると出力値は……
  - 回帰係数 a が正のとき：常に増加する。
  - 回帰係数 a が負のとき：常に減少する。

⇒ 決定木モデルでは、説明変数の増減と出力値の増減が一定とは限らない。



決定木モデル

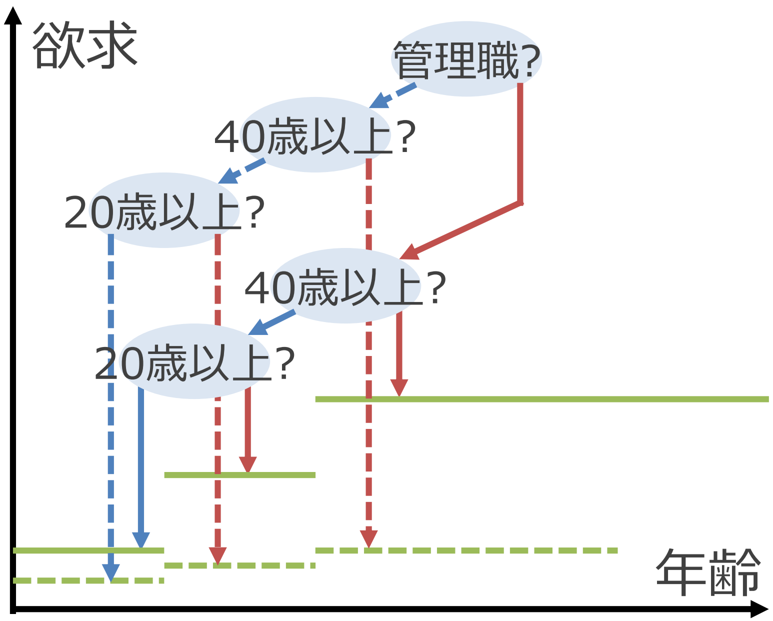


回帰モデル

ロジスティック回帰モデルと決定木モデルの比較 (3/3)

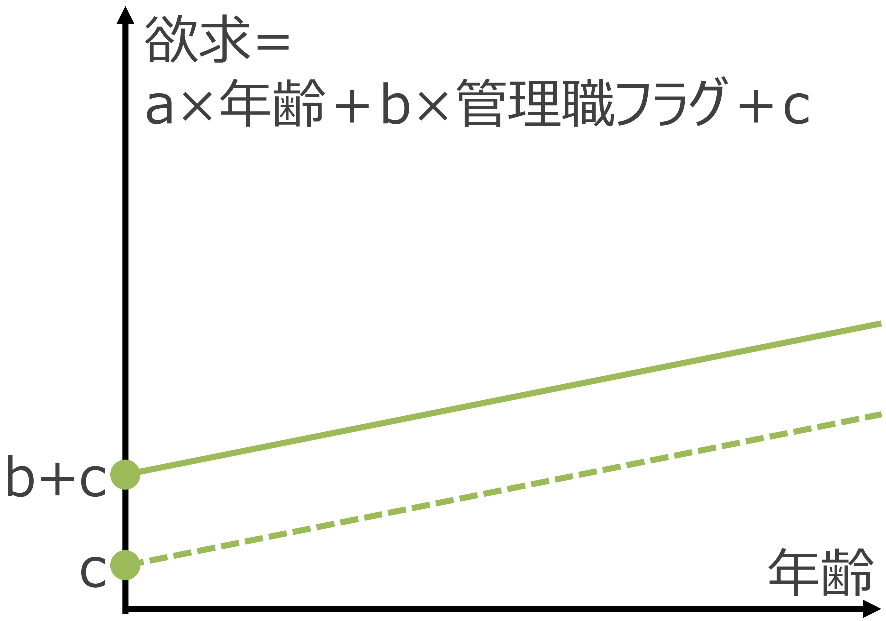
交互作用

- 前述の単調性より、決定木モデルで捉えられた交互作用をロジスティック回帰モデルで表現することは難しい。



決定木モデル

交互作用を表現できるため
役職によって年齢に対する欲求の増え方が
異なるようなモデルを作れる



回帰モデル

各説明変数が独立に単調性を持つため
年齢に対する欲求の増えるスピード (傾き) は
役職に関わらず同じになってしまう

今回の目標 (目次)

前回は決定木モデルによってモデル構築を行ったが、
今回はロジスティック回帰モデルによるモデル構築を行う。

下記の5項目を今回の目標とする。

- ロジスティック回帰モデルについて理解する。
- ロジスティック回帰モデルと決定木モデルの違いを理解する。
- モデル構築を行う時の注意点について理解する。
- モデル構築
- Submit

モデル構築を行う時の注意点 (1/2)

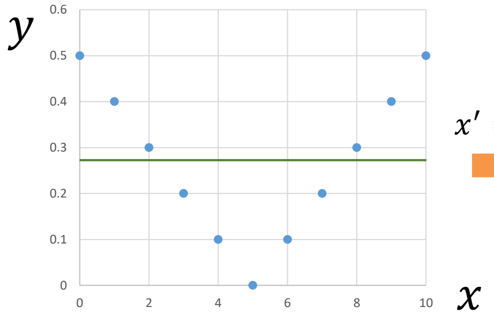
多重共線性

- 線形モデルを用いるときに良く起きる問題で、
  説明変数同士の相関が高いと回帰係数の推定が上手くいかない。
- 相関の高い説明変数の一方を外すなどの工夫が必要。
- 多重共線性の例：満年齢と数え年から申し込み確率を予測する場合
  - モデル
    - $\text{申し込む確率} = 1/(1 + \exp(-\text{欲求}))$
    - $\text{欲求} = a_{\text{満年齢}} \times \text{満年齢} + a_{\text{数え年}} \times \text{数え年} + b$
  - 回帰係数の推定結果
    - $\text{欲求} = 0.10 \times \text{満年齢} + 0.00 \times \text{数え年} - 0.20$
    - $\text{欲求} = 0.00 \times \text{満年齢} + 0.10 \times \text{数え年} - 0.20$
    - $\text{欲求} = 0.05 \times \text{満年齢} + 0.05 \times \text{数え年} - 0.20$
    - ……のように、回帰係数の候補が無限にある。

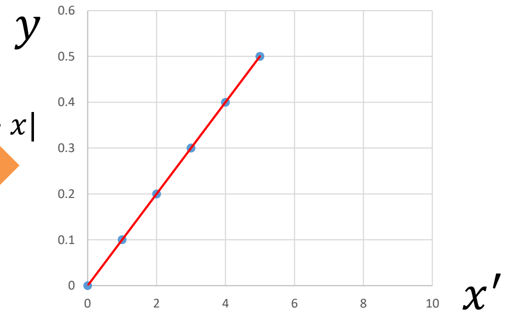
モデル構築を行う時の注意点(2/2)

非線形な変数

- 対数オッズと説明変数の関係が非線形の場合は、変数加工により線形に直すなどの工夫が必要。

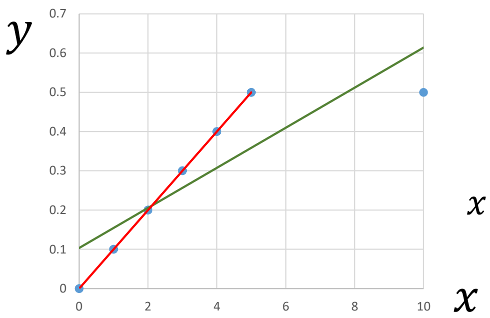


$x' = |5 - x|$



外れ値の影響

- 外れ値がある場合は、外れ値を置き換えるなどの工夫が必要。



$x > 5$
$\Downarrow$
$x = 5$に置換

今回の目標 (目次)

前回は決定木モデルによってモデル構築を行ったが、
今回はロジスティック回帰モデルによるモデル構築を行う。

下記の5項目を今回の目標とする。

- ロジスティック回帰モデルについて理解する。
- ロジスティック回帰モデルと決定木モデルの違いを理解する。
- モデル構築を行う時の注意点について理解する。
- モデル構築
- Submit

→ **別資料に記載**

参考文献

[1] 東京大学教養学部統計学教室：
自然科学の統計学 第18版, 2016年, 東京大学出版会.

[2] 平井有三：
はじめてのパターン認識, 2012年, 森北出版株式会社.