銀行の顧客ターゲティング

15期　石川愛花　中田芽衣

目次

目的

- ある銀行の顧客属性データと過去のキャンペーンでの接触情報を用い、キャンペーンの結果口座を開設したかどうかを予測する。
- マーケティングキャンペーンの効率化を図るためのモデリングを行う。
- 今回は27,168名の顧客データを分類することで、どんな人が口座を開設したのかを分析する

データ説明

・学習用データ：27,168名の顧客データ、キャンペーンの結果（口座開設の有無）

↓

・予測対象：18,083名のキャンペーンに対する反応（結果）

データ説明

カラム	ヘッダ名称	説明	変数種別
1	id	行の通し番号	
2	age	年齢	数値
3	job	職種	カテゴリ
4	martial	未婚/既婚	カテゴリ
5	education	教育水準	カテゴリ
6	default	債務不履行があるか	バイナリ
7	balance	年間平均残高	数値(€)
8	housing	住宅ローン	バイナリ
9	loan	個人ローン	バイナリ
10	contact	連絡方法	カテゴリ
11	day	最終接触日	数値
12	month	最終接触月	カテゴリ
13	duration	最終接触時間	数値(秒)
14	compaign	現キャンペーンにおける接触回数	数値
15	pdays	経過日数：前キャンペーン接触後の日数	数値
16	previous	接触実績：現キャンペーン以前までに顧客に接触した回数	数値
17	poutcome	前回のキャンペーンの成果	カテゴリ
18	y	定額預金申し込み有無	バイナリ(1／0)

出所）Deep Analytics（https://deepanalytics.jp/compe/1?tab=compedetail2016/10/15アクセス）

先行研究

- 単純接触効果

アメリカの心理学者ロバート・ザイアンスによって、『Attitudinal Effects of Mere Exposure』という論文の中で提唱された、複数回接触した単語、文字、図形などに好意を持つようになる効果。

論文中では被験者に知らない単語、漢字、人の顔を繰り返し見せる実験が行われ、被験者は単語や漢字については、提示頻度の高いものほど良い意味だと感じた。しかし人の顔については、接触の回数に関係なく好ましさが変わらないものもあり、被験者が好まないものは単純接触効果が及ばないことが確認された。

（ Ruby Marketing http://rubymarketing.jp/blog/mere-exposure_effect/ ）

事例

実際の預金口座キャンペーンの内容
①金利をあげる
②口座開設費無料口座維持費無料
③一定金額預け入れるとキャッシュバック（現金やポイント）
などのキャンペーンが多い

出所：http://www.woman110.com/200807/ 2016年秋預金金利等のキャンペーン一覧

事例

- 三井住友銀行:毎月口座振り込みがある消費者には月4回までATMの手数料を無料にする
- 三菱UFJ銀行:テレビ窓口、郵送、スマホアプリ、店頭窓口の4つの方法で口座開設サービスを提供している
- ジャパンネット銀行:年間残高が高い人に特典を設けている

出所:

- http://www.smbc.co.jp三井住友銀行公式ホームページ
  http://www.bk.mufg.jp/sp/ 三菱UFJ銀行公式ホームページ
  http://www.japannetbank.co.jpジャパンネット銀行公式ホームページ
  https://archive.ics.uci.edu/ml/datasets/Bank+MarketingUCI

予想

- 先行研究から、変数「compaign」(現キャンペーンでの接触回数）の効果はある程度見込めると考えられる。
- また同様に「previous」（現キャンペーン以前までに顧客に接触した回数）もキャンペーン参加に対して正の影響を及ぼすと予想する。

仮説

変数名	仮説
2 age 年齢	年齢が高いことはことはキャンペーン参加に正の影響を与える
3 job 職種	—
4 martial 未婚/既婚	既婚であることはキャンペーン参加に正の影響を与える
5 education 教育水準	—
6 default 債務不履行があるか	債務不履行がないことはキャンペーン参加に正の影響を与える
7 balance 年間平均残高	年間平均残高が高いことはキャンペーン参加に正の影響を与える
8 housing 住宅ローン	住宅ローン契約はキャンペーン参加に正の影響を与える
9 loan 個人ローン	個人ローン契約はキャンペーン参加に正の影響を与える
10 contact 連絡方法	携帯端末での接触はキャンペーン参加に正の影響を与える

変数名	仮説
11 day 最終接触日	—
12 month 最終接触月	—
13 duration 最終接触時間 最後に接触したときの会話の長さ(秒)	最終接触時間が長いことはキャンペーン参加に正の影響を与える
14 compaign 当キャンペーンにおける接触回数	現キャンペーンへの接触回数が多いことはキャンペーン参加に正の影響を与える
15 pdays 経過日数：前キャンペーン接触後の日数	前キャンペーン接触から日にちが立っていないことはキャンペーン参加に正の影響を与える
16 previous 接触実績：当キャンペーン以前までに顧客に接触した回数	現キャンペーン以前までの顧客接触回数が多いことはキャンペーン参加に正の影響を与える
17 poutcome 前回のキャンペーンの成果	—

※Y（口座開設）を（失敗）,1（成功）に変換した際、欠損値(Unknown)は0と変換した。

出所）Deep Analytics （https://deepanalytics.jp/compe/1?tab=compedetail）を基に筆者作成

分析方法

①ロジスティック回帰分析

単純主効果のみ。一つの説明変数が独立して目的変数に影響する

目的変数が1か0をとる場合に使う手法

今回は銀行口座を開設した（1）開設しなかった（0）

②ツリー（決定木）

交互作用：ある説明変数の効果が、他の説明変数の値により異なる

データを樹形図の形で分類できる

今回の目的は27,168名の顧客データを分類することで、どんな人が口座を開設したのかを分析する→ツリーのほうが詳細な考察に適している

分析方法

ロジスティック回帰と決定木の違い

ロジスティック回帰分析

- 線形回帰分析を修正した分析手法であり、S字型の曲線によって複数変数間の相関関係を理解し、その曲線を導く関数式によって事象の発生確率を予測する手法。

決定木

- IF THEN（もし～が発生したら、その次に～が発生する）形式でデータを分類／予測する手法。各行の類似度合いに基づいて行の分類を行い、その分岐条件を IF THEN のモデルで記述する。当該モデルを用いて分類、もしくは予測に活用可能。

出所：http://blogs.teradata.com/international/ja/hhg8/ 分析手法の種類 | マーケターのための データマイニング・ヒッチハイクガイド

分析方法

まず初めに

- Y（口座開設をしたか）を目的変数とした分析

を行い、口座開設をした顧客の特徴を調べた。

その結果、ツリーの一番上でduration(最終接触時間）が521.5以上であるかどうかで分岐したため、さらに

- duration（521.5以上なら1、未満なら0）を目的変数

として再び分析を行った。

ロジスティック回帰結果（目的変数：口座開設）

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.415e+00  2.187e-01 -11.041  < 2e-16 ***
age               -6.695e-04  2.779e-03  -0.241 0.809636
jobblue-collar    -3.545e-01  9.140e-02  -3.879 0.000105 ***
jobentrepreneur   -4.021e-01  1.556e-01  -2.584 0.009758 **
jobhousemaid      -4.541e-01  1.698e-01  -2.675 0.007482 **
jobmanagement     -1.493e-01  9.229e-02  -1.617 0.105799
jobretired         3.635e-01  1.206e-01   3.014 0.002576 **
jobself-employed  -3.429e-01  1.402e-01  -2.445 0.014486 *
jobservices       -3.170e-01  1.059e-01  -2.995 0.002746 **
jobstudent         5.694e-01  1.354e-01   4.207 2.59e-05 ***
jobtechnician     -3.211e-01  8.700e-02  -3.691 0.000224 ***
jobunemployed     -1.367e-01  1.377e-01  -0.993 0.320765
jobunknown        -4.568e-01  3.054e-01  -1.496 0.134700
maritalmarried    -1.021e-01  7.575e-02  -1.347 0.177877
maritalsingle      2.113e-01  8.584e-02   2.461 0.013851 *
educationsecondary 2.638e-01  8.222e-02   3.209 0.001333 **
educationtertiary  3.778e-01  9.564e-02   3.950 7.80e-05 ***
educationunknown   3.884e-01  1.294e-01   3.003 0.002678 **
defaultyes         1.147e-01  1.932e-01   0.594 0.552805
balance            1.376e-05  6.443e-06   2.135 0.032759 *
housingyes        -8.112e-01  5.145e-02 -15.769  < 2e-16 ***
loanyes           -5.592e-01  7.493e-02  -7.463 8.47e-14 ***
contacttelephone  -7.809e-02  9.468e-02  -0.825 0.409478
contactunknown    -1.166e+00  7.470e-02 -15.614  < 2e-16 ***
day               -4.880e-03  2.764e-03  -1.766 0.077403 .
duration           4.101e-03  8.138e-05  50.394  < 2e-16 ***
campaign          -1.217e-01  1.350e-02  -9.013  < 2e-16 ***
pdays             -4.139e-05  3.932e-04  -0.105 0.916167
previous           7.565e-03  6.539e-03   1.157 0.247319
poutcomeother      1.616e-01  1.134e-01   1.425 0.154239
poutcomesuccess    2.238e+00  1.022e-01  21.902  < 2e-16 ***
poutcomeunknown   -3.339e-01  1.149e-01  -2.905 0.003675 **
--+
```

有意になった変数（青字は負で有意）

- jobblue-collar***　肉体労働
- Jobentrepreneur**　企業家
- Jobhousemaid**　家政婦
- Jobretired***　退職
- Jobself-employed*　自営業
- Jobservices**　サービス業
- Jobstudent***　学生
- Jobtechnician***　技術職
- Maritalsingle*　独身
- educationsecondary**　中高
- educationtertiary***　大学・専門
- educationunknown**　不明
- Balance*　年間平均残高
- housingyes***　住宅ローンあり
- loanyes***　個人ローンあり
- Contactunknown***　連絡方法不明
- Duration***　最終接触時間
- Campaign***　現キャンペーン接触回数
- Poutcomesuccess***　前回成功
- Poutcomeunknown**　前回不明

ロジスティック回帰分析結果（目的変数：口座開設）

	Estimate	Std. Error	z value	Pr (>|z|)
年齢	-6.695e-04	2.779e-03	-0.241	0.809636
肉体労働者	-3.545e-01	9.140e-02	-3.879	0.000105***
企業家	-4.021e-01	1.556e-01	-2.584	0.009758**
家政婦	-4.541e-01	1.698e-01	-2.675	0.007482**
マネジメント	-1.493e-01	9.229e-02	-1.617	0.105799
退職者	3.635e-01	1.206e-01	3.014	0.002576**
自営業	-3.429e-01	1.402e-01	-2.445	0.014486*
サービス業	-3.170e-01	1.059e-01	-2.995	0.002746**
学生	5.694e-01	1.354e-01	4.207	0.0000259***
技術職	-3.211e-01	8.700e-02	-3.691	0.000224***
無職	-1.367e-01	1.377e-01	-0.993	0.320765
職業不明	-4.568e-01	3.054e-01	-1.496	0.134700
既婚	-1.021e-01	7.575e-02	-1.347	0.177877
未婚	2.113e-01	8.584e-02	2.461	0.013851*

	Estimate	Std. Error	z value	Pr (>|z|)
中高卒	2.638e-01	8.222e-02	3.209	0.001333 **
大学、専門学校卒	3.778e-01	9.564e-02	3.950	7.80e-05 ***
学歴不明	3.884e-01	1.294e-01	3.003	0.002678 **
債務不履行あり	1.147e-01	1.932e-01	0.594	0.552805
年間平均残高	1.376e-05	6.443e-06	2.135	0.032759 *
住宅ローンあり	-8.112e-01	5.15E-02	-15.769	< 2e-16 ***
個人ローンあり	-5.592e-01	7.493e-02	-7.463	8.47e-14 ***
連絡先携帯電話	-7.809e-02	9.468e-02	-0.825	0.409478
連絡先不明	-1.166e+00	7.47E-02	-15.614	< 2e-16 ***
最終接触日	-4.880e-03	2.764e-03	-1.766	0.077403 .
最終接触時間	4.101e-03	8.138e-05	50.394	< 2e-16 ***
当キャンペーン接触回数	-1.217e-01	1.350e-02	-9.013	< 2e-16 ***
前キャンペーン接触後の日数	-4.139e-05	3.932e-04	-0.105	0.916167
当キャンペーン以前までの接触回数	7.565e-03	6.539e-03	1.157	0.247319
前回結果その他	1.616e-01	1.134e-01	1.425	0.154239
前回結果成功	2.238e+00	1.022e-01	21.902	< 2e-16 ***
前回結果不明	-3.339e-01	1.149e-01	-2.905	0.003675 **

サンプルサイズ：27,168

ロジスティック回帰分析結果（目的変数：口座開設）

今回のキャンペーンで口座を開設した人は…

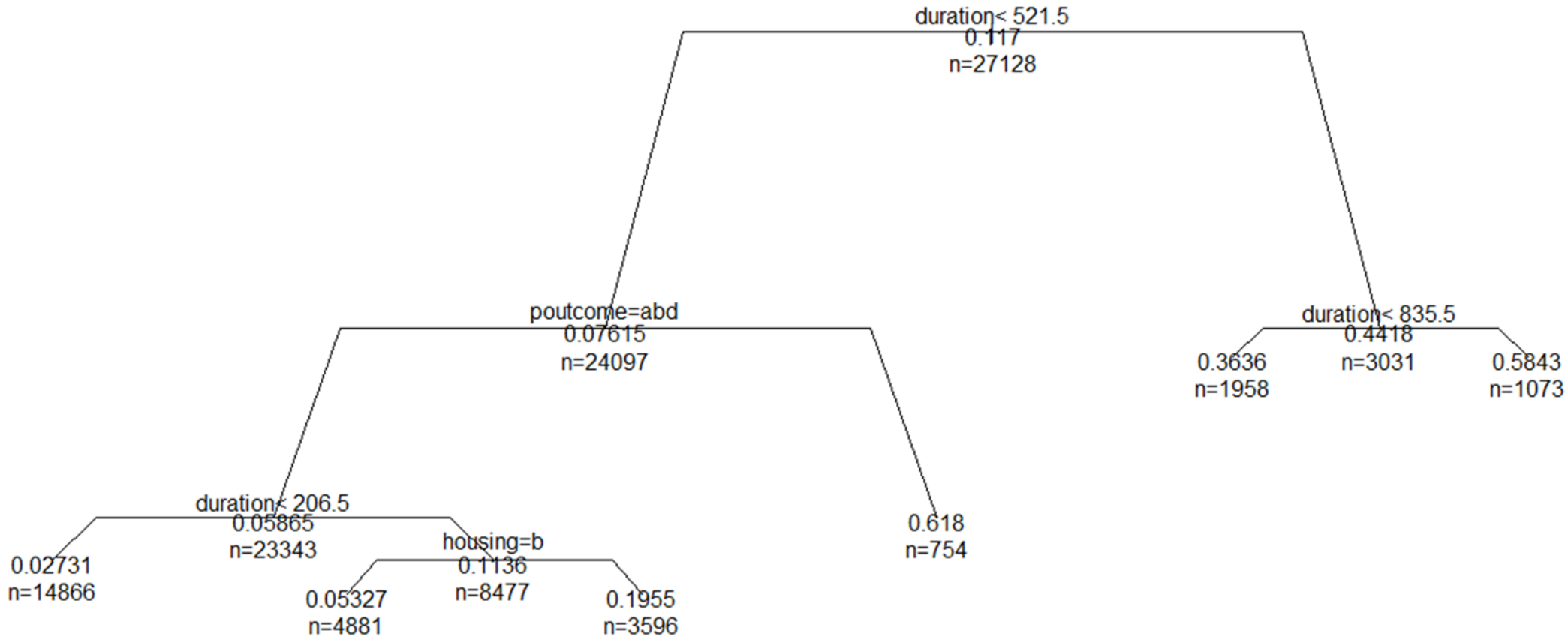
- 退職者や学生である
- 独身である
- 学歴が高い
- 年間平均残高が高い
- 住宅や個人ローンを組んでいない
- 最終接触時間が長い
- 今回のキャンペーンでの接触回数が少ない
- 前回のキャンペーンでも口座を開設している

決定木分析(目的変数:口座開設)

```
> rt <- rpart(y~age+job+marital+education+default+balance+housing+loan+contact+day+duration+campaign+pdays+previous+poutcome, data = tree)
> print(rt)
n= 27128

node), split, n, deviance, yval
      * denotes terminal node

 1) root 27128 2802.6390 0.11700090
   2) duration< 521.5 24097 1695.2640 0.07615056
     4) poutcome=failure,other,unknown 23343 1288.7120 0.05864713
       8) duration< 206.5 14866  394.9119 0.02731064 *
       9) duration>=206.5 8477  853.6017 0.11360150
        18) housing=yes 4881  246.1504 0.05326777 *
        19) housing=no 3596  565.5670 0.19549500 *
     5) poutcome=success 754  177.9947 0.61803710 *
   3) duration>=521.5 3031  747.4721 0.44176840
     6) duration< 835.5 1958  453.0909 0.36363640 *
     7) duration>=835.5 1073  260.6170 0.58434300 *
>
```

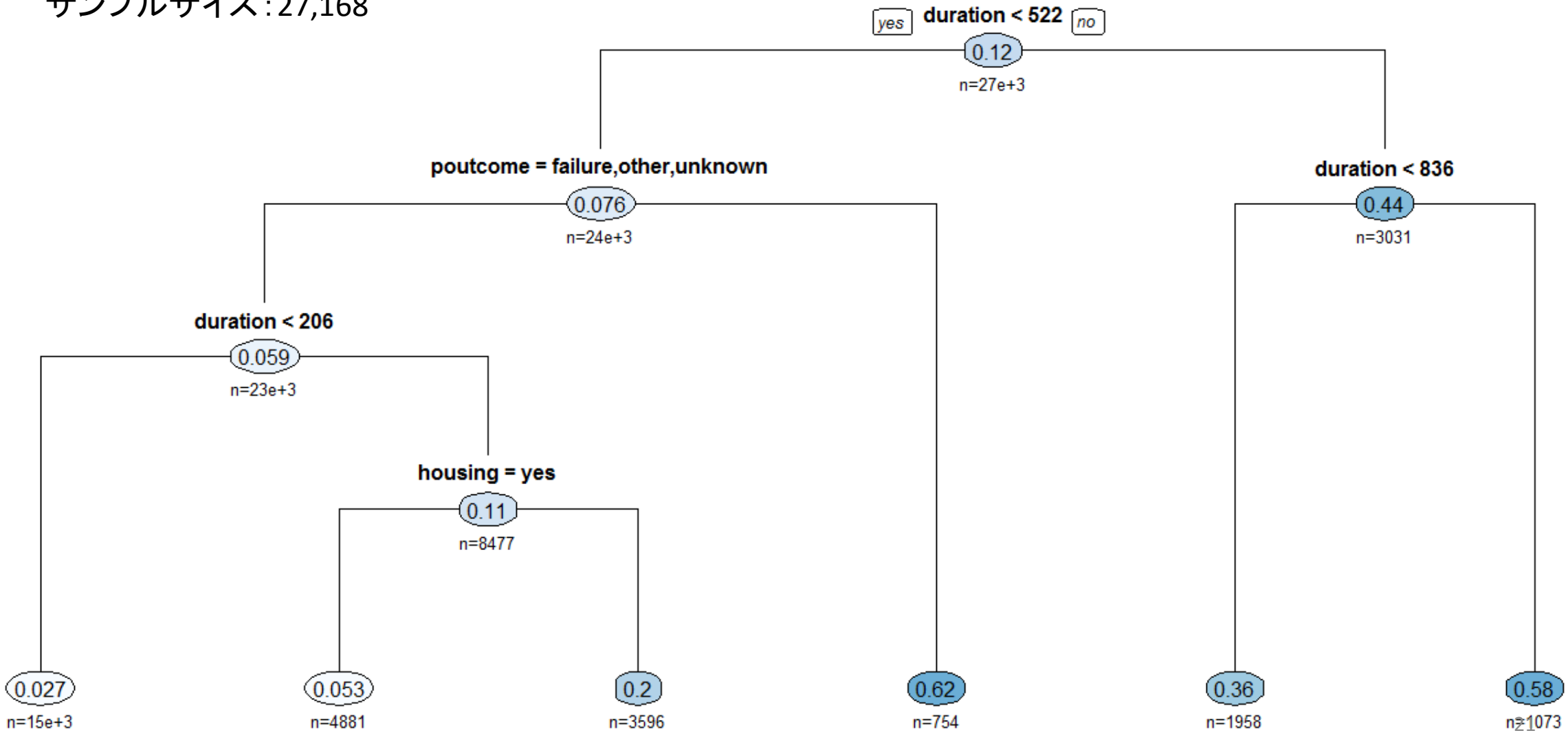
標準プロット

サンプルサイズ：27,168



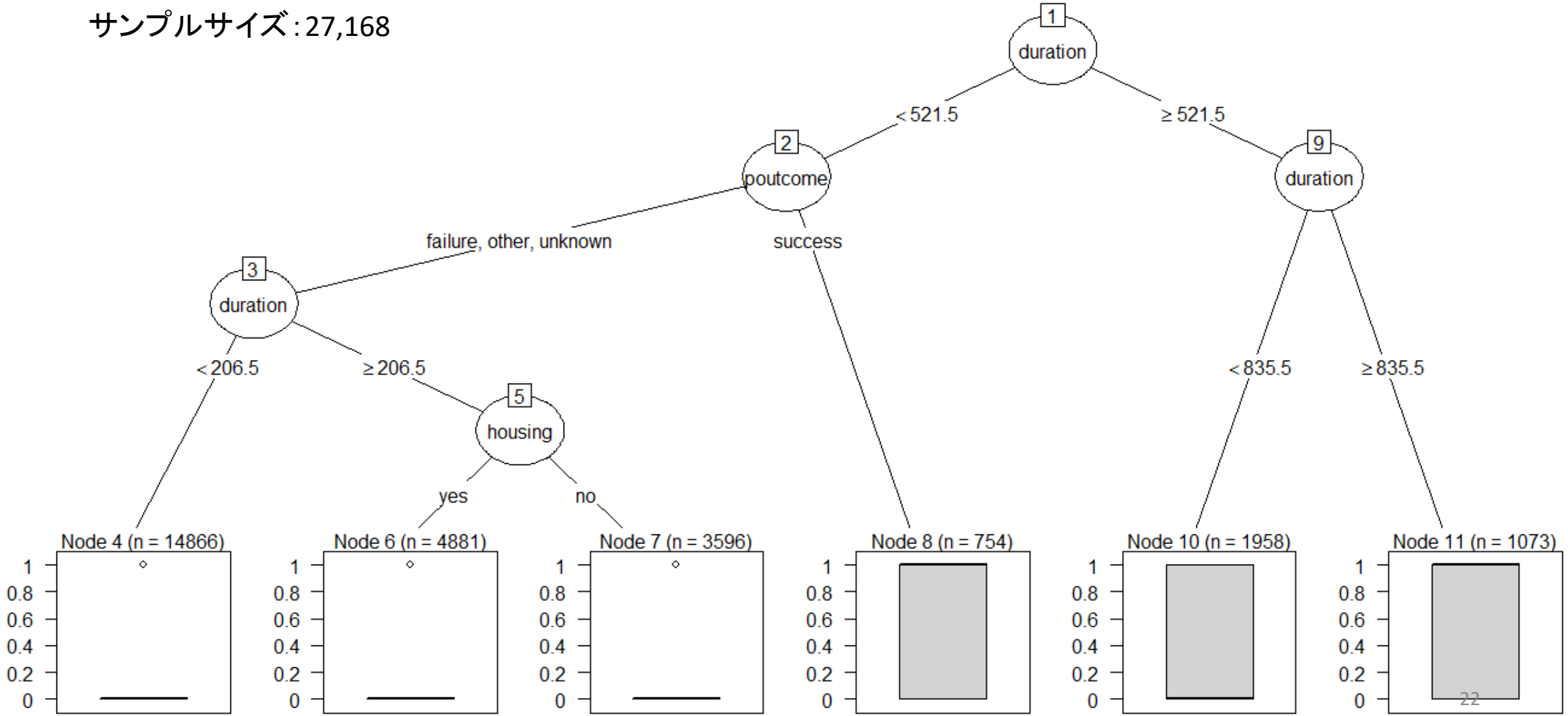
rpart.plot

サンプルサイズ：27,168



as.party

サンプルサイズ：27,168



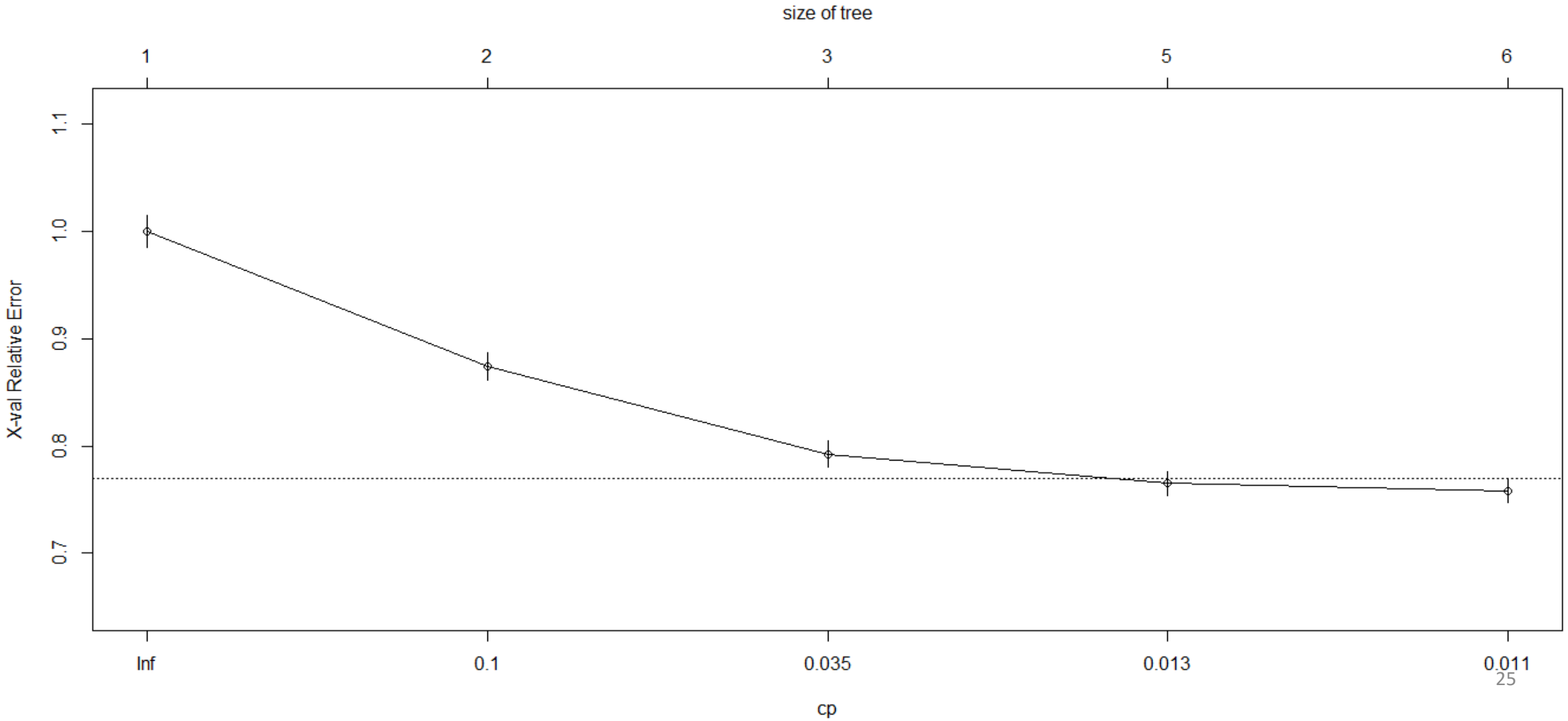
剪定

- 剪定：木が複雑になり過ぎたりした場合に回帰木の枝を切り落として見通しを良くする作業。
- 関数rpartでは、樹木を成長させると同時に交差確認法の結果も計算している
- 関数printcpは、樹木の剪定のための複雑さのパラメータ、つまり複雑度(cp)を返す
- 複雑度(cp)が低いほど、選定の基準が低くなり、より分岐が細かくなる。0.011では構造が単純であるため、より詳細な結果を見るため、cpを0.003まで下げて分析を試した。

剪定

```
> plot(as.party(rt))
> printcp(rt)

Regression tree:
rpart(formula = y ~ age + job + marital + education + default +
    balance + housing + loan + contact + day + duration + campaign +
    pdays + previous + poutcome, data = tree)

Variables actually used in tree construction:
[1] duration housing  poutcome

Root node error: 2802.6/27128 = 0.10331

n= 27128

        CP nsplit rel error  xerror     xstd
1 0.128416      0   1.00000 1.00013 0.014471
2 0.081551      1   0.87158 0.87436 0.012933
3 0.014644      2   0.79003 0.79261 0.012056
4 0.012047      4   0.76075 0.76502 0.011141
5 0.010000      5   0.74870 0.75824 0.011254
```

剪定を考えるため、printcp関数とplotcp関数を実行

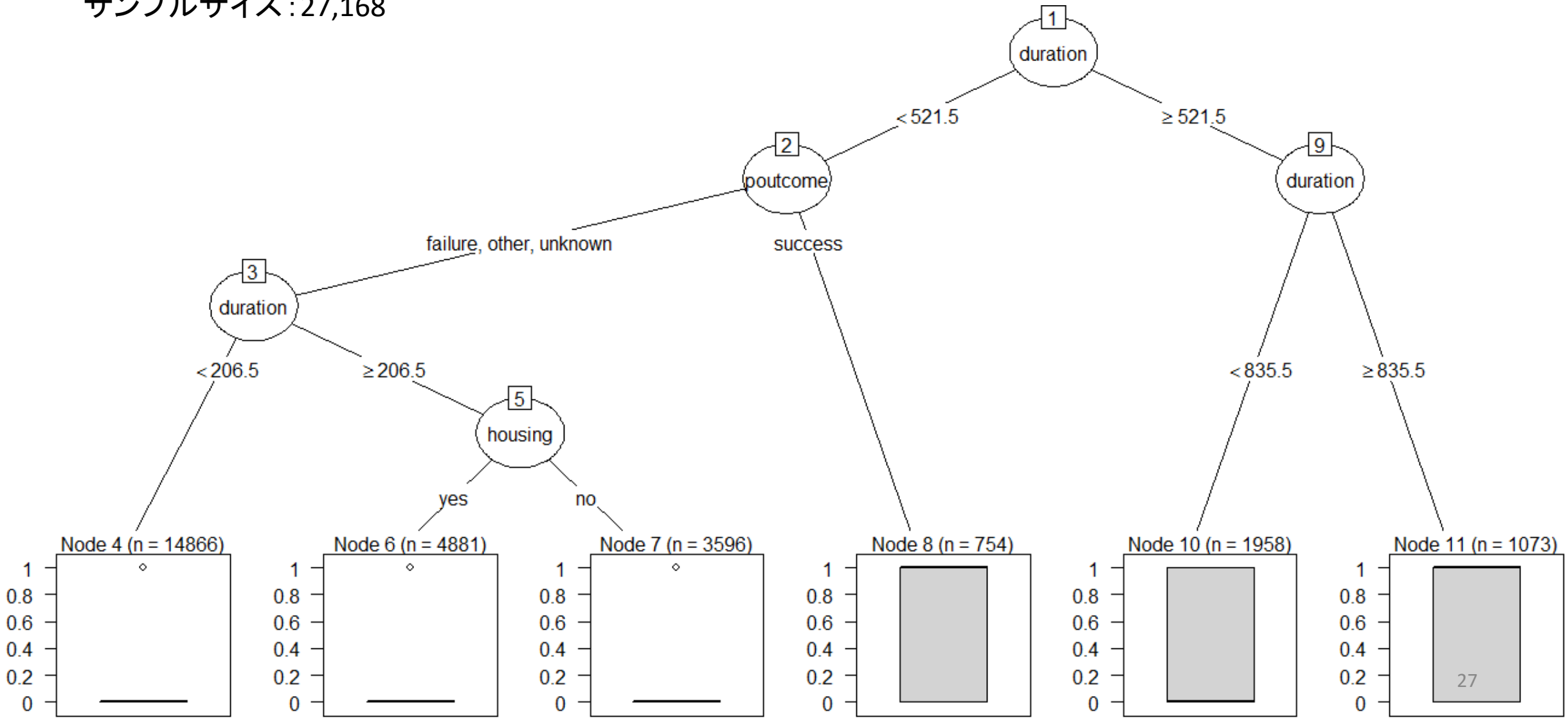


剪定の基準をcp＝0.011として再度分析

```
> rt2 <- rpart(y~age+job+marital+education+default+balance+housing+loan+contact+day+duration+campaign+pdays+previous+poutcome, data = tree, cp = 0.011)
> print(rt2)
n= 27128

node), split, n, deviance, yval
      * denotes terminal node

 1) root 27128 2802.6390 0.11700090
   2) duration< 521.5 24097 1695.2640 0.07615056
     4) poutcome=failure,other,unknown 23343 1288.7120 0.05864713
       8) duration< 206.5 14866  394.9119 0.02731064 *
       9) duration>=206.5 8477  853.6017 0.11360150
        18) housing=yes 4881  246.1504 0.05326777 *
        19) housing=no 3596  565.5670 0.19549500 *
     5) poutcome=success 754  177.9947 0.61803710 *
   3) duration>=521.5 3031  747.4721 0.44176840
     6) duration< 835.5 1958  453.0909 0.36363640 *
     7) duration>=835.5 1073  260.6170 0.58434300 *
> printcp(rt)
```

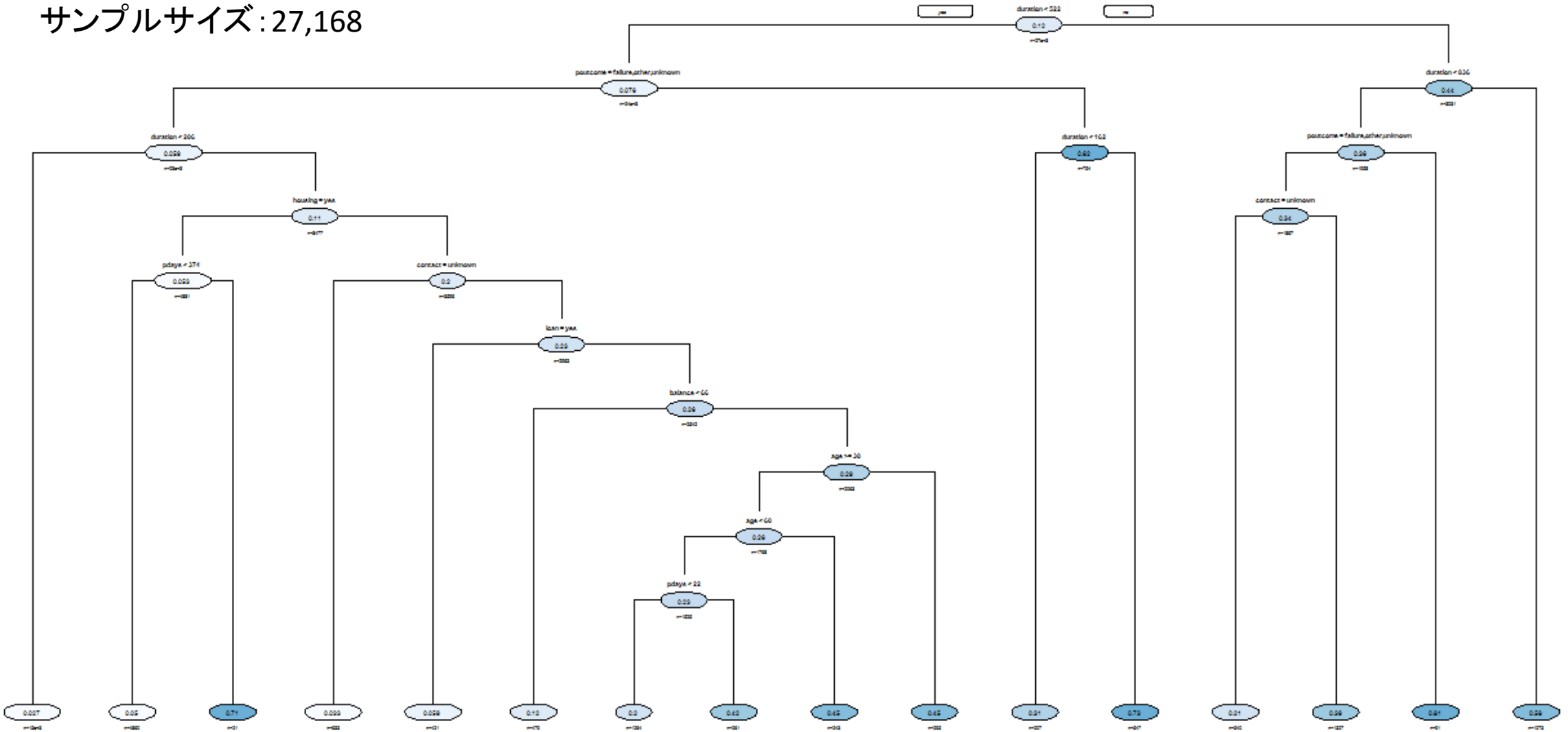
剪定後as.party

サンプルサイズ：27,168



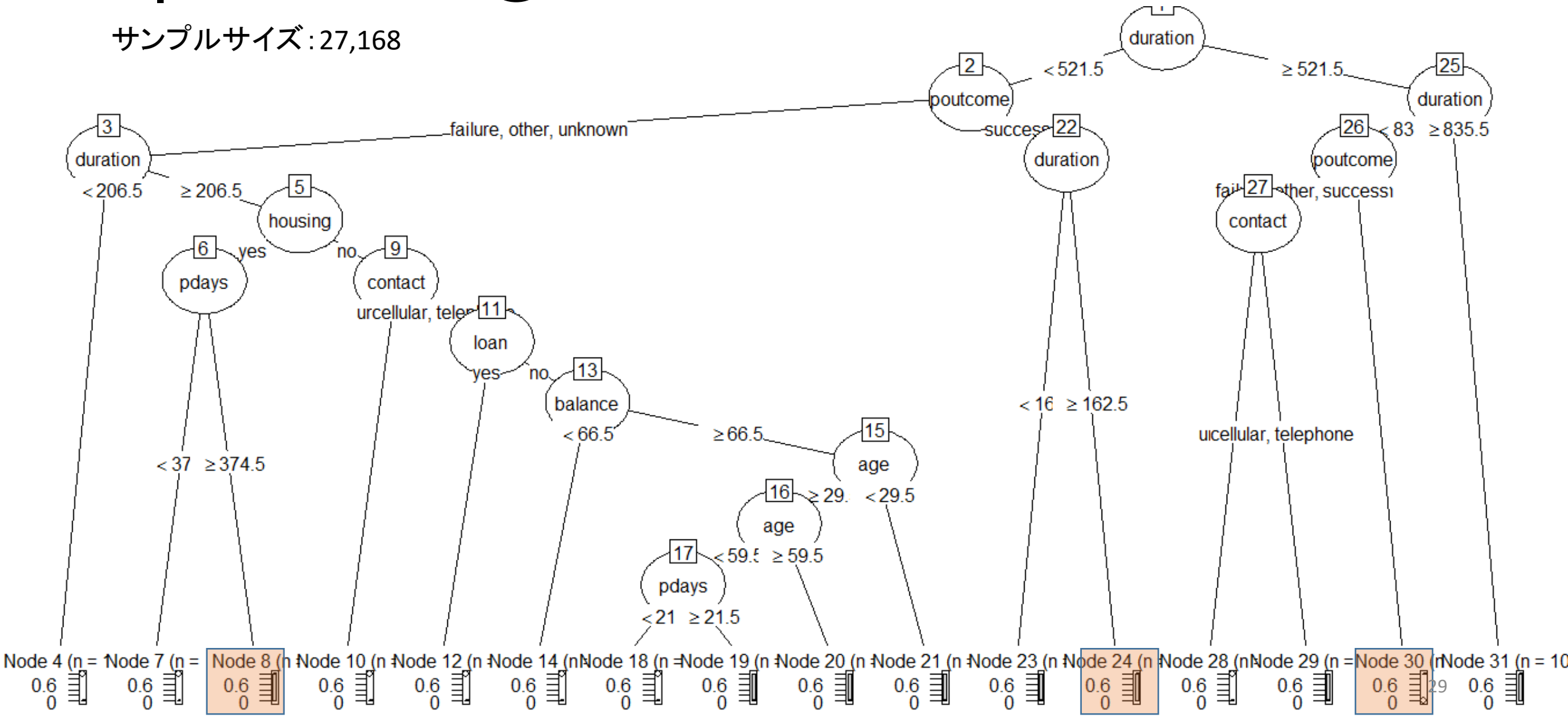
cp＝0.003

サンプルサイズ：27,168



cp＝0.003　②

サンプルサイズ：27,168



ロジスティック回帰分析結果(目的変数:最終接触時間)

	Estimate	Std. Error	z value	Pr (>|z|)
年齢	-3.021e-03	2.395e-03	-1.261	0.20714
肉体労働者	1.789e-01	7.565e-02	2.365	0.01802 *
企業家	2.058e-01	1.208e-01	1.704	0.08842 .
家政婦	2.880e-02	1.395e-01	0.206	0.83643
マネジメント	6.130e-02	8.424e-02	0.728	0.46682
退職者	4.392e-01	1.106e-01	3.969	7.21e-05 ***
自営業	2.139e-01	1.186e-01	1.803	0.07134 .
サービス業	1.756e-02	8.899e-02	0.197	0.84359
学生	-2.293e-01	1.638e-01	-1.400	0.16143
技術職	6.104e-02	7.712e-02	0.791	0.42866
無職	3.368e-01	1.205e-01	2.796	0.00517 **
職業不明	2.408e-03	2.669e-01	0.009	0.99280
既婚	-1.393e-01	6.193e-02	-2.250	0.02445 *
未婚	-2.302e-02	7.150e-02	-0.322	0.74745

	Estimate	Std. Error	z value	Pr (>|z|)
中高卒	-1.505e-02	6.173e-02	-0.244	0.80737
大学、専門学校卒	2.518e-02	7.717e-02	0.326	0.74422
学歴不明	9.829e-02	1.085e-01	0.906	0.36515
債務不履行あり	-1.648e-01	1.573e-01	-1.047	0.29495
年間平均残高	1.598e-05	5.690e-06	2.808	0.00498 **
住宅ローンあり	1.326e-01	4.297e-02	3.086	0.00203 **
個人ローンあり	-2.600e-02	5.421e-02	-0.480	0.63149
連絡先携帯電話	-1.734e-01	8.698e-02	-1.993	0.04624 *
連絡先不明	-1.952e-01	4.758e-02	-4.103	4.08e-05 ***
最終接触日	-5.108e-05	2.367e-03	-0.022	0.98278
当キャンペーン接触回数	-3.104e-02	7.595e-03	-4.087	4.37e-05 ***
前キャンペーン接触後の日数	1.378e-04	4.227e-04	0.326	0.74434
当キャンペーン以前までの接触回数	-6.029e-03	1.270e-02	-0.475	0.63498
前回結果その他	3.424e-02	1.183e-01	0.289	0.77226
前回結果成功	5.108e-01	1.188e-01	4.301	1.70e-05 ***
前回結果不明	2.772e-01	1.284e-01	2.159	0.03087 *

サンプルサイズ：27,168

ロジスティック回帰分析結果（目的変数：最終接触時間)

有意になった変数（青は負で有意）

- 肉体労働者 *
- 企業家 .
- 退職者 ***
- 自営業 .
- 無職 **
- 既婚*
- 年間平均残高 **
- 住宅ローンあり **
- 連絡方法が固定電話 *
- 連絡方法が不明 ***
- 現キャンペーンでの接触回数 ***
- 前回のキャンペーンで成功 ***
- 前回の結果は不明 *

ロジスティック回帰分析結果（目的変数：最終接触時間)

- 有職者では肉体労働者、企業家、自営業
- 退職者、無職
- 年間平均残高 が高い
- 住宅ローンを組んでいる
- 連絡方法が固定電話でも不明でもない
- 現キャンペーンでの接触回数が多くない
- 前回のキャンペーンでも口座を開設している、または結果不明

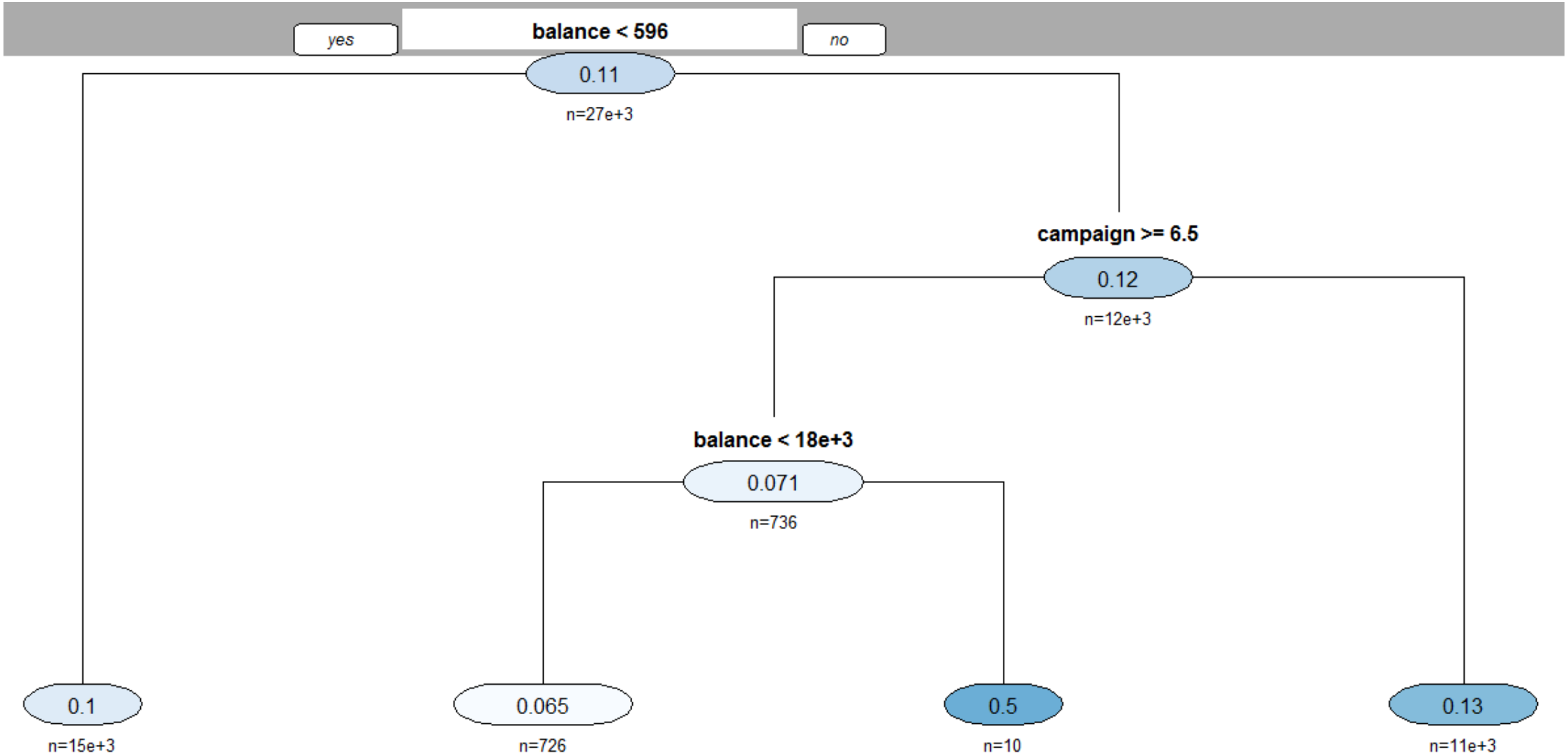
決定木分析(目的変数:最終接触時間)

```
> rpart(duration2~age+job+marital+education+default+balance+housing+loan+contact+day+campaign+pdays+previous+poutcome, data = x)
n= 27128

node), split, n, deviance, yval
      * denotes terminal node

1) root 27128 2692.348 0.1117296 *
> print(rt)
n= 27128

node), split, n, deviance, yval
      * denotes terminal node

1) root 27128 2692.348 0.1117296 *
>
```

決定木分析（目的変数：最終接触時間）

- 最初にcp＝0.001で剪定を行ったが、枝分かれせず根のみになってしまったので、cpを0.0006まで下げた結果、ちょうどよく枝分かれした。

cp=0.006


yes
balance < 596
no
0.11
n=27e+3
campaign >= 6.5
0.12
n=12e+3
balance < 18e+3
0.071
n=736
0.1
n=15e+3
0.065
n=726
0.5
n=10
0.13
n=11e+3

決定木分析結果（目的変数：最終接触時間）

- 分析結果から、最終接触時間には大きく分けて2つの変数が影響を与えることがわかった。
- 年間平均残高が1,800以上
- 当キャンペーン接触回数が6.5回以上

分析結果まとめ（目的変数：口座開設）

	ロジスティック回帰	ツリー
職	退職、学生	30歳未満or60歳以上
結婚	未婚	
学歴	中高以上	
債務不履行		
年間平均残高	高い	高い
住宅ローン	なし	なし
個人ローン	なし	なし
連絡方法		携帯電話
最終接触日		
最終接触時間	長い	長い
当キャンペーン接触回数	多い	
前キャンペーン後日数		長い
接触実績		
前回のキャンペーン成果	成功	成功

分析結果まとめ（目的変数：口座開設）

今回のキャンペーンで口座を開設した人は

- 最終接触時間が長い
- 前回のキャンペーンで口座をひらいた
- 年間平均残高が高い
- ローンなし
- 連絡方法が携帯電話である
- 30歳未満、もしくは60歳以上

という特徴があることが分かった

分析結果まとめ（目的変数：最終接触時間）

	ロジスティック回帰	ツリー
職	肉体労働者、企業家、自営業の人、退職者、無職	
結婚	未婚	
学歴		
債務不履行		
年間平均残高	高い	1,800以上
住宅ローン	あり	
個人ローン		
連絡方法	携帯or不明	
最終接触日		
最終接触時間		
当キャンペーン接触回数	少ない	6回以上
前キャンペーン後日数		
接触実績		
前回のキャンペーン成果	成功or不明	

分析結果まとめ　目的変数：最終接触時間

- ロジスティック回帰分析と決定木で重視されたのは、年間平均残高と、当キャンペーンでの接触回数
- 年間平均残高が高い人

→これは4つ行った全ての分析で有意

- 当キャンペーンでの接触回数は、分析ごとに結果が分かれてしまった。

実務へのインプリケーション

1. 接触時間を長くする
   - 毎月口座に振り込みのある人に対して、ATM利用手数料を月数回無料にする
2. 年間平均残高の高い人を取り込む
   - 金利を預金金額に応じて高くする
   - アプリを使ってその人の収入から最適な毎月の貯金額を提案したり、家計簿の機能を提供するなど、銀行側でも年間平均残高の増加を促す活動を行う。
3. 30歳未満の若年層や60歳以上の退職者を取り込む
   - ネット上で取引を行えるようにする
   - スマホアプリなど口座開設の方法を増やし、手軽さを印象付ける
   - 大学や地域センターへの出張説明会を行う
   - 友人紹介キャンペーンを行う
   - 子供、もしくは孫名義での口座を作らせる
4. 前回のキャンペーンで開設した層を取り込む
   - 複数の口座を持つことで特典が得られるようにする

実務へのインプリケーション

予想ではcampaign(現キャンペーンでの接触回数）とprevious(現キャンペーン以前までに顧客に接触した回数）が有意と予測した。campaignは分析ごとに結果が分かれ、　previousは一度も有意にならなかった。

- キャンペーン前に何らかの接触をしていても、その後のキャンペーン参加意図にはつながらない。コンスタントに顧客に接触していくことが必要
- 分析でduration(最終接触時間）が有意なのを見ると、量より質。いかに一度に長く話を聞いてもらうかが大切
- ATM画面に広告を出すなどして、自分から興味を持ってもらうことが重要ではないか

今後の課題

- より実務に役立てるため、今回使用した以外の変数、例えば性別や連絡時間帯などでも分析してみたい
- 次回はランダムフォレスト分析も行いたい

謝辞

今回の研究では、Deep Analytics（https://deepanalytics.jp/compe/1?tab=compedetail2016/10/15）のデータを使わせていただいた。

無償でのデータ公開に感謝いたします。

参考文献

- Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
- Bank+MarketingUCI https://archive.ics.uci.edu/ml/datasets/
- Deep Analytics https://deepanalytics.jp/compe/1?tab=compedetail （2016/10/15アクセス）
- [Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, GuimarÃ£es, Portugal, October, 2011. EUROSIS.
- Ruby Marketing http://rubymarketing.jp/blog/mere-exposure_effect/ (2016/10/17アクセス）
- Slide share http://www.slideshare.net/hamadakoichi/randomforest-web（2016/10/15アクセス）
- てっく煮ブログ　http://tech.nitoyon.com/ja/blog/2009/04/09/kmeans-visualise(2016/11/7アクセス)
- トライフィールズ　http://www.trifields.jp/decision-tree-classification-tree-1012(2016/11/22アクセス)
- 三井住友銀行公式ホームページ http://www.smbc.co.jp
- 三菱UFJ銀行公式ホームページ http://www.bk.mufg.jp/sp/
- 六本木ではたらくデータサイエンティストのブログhttp://tjo.hatenablog.com/entry/2013/12/24/190000（2016/10/15アクセス)
- ジャパンネット銀行公式ホームページ http://www.japannetbank.co.jp

付録

- K平均法
- R分析画面

K平均法

k平均法

・k平均法：データを似たもの同士でグループ分けする、というクラスタ分析の1つ。

方法

1. 各点にランダムにクラスタを割り当てる
2. クラスタの重心を計算する。
3. 点のクラスタを、一番近い重心のクラスタに変更する
4. 変化がなければ終了。変化がある限りは 2. に戻る

出所)てっく煮ブログ

※もとのデータ個数27129から100のサンプルを無作為抽出し、k平均法を行った。クラスタ数は10に設定した。

K平均法結果

```
> KM<-kmeans(x[,1:15],10,nstart=100,algorithm="Hartigan-Wong")
> KM
K-means clustering with 10 clusters of sizes 30, 1, 2, 4, 3, 11, 8, 14, 24, 3

Cluster means:
         job   marital education   default    balance   housing      loan   contact      day  duration campaign    pdays  previous   poutcome          y
1  0.9000000 0.4333333  0.900000 0.9333333  -122.2000 0.3666667 0.7000000 0.6333333 17.60000  137.1333 2.700000 56.80000 1.0333333 0.03333333 0.06666667
2  1.0000000 0.0000000  1.000000 1.0000000   244.0000 0.0000000 1.0000000 1.0000000 12.00000 1735.0000 4.000000 -1.00000 0.0000000 0.00000000 1.00000000
3  0.5000000 0.5000000  0.500000 1.0000000 20153.0000 0.5000000 1.0000000 1.0000000 11.50000  272.0000 2.000000 -1.00000 0.0000000 0.00000000 0.00000000
4  1.0000000 0.7500000  1.000000 1.0000000  5239.7500 0.7500000 0.7500000 0.7500000 21.00000  155.7500 1.500000 -1.00000 0.0000000 0.00000000 0.00000000
5  1.0000000 0.3333333  1.000000 1.0000000 12253.0000 0.6666667 1.0000000 1.0000000 17.00000  114.0000 2.000000 56.00000 0.3333333 0.00000000 0.00000000
6  0.8181818 0.3636364  1.090909 1.0000000  1865.7273 0.8181818 0.9090909 0.5454545 14.54545  230.7273 2.272727 -1.00000 0.0000000 0.00000000 0.18181818
7  0.8750000 0.5000000  0.875000 1.0000000  2860.7500 0.3750000 1.0000000 0.8750000 14.87500  339.6250 2.000000 42.25000 0.3750000 0.00000000 0.37500000
8  1.0000000 0.6428571  1.142857 1.0000000  1033.5714 0.2857143 0.8571429 0.4285714 10.50000  248.7143 2.714286 80.14286 1.0000000 0.00000000 0.07142857
9  0.8750000 0.5000000  1.208333 1.0000000   395.7917 0.3750000 0.9583333 0.6250000 14.16667  234.8333 2.458333 34.45833 0.4583333 0.04166667 0.16666667
10 1.0000000 0.3333333  1.333333 1.0000000  8039.3333 0.6666667 0.6666667 1.0000000 19.33333  278.3333 1.333333 -1.00000 0.0000000 0.00000000 0.00000000

Clustering vector:
  [1]  9  8  9  8  1  6  8  1  8  7  1  1  3  6  9  7  1  7  1  1  1  1  7  9  6  9  1  1  8  9  4  1  4  1  6  9  6  1  6  6  3  9  8  1  9  1  9  1  1  1
 [51]  1  9  8 10  8  1  1 10  1  7  4  9  4  8  8  6  8  9  9  9  9  2  9  1  5  7  7  9  6  5  5  8  9  7  1  9  1  6  9  9  6  1  9  1  8  8 10  1  9  1

Within cluster sum of squares by cluster:
 [1] 2492547.2       0.0  442984.5  497087.5 3695162.0 1160433.3 1313948.0 1271240.2 1790873.6 1532889.3
 (between_SS / total_SS =  99.0 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"
> |
```

K平均法考察

・yの値が最も高いクラスタは2(y=1　前回口座開設した)
・各項目詳細:
　職　あり
　未婚
　中・高卒
　債務不履行　なし
　年間平均残高244
　住宅ローン　あり
　個人ローン　なし
　連絡方法　携帯
　最終接触日　12日
　最終接触時間　1735秒
　現キャンペーンにおける接触回数　4回
　前キャンペーン後の日数　-1日
　接触実績　なし
　前回のキャンペーンの成果　失敗・不明

K平均法考察

・クラスタ2の特徴：未婚・年間平均残高低い・住宅ローンがある・現キャンペーンでの接触回数が多く、前キャンペーンから日数も経過していない

→未婚で年間平均残高が低く、ローンも残っているということから、将来の多くのお金を必要とする時のために貯金が必要だと考えている消費者が多いのではないか。

→現キャンペーンでの接触回数が多いほど口座開設しやすいのではないか。

K平均法考察

・逆に、yの値が低かったのはクラスタ3，4，5，10(y=0　前回口座開設せず)

・クラスタ3，4，5に共通している特徴：年間平均残高が高い・現キャンペーンでの接触回数が少ない

→年間平均残高が高いことから、生活に余裕がある消費者が多いと考えられる。将来への不安もそれほどなく口座開設を必要としていない、もしくは既に他の銀行の口座を持っているため、口座開設しなかったのではないか。

→前キャンペーンからの日数はクラスタ5を除き、クラスタ2と同じ-1日だが、接触回数がクラスタ2は4回であったのに比べ、3，4，5，10では1.2回と少ない。これからも現キャンペーンでの接触回数が多い方が口座開設をしやすいことがわかる。

```
> m1<-glm(y~age+job+marital+education+default+balance+housing+loan+contact+day+duration+campaign+pdays+previous+poutcome,data=y,  binomial)
> summary(m1)

Call:
glm(formula = y ~ age + job + marital + education + default +
    balance + housing + loan + contact + day + duration + campaign +
    pdays + previous + poutcome, family = binomial, data = y)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.7212  -0.4047  -0.2686  -0.1583   3.0275

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.415e+00  2.187e-01 -11.041  < 2e-16 ***
age                 -6.695e-04  2.779e-03  -0.241 0.809636
jobblue-collar      -3.545e-01  9.140e-02  -3.879 0.000105 ***
jobentrepreneur     -4.021e-01  1.556e-01  -2.584 0.009758 **
jobhousemaid        -4.541e-01  1.698e-01  -2.675 0.007482 **
jobmanagement       -1.493e-01  9.229e-02  -1.617 0.105799
jobretired           3.635e-01  1.206e-01   3.014 0.002576 **
jobself-employed    -3.429e-01  1.402e-01  -2.445 0.014486 *
jobservices         -3.170e-01  1.059e-01  -2.995 0.002746 **
jobstudent           5.694e-01  1.354e-01   4.207 2.59e-05 ***
jobtechnician       -3.211e-01  8.700e-02  -3.691 0.000224 ***
jobunemployed       -1.367e-01  1.377e-01  -0.993 0.320765
jobunknown          -4.568e-01  3.054e-01  -1.496 0.134700
maritalmarried      -1.021e-01  7.575e-02  -1.347 0.177877
maritalsingle        2.113e-01  8.584e-02   2.461 0.013851 *
educationsecondary   2.638e-01  8.222e-02   3.209 0.001333 **
educationtertiary    3.778e-01  9.564e-02   3.950 7.80e-05 ***
educationunknown     3.884e-01  1.294e-01   3.003 0.002678 **
defaultyes           1.147e-01  1.932e-01   0.594 0.552805
balance              1.376e-05  6.443e-06   2.135 0.032759 *
housingyes          -8.112e-01  5.145e-02 -15.769  < 2e-16 ***
loanyes             -5.592e-01  7.493e-02  -7.463 8.47e-14 ***
contacttelephone    -7.809e-02  9.468e-02  -0.825 0.409478
contactunknown      -1.166e+00  7.470e-02 -15.614  < 2e-16 ***
day                 -4.880e-03  2.764e-03  -1.766 0.077403 .
duration             4.101e-03  8.138e-05  50.394  < 2e-16 ***
campaign            -1.217e-01  1.350e-02  -9.013  < 2e-16 ***
pdays               -4.139e-05  3.932e-04  -0.105 0.916167
previous             7.565e-03  6.539e-03   1.157 0.247319
poutcomeother        1.616e-01  1.134e-01   1.425 0.154239
poutcomesuccess      2.238e+00  1.022e-01  21.902  < 2e-16 ***
poutcomeunknown     -3.339e-01  1.149e-01  -2.905 0.003675 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 19581  on 27127  degrees of freedom
Residual deviance: 13468  on 27096  degrees of freedom
AIC: 13532

Number of Fisher Scoring iterations: 6
```

```
Call:
glm(formula = duration2 ~ age + job + marital + education + default +
    balance + housing + loan + contact + day + campaign + pdays +
    previous + poutcome, family = binomial, data = x)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0209  -0.5075  -0.4757  -0.4400   2.7836

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.167e+00  2.027e-01 -10.689  < 2e-16 ***
age               -3.021e-03  2.395e-03  -1.261  0.20714
jobblue-collar     1.789e-01  7.565e-02   2.365  0.01802 *
jobentrepreneur    2.058e-01  1.208e-01   1.704  0.08842 .
jobhousemaid       2.880e-02  1.395e-01   0.206  0.83643
jobmanagement      6.130e-02  8.424e-02   0.728  0.46682
jobretired         4.392e-01  1.106e-01   3.969 7.21e-05 ***
jobself-employed   2.139e-01  1.186e-01   1.803  0.07134 .
jobservices        1.756e-02  8.899e-02   0.197  0.84359
jobstudent        -2.293e-01  1.638e-01  -1.400  0.16143
jobtechnician      6.104e-02  7.712e-02   0.791  0.42866
jobunemployed      3.368e-01  1.205e-01   2.796  0.00517 **
jobunknown         2.408e-03  2.669e-01   0.009  0.99280
maritalmarried    -1.393e-01  6.193e-02  -2.250  0.02445 *
maritalsingle     -2.302e-02  7.150e-02  -0.322  0.74745
educationsecondary -1.505e-02  6.173e-02  -0.244  0.80737
educationtertiary  2.518e-02  7.717e-02   0.326  0.74422
educationunknown   9.829e-02  1.085e-01   0.906  0.36515
defaultyes        -1.648e-01  1.573e-01  -1.047  0.29495
balance            1.598e-05  5.690e-06   2.808  0.00498 **
housingyes         1.326e-01  4.297e-02   3.086  0.00203 **
loanyes           -2.600e-02  5.421e-02  -0.480  0.63149
contacttelephone  -1.734e-01  8.698e-02  -1.993  0.04624 *
contactunknown    -1.952e-01  4.758e-02  -4.103 4.08e-05 ***
day               -5.108e-05  2.367e-03  -0.022  0.98278
campaign          -3.104e-02  7.595e-03  -4.087 4.37e-05 ***
pdays              1.378e-04  4.227e-04   0.326  0.74434
previous          -6.029e-03  1.270e-02  -0.475  0.63498
poutcomeother      3.424e-02  1.183e-01   0.289  0.77226
poutcomesuccess    5.108e-01  1.188e-01   4.301 1.70e-05 ***
poutcomeunknown    2.772e-01  1.284e-01   2.159  0.03087 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 18996  on 27127  degrees of freedom
Residual deviance: 18880  on 27097  degrees of freedom
AIC: 18942

Number of Fisher Scoring iterations: 5
```

```
> rt2 <- rpart(y~age+job+marital+education+default+balance+housing+loan+contact+day+duration+campaign+pdays+previous+poutcome, data = x, cp = 0.003)
> print(rt2)
n= 27128

node), split, n, deviance, yval
      * denotes terminal node

   1) root 27128 2802.639000 0.11700090
     2) duration< 521.5 24097 1695.264000 0.07615056
       4) poutcome=failure,other,unknown 23343 1288.712000 0.05864713
         8) duration< 206.5 14866  394.911900 0.02731064 *
         9) duration>=206.5 8477  853.601700 0.11360150
          18) housing=yes 4881  246.150400 0.05326777
            36) pdays< 374.5 4860  232.649200 0.05041152 *
            37) pdays>=374.5 21    4.285714 0.71428570 *
          19) housing=no 3596  565.567000 0.19549500
            38) contact=unknown 633   20.303320 0.03317536 *
            39) contact=cellular,telephone 2963  525.022600 0.23017210
              78) loan=yes 421   23.515440 0.05938242 *
              79) loan=no 2542  487.193200 0.25845790
               158) balance< 66.5 479   49.453030 0.11691020 *
               159) balance>=66.5 2063  425.914700 0.29132330
                 318) age>=29.5 1768  343.646500 0.26414030
                   636) age< 59.5 1525  273.426900 0.23409840
                    1272) pdays< 21.5 1264  199.341800 0.19620250 *
                    1273) pdays>=21.5 261   63.478930 0.41762450 *
                   637) age>=59.5 243   60.205760 0.45267490 *
                 319) age< 29.5 295   73.132200 0.45423730 *
       5) poutcome=success 754  177.994700 0.61803710
        10) duration< 162.5 207   44.212560 0.30917870 *
        11) duration>=162.5 547  106.563100 0.73491770 *
     3) duration>=521.5 3031  747.472100 0.44176840
       6) duration< 835.5 1958  453.090900 0.36363640
        12) poutcome=failure,other,unknown 1867  419.979600 0.34172470
          24) contact=unknown 540   89.933330 0.21111110 *
          25) contact=cellular,telephone 1327  317.085200 0.39487570 *
        13) poutcome=success 91   13.824180 0.81318680 *
       7) duration>=835.5 1073  260.617000 0.58434300 *
```