Latent Class Modeling in R

Lito P. Cruz

https://www.linkedin.com/in/l-p-cruz-phd/

https://unpocologico.wordpress.com/

Talk Outline

- Introduction
- What are Mixture Models?
- Latent Classes Species
- poLCA on Portugese Bank Marketing Dataset
  - Preparing the Data
  - Formulating the Models
  - Finding the Right One
- Using the Model
- Further on w depmixS4
- Conclusion Q & A

Note: A bit of Bayesian Statistics but don’t be scared

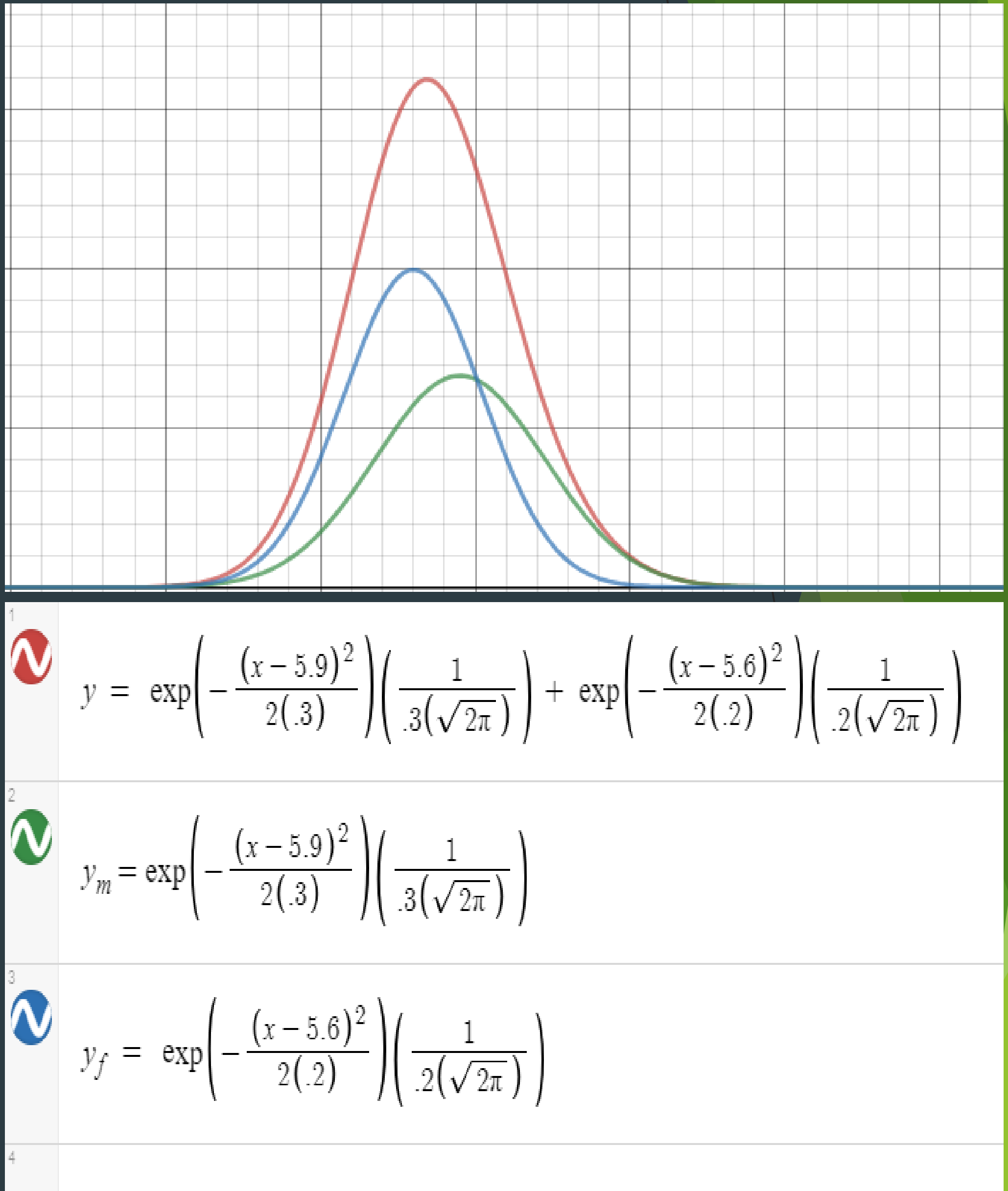
Introduction

- Latent Class Models are part of the family of Mixture Models (MM).
- Latent implies “hiddenness”, “concealed”, “unmanifested”, notion of “undisclosed” or “unobserved”
- In this talk we will focus on one – the Latent Class Analysis (LCA)
- It is generally a segmentation technique but since it is probabilistic (it is MM after all) it can be used for prediction too.
- Does not use distances for similarities *but uses membership probabilities estimated directly from the model.*
- Where has it been used for:
  - Health Sciences
  - Behavioral/Political Sciences
  - Economics/Econometrics
  - **Marketing Sciences (heavily used)**
    - Segmentation w/ Customer Lifetime Values, Recency/Frequency
- It has sister models w similar names that is why it gets confusing in the literature.

https://www.linkedin.com/in/l-p-cruz-phd/
https://unpocologico.wordpress.com/

What are Mixture Models?

- Assume we are studying the heights of people in a city like Melbourne.
- We get their heights and found out that their height distribution looks like the one on RED. This is what we see as the result of our data gathering.
- Unknown to us is that there are male and female heights that are causing that total sample height distribution to behave that way.
- The job of a MM is to find those HIDDEN hence, LATENT variables that are causing the RED to behave that way.
- These LATENT variables are the Classes LCA will discover!
- THAT IS GOOD ISN'T IT?



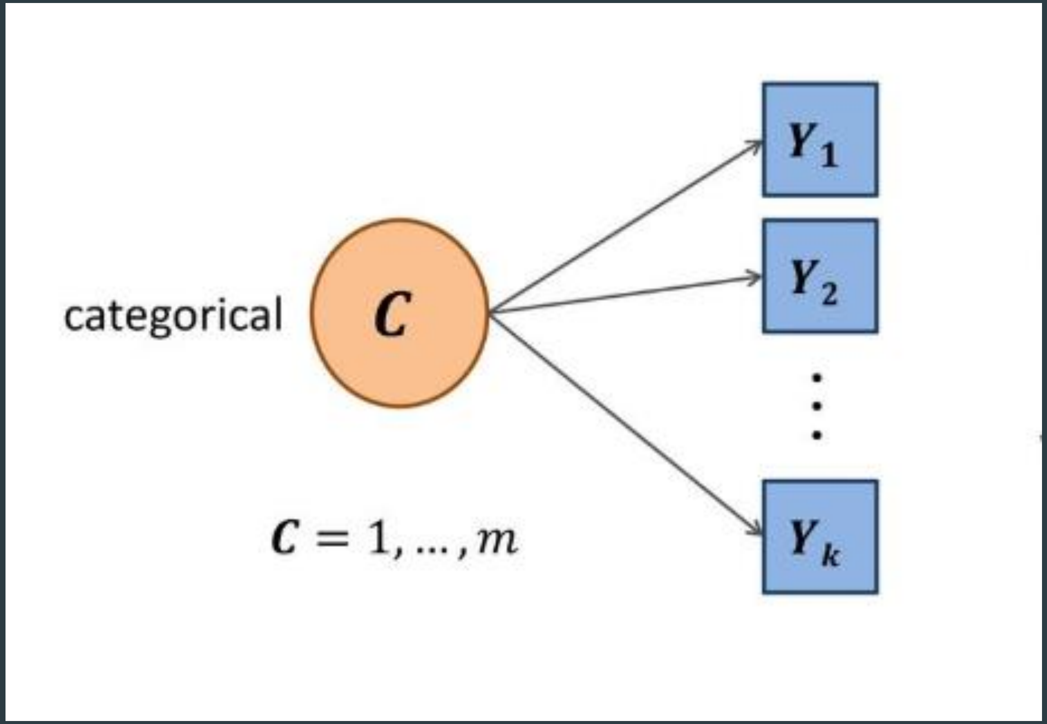
$$y = \exp\left(-\frac{(x-5.9)^2}{2(.3)}\right)\left(\frac{1}{.3\left(\sqrt{2\pi}\right)}\right) + \exp\left(-\frac{(x-5.6)^2}{2(.2)}\right)\left(\frac{1}{.2\left(\sqrt{2\pi}\right)}\right)$$

$$y_m = \exp\left(-\frac{(x-5.9)^2}{2(.3)}\right)\left(\frac{1}{.3\left(\sqrt{2\pi}\right)}\right)$$

$$y_f = \exp\left(-\frac{(x-5.6)^2}{2(.2)}\right)\left(\frac{1}{.2\left(\sqrt{2\pi}\right)}\right)$$

https://www.linkedin.com/in/l-p-cruz-phd/
https://unpocologico.wordpress.com/

VISUALIZING LATENT CLASSES



$Y_i = Categorical\ Variables$

$C = Latent\ Classes\ - suspected\ categorical\ variables\ influencing\ the\ Y_i$

https://www.linkedin.com/in/l-p-cruz-phd/
https://unpocologico.wordpress.com/

Latent Concepts and its Species

- We need known category variables in our dataset
  - Sometimes called indicators
  - Sometimes called manifest variables
- Covariates – these are variables that can secondarily affect the values of our indicators
- Latent Class - hidden or unknown variables, which should be named later
- HOW THEY ALL RELATE :

Indicator variables	Latent variable(s)	
	Discrete	***Continuous***
Discrete	**Latent Class Analysis**	Latent Trait Analysis
Continuous	Latent Profile Analysis	Factor Analysis

https://www.linkedin.com/in/l-p-cruz-phd/
https://unpocologico.wordpress.com/

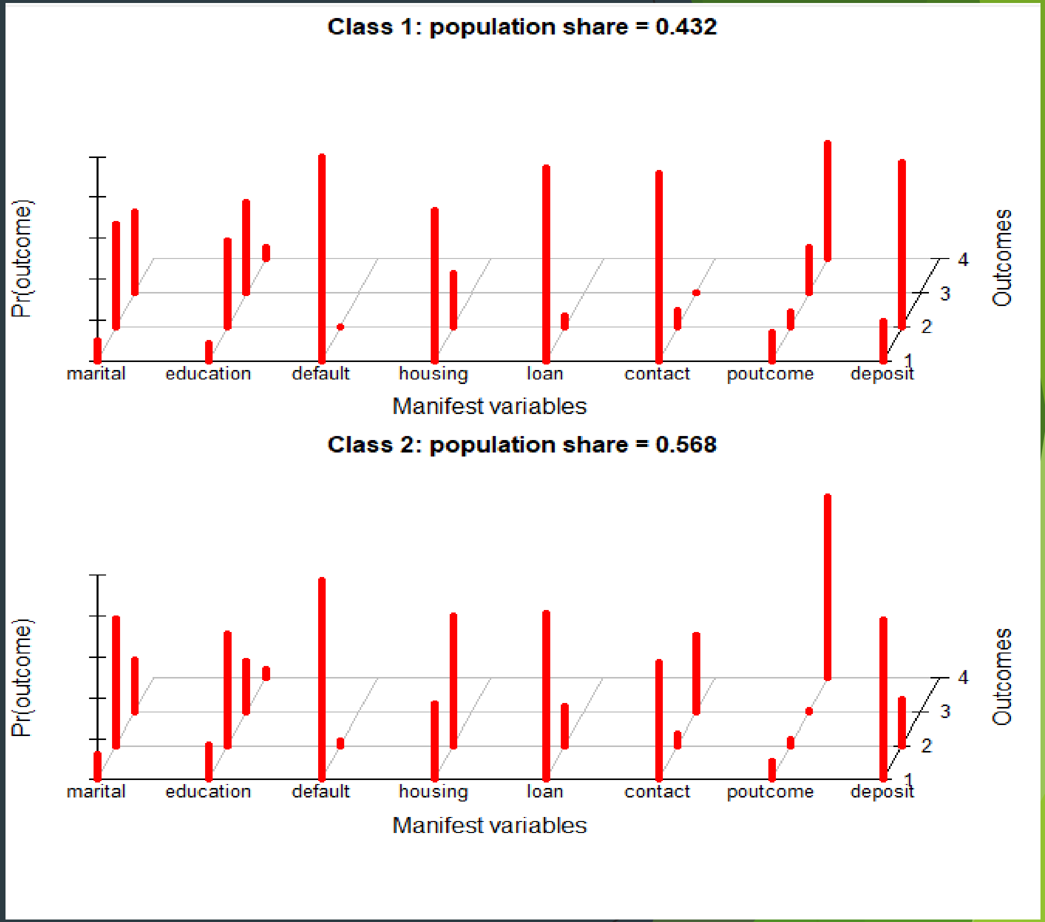
LCA in R

- MCLUST – Model based Clustering
- poLCA – Polytomous LCA – Many Valued Category Variables
- LCCA – Latent Class Causal Analysis
- randomLCA – Random Effects LCA
- BayesLCA – Bayesian LCA
- depmixS4 - Dependent Mixture Models - Hidden Markov Models of GLMs and Other Distributions in S4

https://www.linkedin.com/in/l-p-cruz-phd/
https://unpocologico.wordpress.com/

poLCA R Package

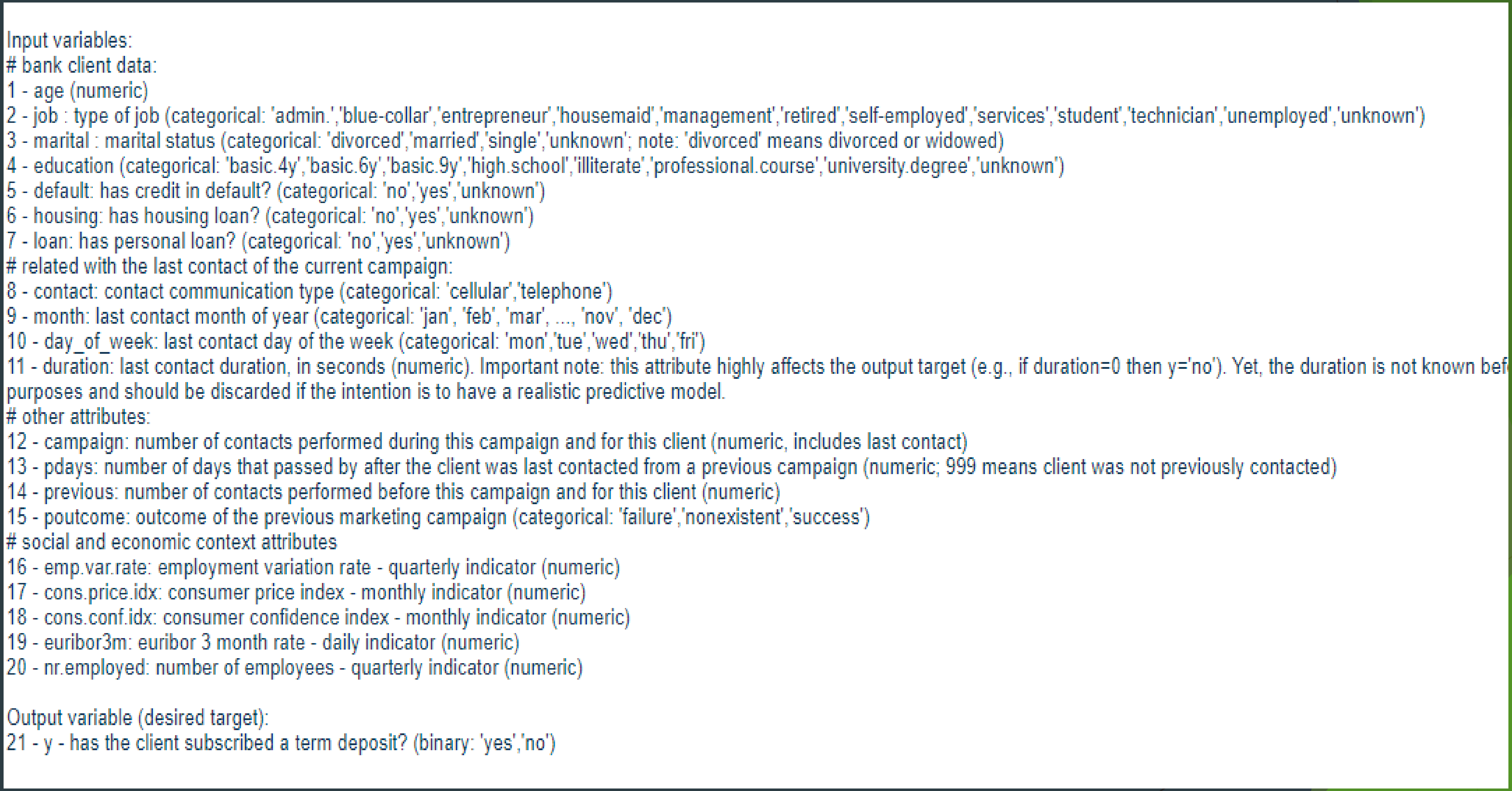
- It means the manifest variables have multi-level factor values or category levels - so multi-nominal vs binomial
- Has no modeling assumptions ie does not rely on knowledge of the distribution of the variables
- $P(Y_i|C_j)$ - given the data came from Class J what is the probability of Y_i obtaining a certain value?
- Supports latent class “regression” for class membership prediction
- No need for Principal Component Analysis then Linear Discriminant Analysis.
- Offers goodness of fit tests information



https://www.linkedin.com/in/l-p-cruz-phd/
https://unpocologico.wordpress.com/

The Portugese Bank Dataset for Marketing

https://archive.ics.uci.edu/ml/datasets/Bank+Marketing (modified Kaggle version)



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')
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). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known bef
purposes and should be discarded if the intention is to have a realistic predictive model.
# 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)
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)

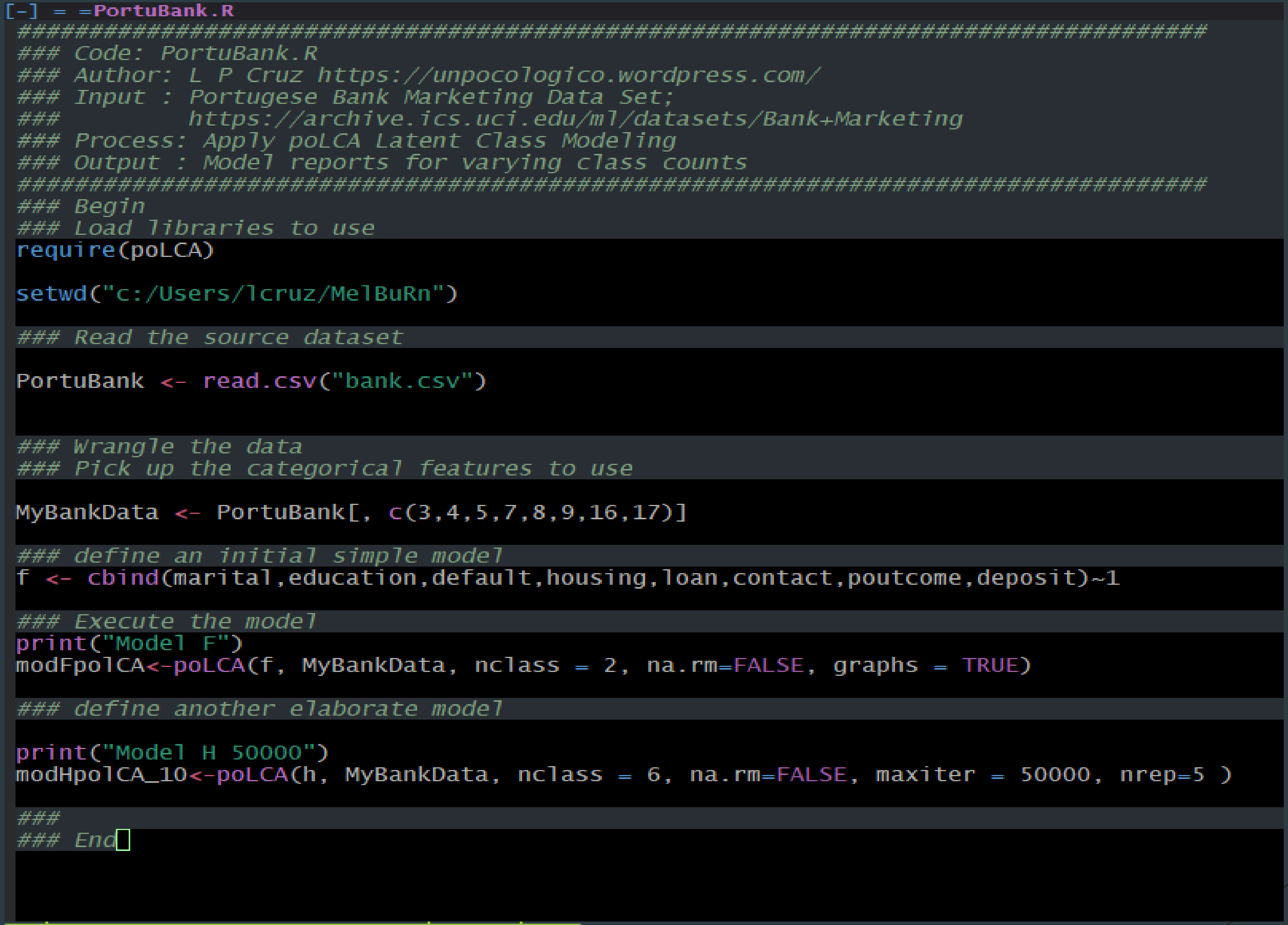
Output variable (desired target):
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

The Portugese Bank Dataset for Marketing poLCA Modeling Context

- The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.
- **Q : We want to know the types of people who say YES/NO to the term deposit deal.**
- poLCA can only consider Category/Factor Manifest Variables/Features/Attributes. For our case study we are going to examine the following
- Marital – divorced/married/single – marital status
- Education – primary/secondary/tertiary/unknown – highest level of education
- Default – yes/no – did the person have a credit in default?
- Housing – yes/no – does the person have a housing loan?
- Loan – yes/no – does the person have a personal loan?
- Contact – cellular/telephone/unknown – how the person got contacted
- Poutcome – failure/other/success/unknown
- Deposit – yes/no – the response to the current marketing campaign – did the person subscribe to the term deposit?

https://www.linkedin.com/in/l-p-cruz-phd/
https://unpocologico.wordpress.com/

The Code



```
###################################################################################
### Code: PortuBank.R
### Author: L P Cruz https://unpocologico.wordpress.com/
### Input : Portugese Bank Marketing Data Set;
###         https://archive.ics.uci.edu/ml/datasets/Bank+Marketing
### Process: Apply poLCA Latent Class Modeling
### Output : Model reports for varying class counts
###################################################################################
### Begin
### Load libraries to use
require(poLCA)

setwd("c:/Users/lcruz/MelBuRn")

### Read the source dataset

PortuBank <- read.csv("bank.csv")


### Wrangle the data
### Pick up the categorical features to use

MyBankData <- PortuBank[, c(3,4,5,7,8,9,16,17)]

### define an initial simple model
f <- cbind(marital,education,default,housing,loan,contact,poutcome,deposit)~1

### Execute the model
print("Model F")
modFpolCA<-poLCA(f, MyBankData, nclass = 2, na.rm=FALSE, graphs = TRUE)

### define another elaborate model

print("Model H 50000")
modHpolCA_10<-poLCA(h, MyBankData, nclass = 6, na.rm=FALSE, maxiter = 50000, nrep=5 )

###
### End
```

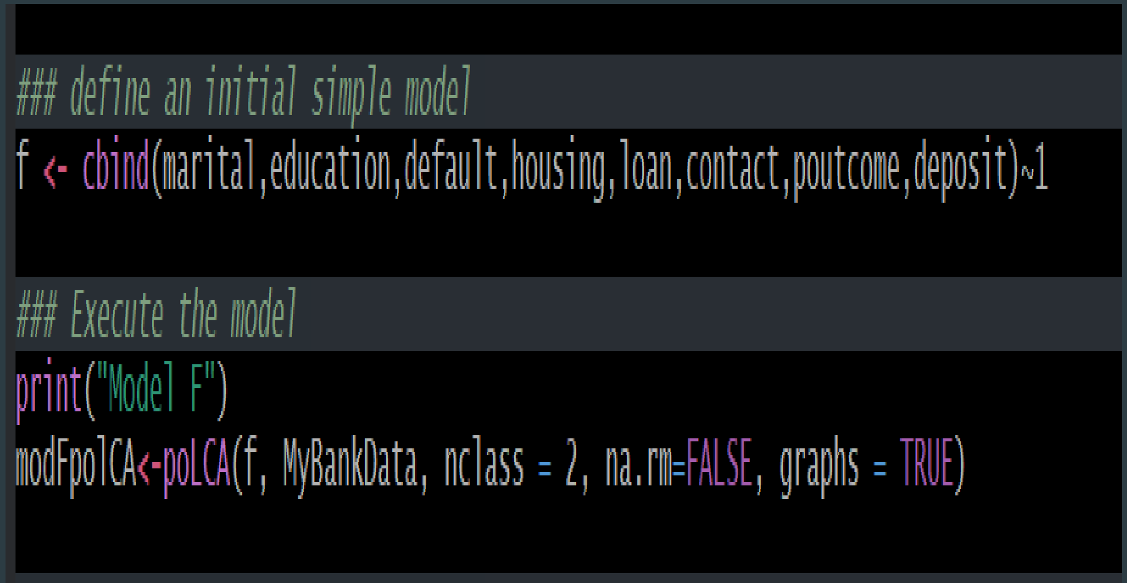
Notice the simplicity of the steps.

Thanks to R, ML codes tend to be short.

Short codes have less moving parts subject to breakages

https://unpocologico.wordpress.com/

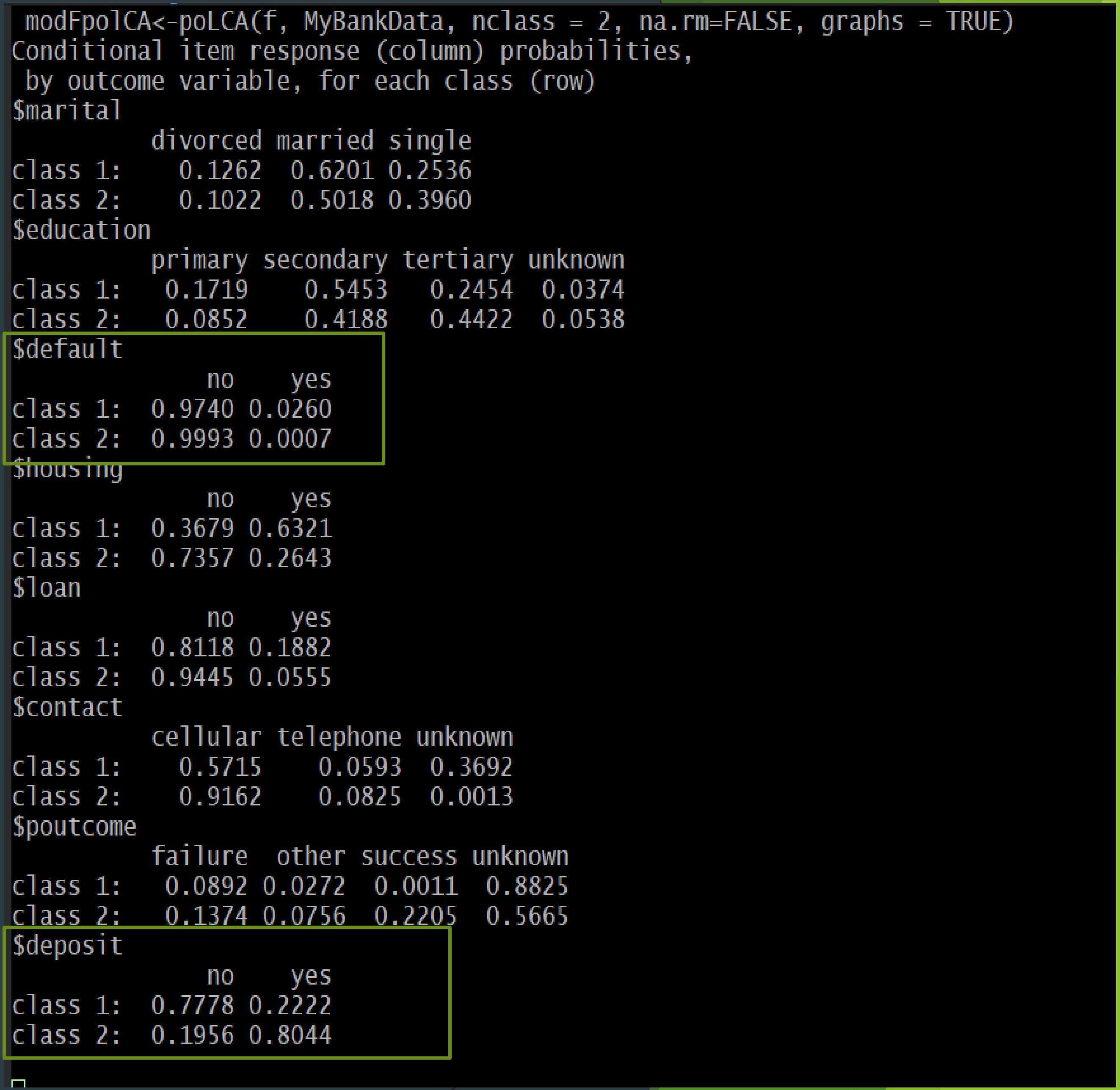
Sample Model Report



```
### define an initial simple model
f <- cbind(marital,education,default,housing,loan,contact,poutcome,deposit)~1

### Execute the model
print("Model F")
modFpolCA<-poLCA(f, MyBankData, nclass = 2, na.rm=FALSE, graphs = TRUE)
```

We get conditional probabilities

$$P(default = yes|class = 1) = 0.0260$$
$$P(deposit = yes|class = 1) = 0.2222$$

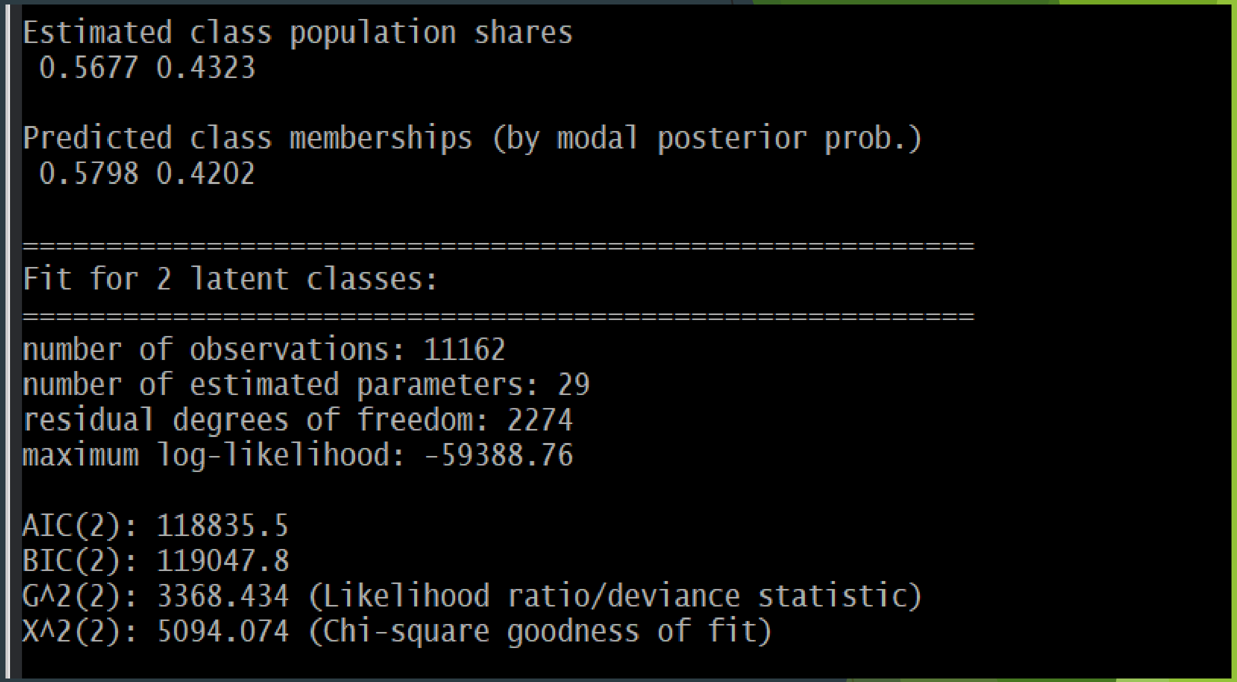


```
 modFpolCA<-poLCA(f, MyBankData, nclass = 2, na.rm=FALSE, graphs = TRUE)
Conditional item response (column) probabilities,
 by outcome variable, for each class (row)
$marital
          divorced married single
class 1:    0.1262  0.6201 0.2536
class 2:    0.1022  0.5018 0.3960
$education
          primary secondary tertiary unknown
class 1:   0.1719    0.5453   0.2454  0.0374
class 2:   0.0852    0.4188   0.4422  0.0538
$default
              no    yes
class 1:  0.9740 0.0260
class 2:  0.9993 0.0007
$housing
              no    yes
class 1:  0.3679 0.6321
class 2:  0.7357 0.2643
$loan
              no    yes
class 1:  0.8118 0.1882
class 2:  0.9445 0.0555
$contact
          cellular telephone unknown
class 1:    0.5715    0.0593  0.3692
class 2:    0.9162    0.0825  0.0013
$poutcome
          failure  other success unknown
class 1:   0.0892 0.0272  0.0011  0.8825
class 2:   0.1374 0.0756  0.2205  0.5665
$deposit
              no    yes
class 1:  0.7778 0.2222
class 2:  0.1956 0.8044
```

https://www.linkedin.com/in/l-p-cruz-phd/
https://unpocologico.wordpress.com/

Detecting a Suitable Model

1. When Estimated Class $\approx$ Predicted Class $\Longrightarrow$ Good Indication
2. We want the AIC, BIC, G^2 and χ^2 to be small
3. The ideal situation - all of #1 and #2 are satisfied
4. If mixed, still useful specially when #2 is met
5. Tuning – try varying the class numbers, number of repetitions, maximum iterations



```
Estimated class population shares
 0.5677 0.4323

Predicted class memberships (by modal posterior prob.)
 0.5798 0.4202

=========================================================
Fit for 2 latent classes:
=========================================================
number of observations: 11162
number of estimated parameters: 29
residual degrees of freedom: 2274
maximum log-likelihood: -59388.76

AIC(2): 118835.5
BIC(2): 119047.8
G^2(2): 3368.434 (Likelihood ratio/deviance statistic)
X^2(2): 5094.074 (Chi-square goodness of fit)
```

χ^2Chi-Square Test for model goodness of fit

H_0 the expected = predicted values

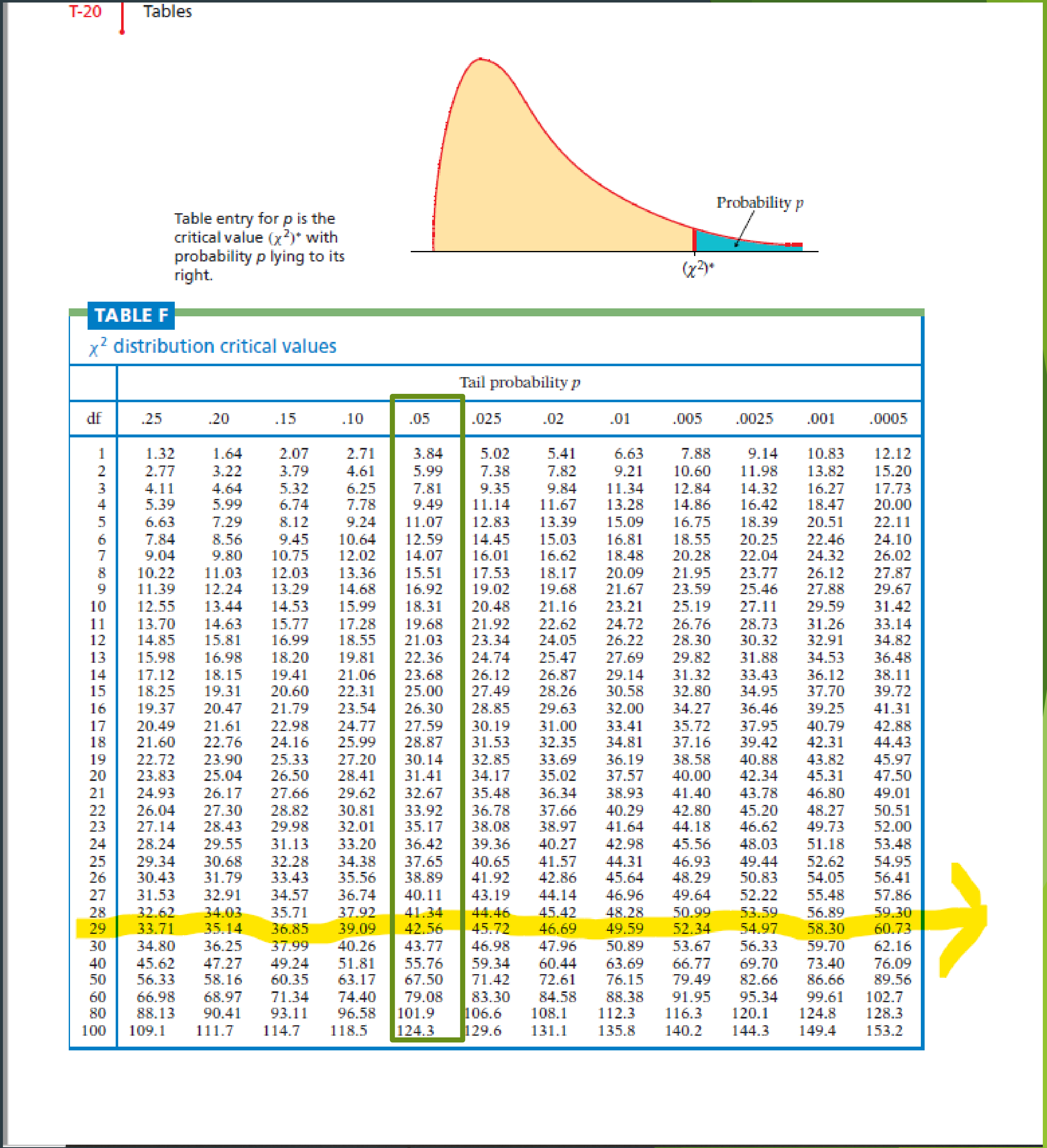
df= 29, χ^2= 5094.074

https://www.linkedin.com/in/l-p-cruz-phd/
https://unpocologico.wordpress.com/

H_0 the expected = predicted values
df= 29, χ^2= 5094.074

At this result we are rejecting H_0 at $\alpha = 0.5$. We are way far to the right

Therefore, this suggests we should try and do better with our model



T-20 Tables

Probability p
$(\chi^2)^*$

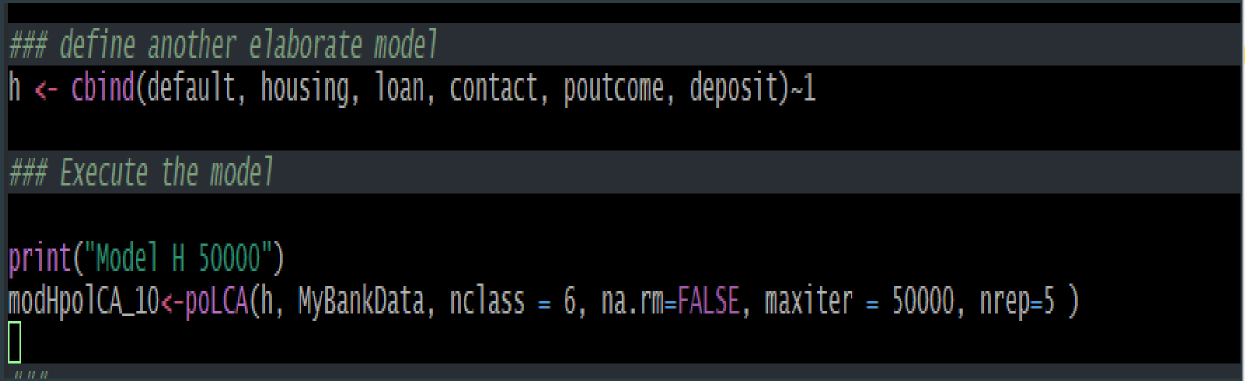
Table entry for p is the critical value $(\chi^2)^*$ with probability p lying to its right.

TABLE F

χ^2 distribution critical values

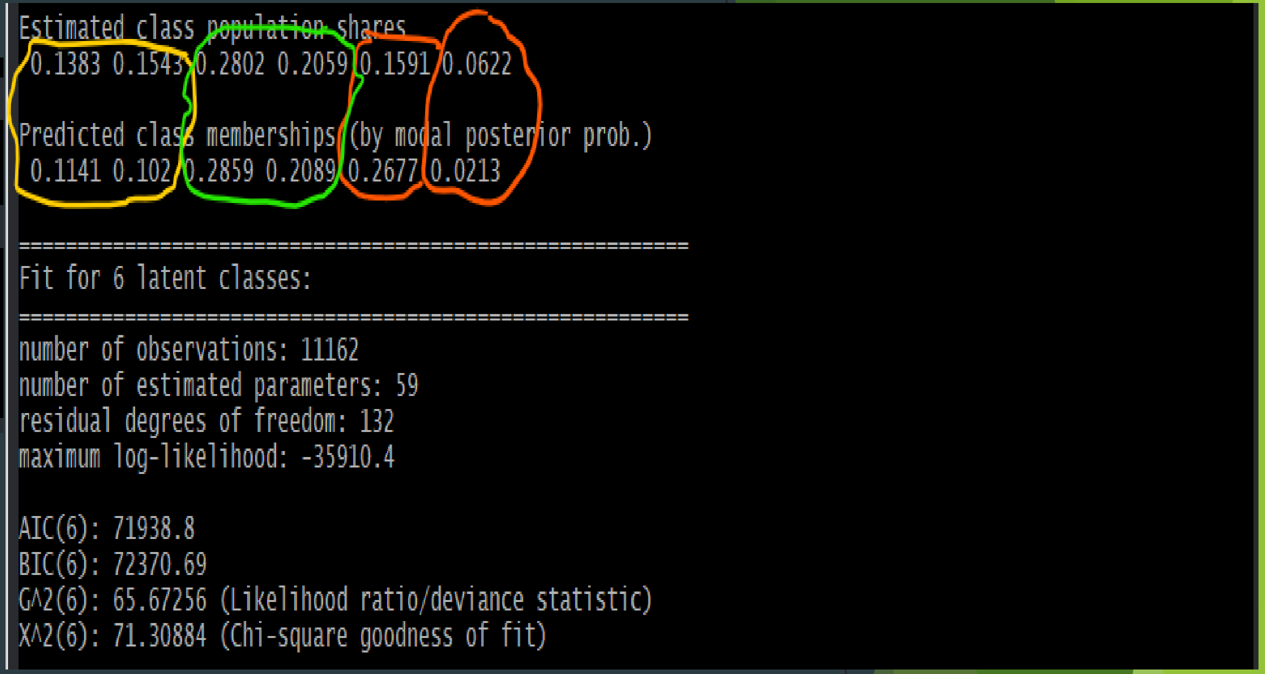
	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91	34.82
13	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48
14	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70	39.72
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25	41.31
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79	42.88
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31	44.43
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82	45.97
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31	47.50
21	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80	49.01
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27	50.51
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73	52.00
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18	53.48
25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62	54.95
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05	56.41
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48	57.86
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89	59.30
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30	60.73
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70	62.16
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40	76.09
50	56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66	89.56
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61	102.7
80	88.13	90.41	93.11	96.58	101.9	106.6	108.1	112.3	116.3	120.1	124.8	128.3
100	109.1	111.7	114.7	118.5	124.3	129.6	131.1	135.8	140.2	144.3	149.4	153.2

Modified Model

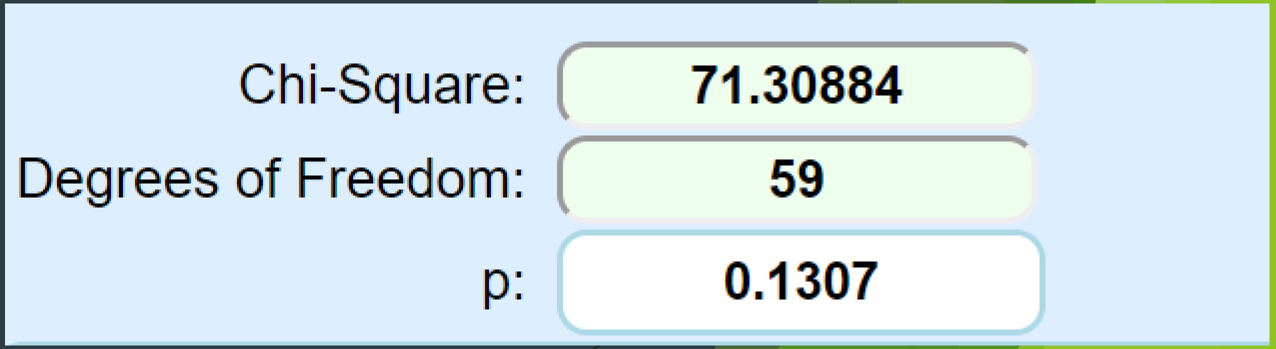


```
### define another elaborate model
h <- cbind(default, housing, loan, contact, poutcome, deposit)~1

### Execute the model

print("Model H 50000")
modHpolCA_10<-poLCA(h, MyBankData, nclass = 6, na.rm=FALSE, maxiter = 50000, nrep=5 )
```

- We have good estimate and predicted w Classes 3 & 4
- We have acceptable numbers also for Classes 1 & 2
- We have bad numbers for Classes 5 & 6
- We got good numbers AIC, BIC, G^2 and χ^2
- As can be seen we have $p = 0.13 > a = 0.05$, so we are accepting the hypothesis that the observed are not significantly different from the expected values - there is a good fit



```
Estimated class population shares
 0.1383 0.1543 0.2802 0.2059 0.1591 0.0622

Predicted class memberships (by modal posterior prob.)
 0.1141 0.102 0.2859 0.2089 0.2677 0.0213

=========================================================
Fit for 6 latent classes:
=========================================================
number of observations: 11162
number of estimated parameters: 59
residual degrees of freedom: 132
maximum log-likelihood: -35910.4

AIC(6): 71938.8
BIC(6): 72370.69
G^2(6): 65.67256 (Likelihood ratio/deviance statistic)
X^2(6): 71.30884 (Chi-square goodness of fit)
```



Chi-Square: 71.30884

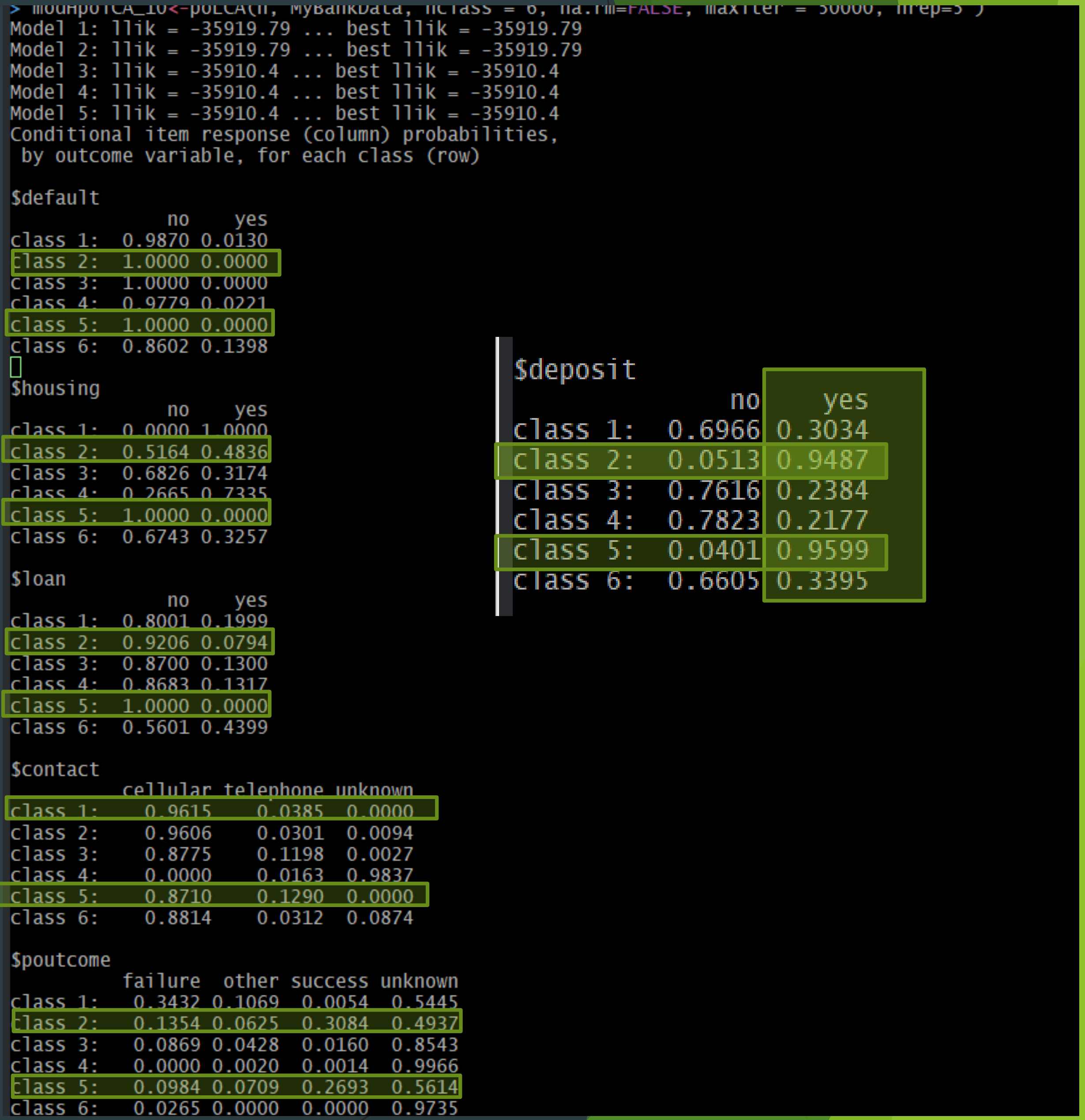
Degrees of Freedom: 59

p: 0.1307

Qualitative Analysis

- Q: What kind of people say YES to the offer?
- A: Class 2 & 5 people are likely to say YES to the offer
- Q: What are their characteristics?
  - Never had credits in default
  - No housing loan
  - No personal loan
  - Contacted by Cell phone
  - Undetermined reaction to previous campaign

Q. If the person said YES before how likely is the person to say YES again



```
> modHpoLCA_10<-poLCA(f, MyBankData, nclass = 6, na.rm=FALSE, maxiter = 30000, nrep=5 )
Model 1: llik = -35919.79 ... best llik = -35919.79
Model 2: llik = -35919.79 ... best llik = -35919.79
Model 3: llik = -35910.4 ... best llik = -35910.4
Model 4: llik = -35910.4 ... best llik = -35910.4
Model 5: llik = -35910.4 ... best llik = -35910.4
Conditional item response (column) probabilities,
 by outcome variable, for each class (row)

$default
               no    yes
class 1:  0.9870 0.0130
class 2:  1.0000 0.0000
class 3:  1.0000 0.0000
class 4:  0.9779 0.0221
class 5:  1.0000 0.0000
class 6:  0.8602 0.1398

$housing
               no    yes
class 1:  0.0000 1.0000
class 2:  0.5164 0.4836
class 3:  0.6826 0.3174
class 4:  0.2665 0.7335
class 5:  1.0000 0.0000
class 6:  0.6743 0.3257

$loan
               no    yes
class 1:  0.8001 0.1999
class 2:  0.9206 0.0794
class 3:  0.8700 0.1300
class 4:  0.8683 0.1317
class 5:  1.0000 0.0000
class 6:  0.5601 0.4399

$contact
          cellular telephone unknown
class 1:    0.9615    0.0385  0.0000
class 2:    0.9606    0.0301  0.0094
class 3:    0.8775    0.1198  0.0027
class 4:    0.0000    0.0163  0.9837
class 5:    0.8710    0.1290  0.0000
class 6:    0.8814    0.0312  0.0874

$poutcome
          failure  other success unknown
class 1:   0.3432 0.1069  0.0054  0.5445
class 2:   0.1354 0.0625  0.3084  0.4937
class 3:   0.0869 0.0428  0.0160  0.8543
class 4:   0.0000 0.0020  0.0014  0.9966
class 5:   0.0984 0.0709  0.2693  0.5614
class 6:   0.0265 0.0000  0.0000  0.9735
```

```
$deposit
              no    yes
class 1:  0.6966 0.3034
class 2:  0.0513 0.9487
class 3:  0.7616 0.2384
class 4:  0.7823 0.2177
class 5:  0.0401 0.9599
class 6:  0.6605 0.3395
```

https://www.linkedin.com/in/l-p-cruz-phd/
https://unpocologico.wordpress.com/

Quantitative Analysis

Q. If the person said YES before, how likely is the person to say YES again

$$P(deposit = YES|poutcome = success) = ?$$

A. We apply Conditional Probs & Bayes' Theorem

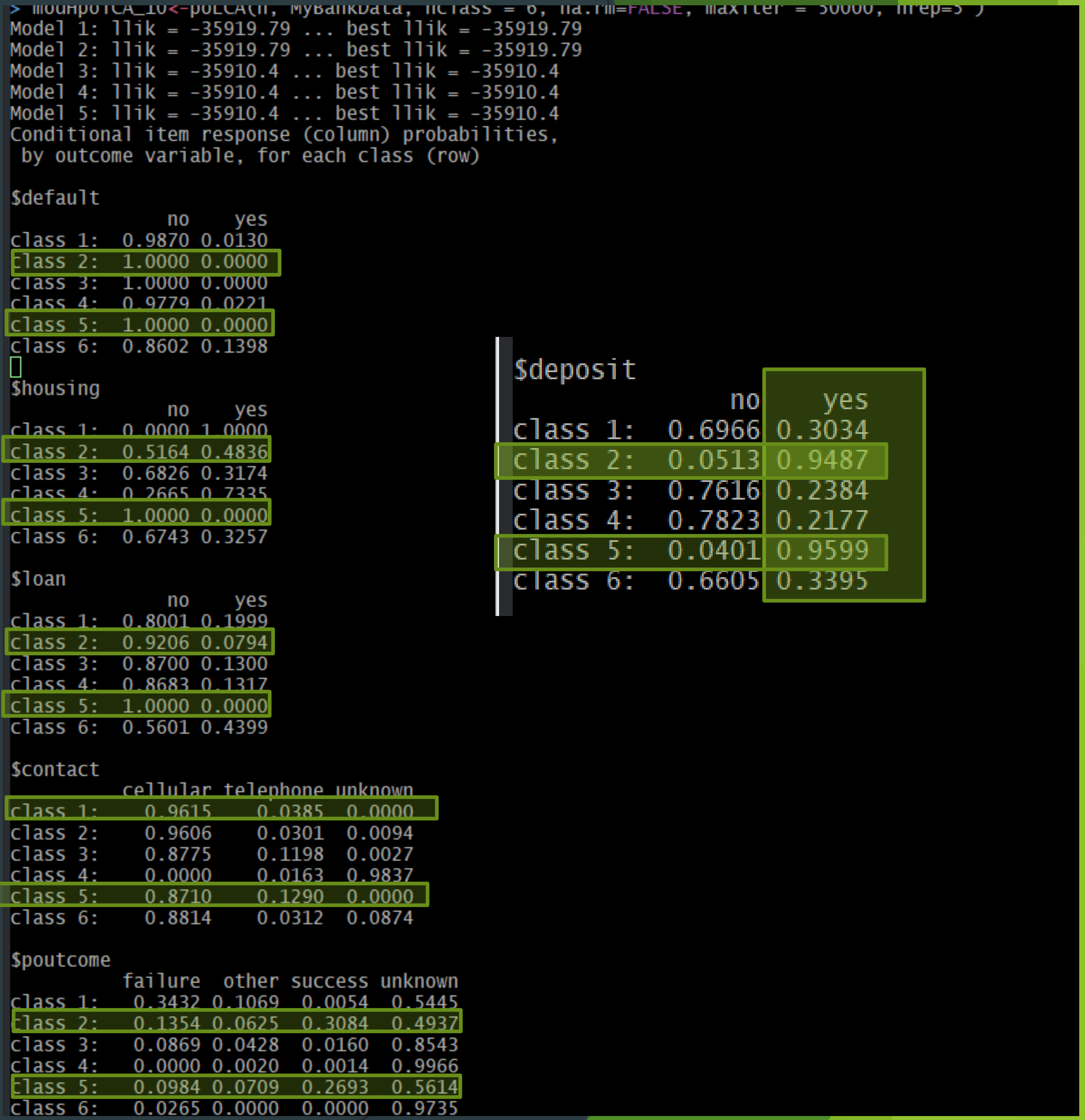
$$P(deposit = YES|poutcome = success) = \frac{P(deposit = YES\ \cap poutcone = success)}{P(poutcome = success)}$$

$$= \frac{P(deposit{=}YES\ \cap poutcone{=}success|class{=}1)}{P(poutcome{=}success)} + \cdots + \frac{P(deposit{=}YES\ \cap poutcone{=}success|class{=}6)}{P(poutcome{=}success)}$$

$$\frac{(.3034) \times (.0054) \times (.1383) + \cdots + (.3395) \times (.0000) \times (.0622)}{(.0054) \times (.1383) + \cdots + (.0000) \times (.0622)}$$

$$= \frac{0.087631}{0.095203} = 0.920459$$

Therefore 92% chance, we should comeback to those who subscribed before and re-market to them.



```
> modHpoLCA_10<-poLCA(f, MyBankData, nclass = 6, na.rm=FALSE, maxiter = 30000, nrep=5 )
Model 1: llik = -35919.79 ... best llik = -35919.79
Model 2: llik = -35919.79 ... best llik = -35919.79
Model 3: llik = -35910.4 ... best llik = -35910.4
Model 4: llik = -35910.4 ... best llik = -35910.4
Model 5: llik = -35910.4 ... best llik = -35910.4
Conditional item response (column) probabilities,
 by outcome variable, for each class (row)

$default
           no    yes
class 1:  0.9870 0.0130
class 2:  1.0000 0.0000
class 3:  1.0000 0.0000
class 4:  0.9779 0.0221
class 5:  1.0000 0.0000
class 6:  0.8602 0.1398

$housing
           no    yes
class 1:  0.0000 1.0000
class 2:  0.5164 0.4836
class 3:  0.6826 0.3174
class 4:  0.2665 0.7335
class 5:  1.0000 0.0000
class 6:  0.6743 0.3257

$loan
           no    yes
class 1:  0.8001 0.1999
class 2:  0.9206 0.0794
class 3:  0.8700 0.1300
class 4:  0.8683 0.1317
class 5:  1.0000 0.0000
class 6:  0.5601 0.4399

$contact
          cellular telephone unknown
class 1:    0.9615    0.0385  0.0000
class 2:    0.9606    0.0301  0.0094
class 3:    0.8775    0.1198  0.0027
class 4:    0.0000    0.0163  0.9837
class 5:    0.8710    0.1290  0.0000
class 6:    0.8814    0.0312  0.0874

$poutcome
          failure  other success unknown
class 1:   0.3432 0.1069  0.0054  0.5445
class 2:   0.1354 0.0625  0.3084  0.4937
class 3:   0.0869 0.0428  0.0160  0.8543
class 4:   0.0000 0.0020  0.0014  0.9966
class 5:   0.0984 0.0709  0.2693  0.5614
class 6:   0.0265 0.0000  0.0000  0.9735
```

```
$deposit
           no    yes
class 1:  0.6966 0.3034
class 2:  0.0513 0.9487
class 3:  0.7616 0.2384
class 4:  0.7823 0.2177
class 5:  0.0401 0.9599
class 6:  0.6605 0.3395
```

https://www.linkedin.com/in/l-p-cruz-phd/
https://unpocologico.wordpress.com/

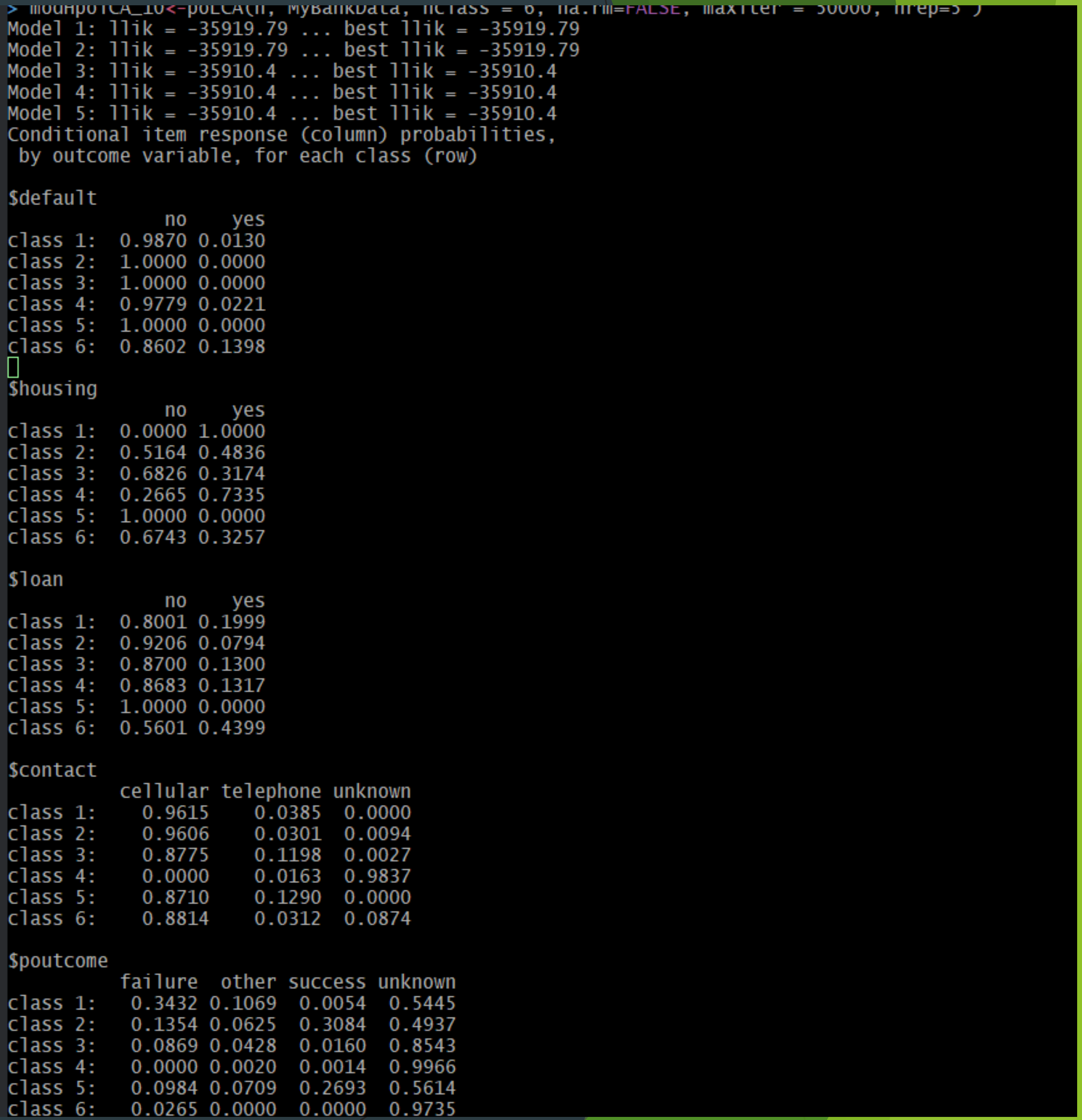
Prediction

- Because of Conditional Probs & Bayes Theorem, we can make predictions if a client will buy or not based on how more or how less of information we have about the client.
- For example, we can ask the question, what is the probability this person whom I know who does not have a housing loan and no personal loan will say YES to the deal?
- Even if these are the only information we know we can still proceed, this is asking
- $P(deposit = YES|housing = YES, loan = YES)$
- Likewise since we have $P(default = YES|class = 1)$, by Bayes, we can have $P(class = 1|default = YES)$
- This is the nice thing about Bayesianism, you can proceed even if you do not have all the information on a customer.

https://www.linkedin.com/in/l-p-cruz-phd/
https://unpocologico.wordpress.com/

Naming the Classes

- The difficult part of Latent Class Modeling, the interpretation of the Classes or how to name them
- Some quick rules of thumb:
    - The latent probabilities vary for a manifest variable
    - Then also, they are closer to 0 or 1
- Examples, Class 1 are people w no defaults but has housing loan and no personal loan uses cell phone with no known previous outcome.
- How should we name this class? Lots of possibilities here, and we can make it descriptive
- Note: This gives us a notional profile of a customer.

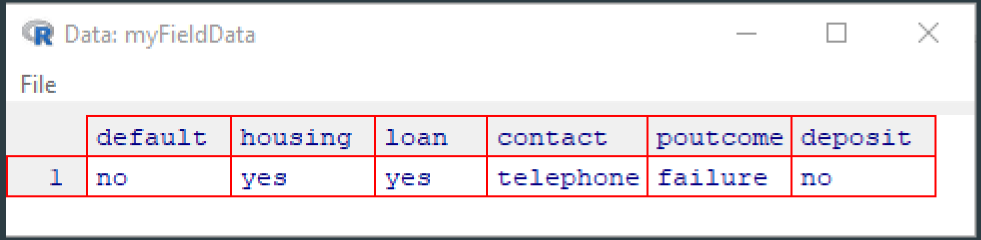


```
> modHpoLCA_10<-poLCA(f, MyBankData, nclass = 6, na.rm=FALSE, maxiter = 50000, nrep=5 )
Model 1: llik = -35919.79 ... best llik = -35919.79
Model 2: llik = -35919.79 ... best llik = -35919.79
Model 3: llik = -35910.4 ... best llik = -35910.4
Model 4: llik = -35910.4 ... best llik = -35910.4
Model 5: llik = -35910.4 ... best llik = -35910.4
Conditional item response (column) probabilities,
 by outcome variable, for each class (row)

$default
             no    yes
class 1:  0.9870 0.0130
class 2:  1.0000 0.0000
class 3:  1.0000 0.0000
class 4:  0.9779 0.0221
class 5:  1.0000 0.0000
class 6:  0.8602 0.1398

$housing
             no    yes
class 1:  0.0000 1.0000
class 2:  0.5164 0.4836
class 3:  0.6826 0.3174
class 4:  0.2665 0.7335
class 5:  1.0000 0.0000
class 6:  0.6743 0.3257

$loan
             no    yes
class 1:  0.8001 0.1999
class 2:  0.9206 0.0794
class 3:  0.8700 0.1300
class 4:  0.8683 0.1317
class 5:  1.0000 0.0000
class 6:  0.5601 0.4399

$contact
          cellular telephone unknown
class 1:    0.9615    0.0385  0.0000
class 2:    0.9606    0.0301  0.0094
class 3:    0.8775    0.1198  0.0027
class 4:    0.0000    0.0163  0.9837
class 5:    0.8710    0.1290  0.0000
class 6:    0.8814    0.0312  0.0874

$poutcome
          failure  other success unknown
class 1:   0.3432 0.1069  0.0054  0.5445
class 2:   0.1354 0.0625  0.3084  0.4937
class 3:   0.0869 0.0428  0.0160  0.8543
class 4:   0.0000 0.0020  0.0014  0.9966
class 5:   0.0984 0.0709  0.2693  0.5614
class 6:   0.0265 0.0000  0.0000  0.9735
```

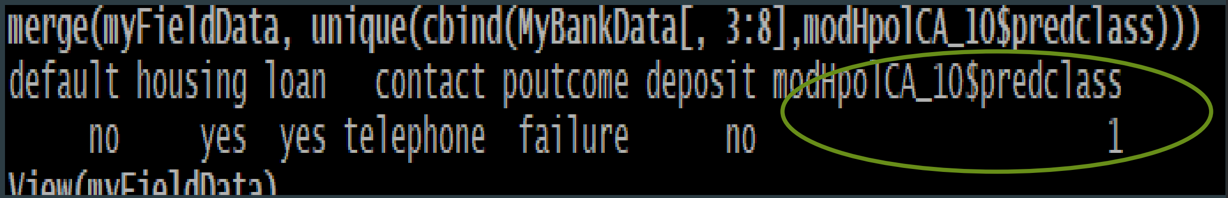
https://www.linkedin.com/in/l-p-cruz-phd/
https://unpocologico.wordpress.com/

Can we predict which Class a data belongs?

- Yes, there are two ways.
- We can use the `predclass` or `posterior` functions of poLCA
- You give it a Y combination and it will produce back a C.
- Assume we have, a person, w default = NO, housing = YES, loan = YES, contact = TELEPHONE, poutcome = FAILURE, desposit = NO

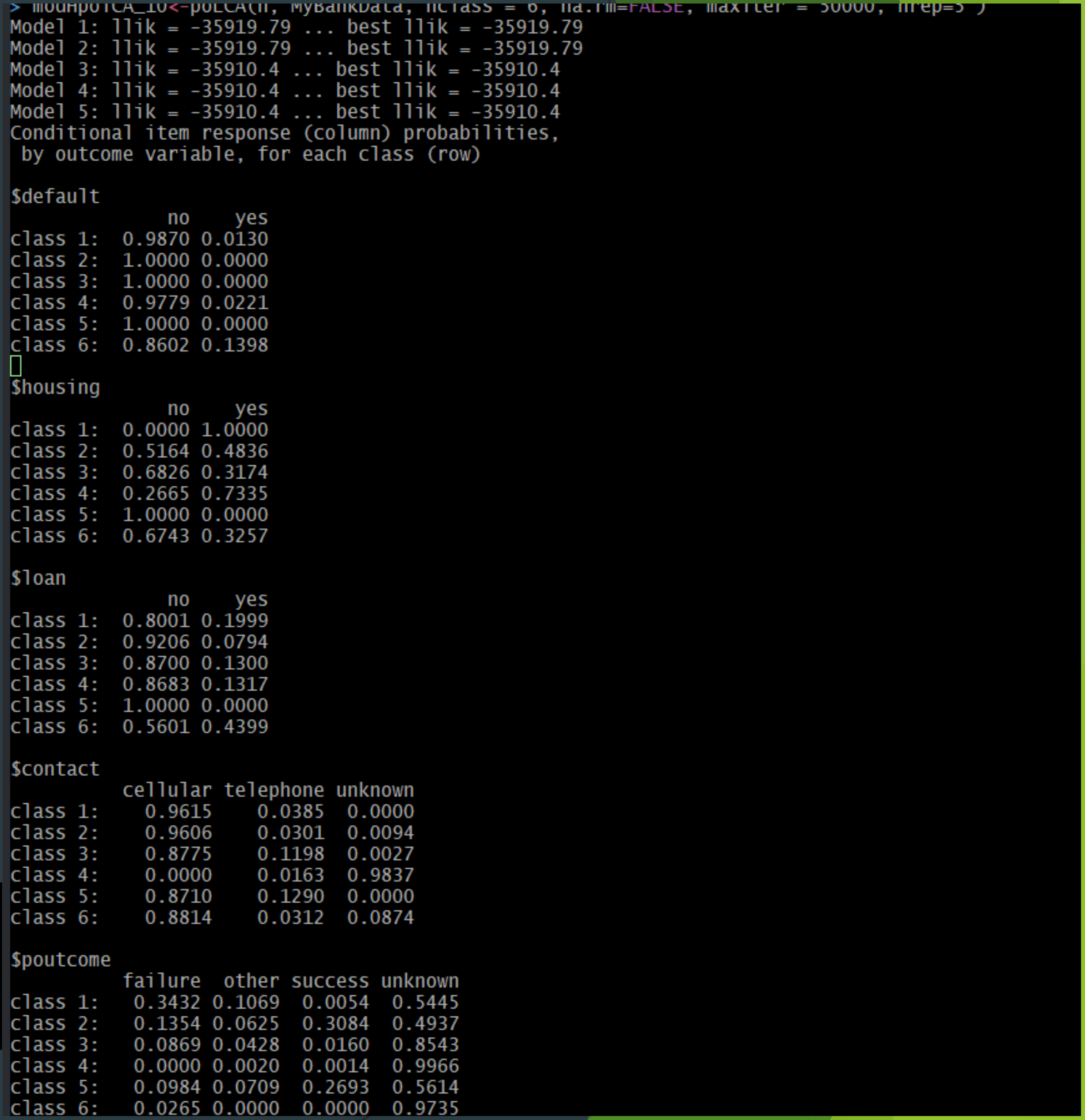


	default	housing	loan	contact	poutcome	deposit
1	no	yes	yes	telephone	failure	no



```
merge(myFieldData, unique(cbind(MyBankData[, 3:8],modHpoLCA_10$predclass)))
 default housing loan   contact poutcome deposit modHpoLCA_10$predclass
      no      yes  yes telephone  failure      no                      1
View(myFieldData)
```

https://www.linkedin.com/in/l-p-cruz-phd/
https://unpocologico.wordpress.com/



```
> modHpoLCA_10<-poLCA(f, MyBankData, nclass = 6, na.rm=FALSE, maxiter = 30000, nrep=5 )
Model 1: llik = -35919.79 ... best llik = -35919.79
Model 2: llik = -35919.79 ... best llik = -35919.79
Model 3: llik = -35910.4 ... best llik = -35910.4
Model 4: llik = -35910.4 ... best llik = -35910.4
Model 5: llik = -35910.4 ... best llik = -35910.4
Conditional item response (column) probabilities,
 by outcome variable, for each class (row)

$default
              no    yes
class 1:  0.9870 0.0130
class 2:  1.0000 0.0000
class 3:  1.0000 0.0000
class 4:  0.9779 0.0221
class 5:  1.0000 0.0000
class 6:  0.8602 0.1398

$housing
              no    yes
class 1:  0.0000 1.0000
class 2:  0.5164 0.4836
class 3:  0.6826 0.3174
class 4:  0.2665 0.7335
class 5:  1.0000 0.0000
class 6:  0.6743 0.3257

$loan
              no    yes
class 1:  0.8001 0.1999
class 2:  0.9206 0.0794
class 3:  0.8700 0.1300
class 4:  0.8683 0.1317
class 5:  1.0000 0.0000
class 6:  0.5601 0.4399

$contact
          cellular telephone unknown
class 1:    0.9615    0.0385  0.0000
class 2:    0.9606    0.0301  0.0094
class 3:    0.8775    0.1198  0.0027
class 4:    0.0000    0.0163  0.9837
class 5:    0.8710    0.1290  0.0000
class 6:    0.8814    0.0312  0.0874

$poutcome
          failure  other success unknown
class 1:   0.3432 0.1069  0.0054  0.5445
class 2:   0.1354 0.0625  0.3084  0.4937
class 3:   0.0869 0.0428  0.0160  0.8543
class 4:   0.0000 0.0020  0.0014  0.9966
class 5:   0.0984 0.0709  0.2693  0.5614
class 6:   0.0265 0.0000  0.0000  0.9735
```

Further On

- depmixS4 – you might try this out if you have a combination of continuous and discrete manifest variables to study
- Accommodates mixed multivariate data
- However, not as extensive report but is offered for extension

https://www.linkedin.com/in/l-p-cruz-phd/
https://unpocologico.wordpress.com/

Conclusion

- We talked about mixture models
- We covered the varied ideas on latent classes
- We demonstrated using one technique – poLCA
- We used this to attack a marketing dataset
- We defined how to spot a fitting model
- How to predict using poLCA
- How to cluster using poLCA
- The way ahead

https://www.linkedin.com/in/l-p-cruz-phd/
https://unpocologico.wordpress.com/

Thank you for coming & having me
Questions or Comments - Welcome

https://www.linkedin.com/in/l-p-cruz-phd/

https://unpocologico.wordpress.com/