Statistical Regression for Bank Marketing

Michael Zetune, Komal Malik, Jeremy Greene, Hannah Smilansky

Overview

The Data

Context

Analysis

Questions and Predictions

Methods

Data Cleaning

Assumptions

Regression

Results

Findings

Conclusion

The Data

Context

- Predicted whether or not the client created an account based on a direct marketing campaign done by a Portuguese banking institution.
- Assigned probability of success to the campaign by using predictor and quantitative variables.
- Created dummy variables for the campaign's outcome (1 for made a term deposit, 0 for failure to secure a term deposit from the client) and basing the multivariate regression on that.

Analysis

The dataset we are using lists a randomly selected 4000 bank customers from a larger dataset. Attributes include age, education, marital status, and more. The output variable is whether or not a given client subscribed to a term deposit.

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 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.
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 - **make.account** - has the client subscribed a term deposit? (binary: 'yes','no')

Question and Prediction

- Which predictor variable has the highest practical significance in predicting whether or not an account is made?

- We predict that the number of times an individual was previously contacted will play a significant role in whether or not they open an account.

Methods

Data Cleaning

01

Remove marketing$duration column:

```
marketing$duration <- NULL
```

This is because duration is extremely correlated with the output target, but duration isn't actually known until a call is performed.

Data Cleaning

02

Modify coded marketing$pdays column:

```
marketing$pdays[marketing$pdays == 999] <- NA
```

The dataset repository tells us 999 means the client wasn't previously contacted, so they coded 999. After this step, most clients were never contacted, so we remove the column:

```
marketing$pdays <- NULL
```

Data Cleaning
03

Lastly, we change the output column from “yes” or “no” to 1 or 0 to assist with logistic regression:

```
marketing$made.account[marketing$make.account == 'yes'] <- 1

marketing$made.account[marketing$make.account == 'no'] <- 0

marketing$make.account <- NULL
```

Data Cleaning

04

Using vif and alias functions we found that the `loan` variable is “unknown” if and only if the `housing` variable is “unknown”. Therefore we needed to remove rows with `loan` = “unknown”.

```
marketing <- subset(marketing, marketing$housing != 'unknown')

model1 <- glm(made.account ~ ., data=marketing, family='binomial')

alias(model1)
```

Assumptions

- Binary output
- No multicollinearity
- Independence of observations
- Large dataset

Regression

01

We started with a simple logistic model that considers all variables:

```
model1 <- glm(made.account ~ ., data=marketing, family='binomial')
```

Regression
02

Next, we performed backwards, forwards, and both step regression:

```
null <- glm(made.account ~ 1, data=marketing, family='binomial')
full <- glm(made.account ~ ., data=marketing, family='binomial')

backward.model <- step(full, scope=list(lower=null, upper=full),
direction='backward')

forward.model <- step(null, scope = list(lower=null, upper=full), direction =
'forward')

both.model <- step(null, scope=list(lower=null, upper=full),
direction='both')
```

Regression

03

Lastly, we tried using regsubsets.output to find the optimal set of variables:

```
subset.model <- glm(made.account ~ age + job + month + campaign + previous +
poutcome + emp.var.rate + euribor3m + nr.employed, data=marketing,
family='binomial')
```

Results

Findings

01

We have four models to consider (forward.model, backward.model, both.model, and subset.model). Our team compared the AICs to test the relative quality of each statistical model.

```
AIC(forward.model)     # 2230.536
AIC(backward.model)    # 2224.604
AIC(both.model)        # 2230.536
AIC(subset.model)      # 2256.057
```

Findings

02

After that, we tested for goodness-of-fit by finding the pseudo R^2 of each model.

```
# forward.model pseudo R^2
1-(2196.5/2783.7)

# backward.model pseudo R^2
1-(2188.6/2783.7)

# both.model pseudo R^2
1-(2196.5/2783.7)

# subset.model pseudo R^2
1-(2198.1/2783.7)
```

```
# forward.model.rsq = .2109

# backward.model.rsq = .2138

# both.model.rsq = .2109

# subset.model.rsq = .2104
```

Findings

03

Finally, we matched predictive accuracy to each model. (1/2)

```
# Naive model
sum(marketing$made.account == 0)/nrow(marketing)          # 0.8899

# Forward model
predicted.frwd <- (predict(forward.model, type = 'response') >= 0.5)
actual.frwd <- (marketing$made.account == 1)
sum(predicted.frwd == actual.frwd) / nrow(marketing)    # 0.9041
```

Findings

03 continued

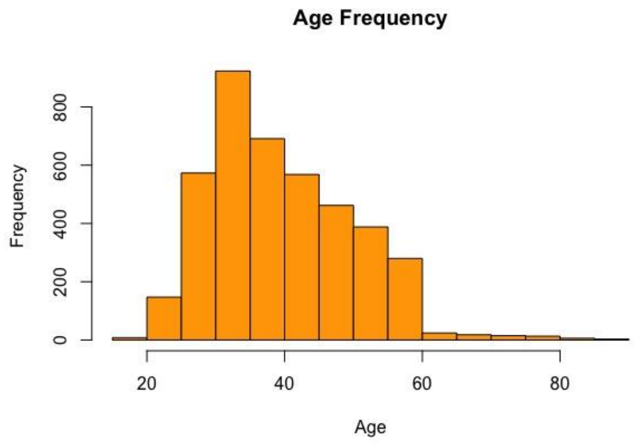
Finally, we matched predictive accuracy to each model. (2/2)

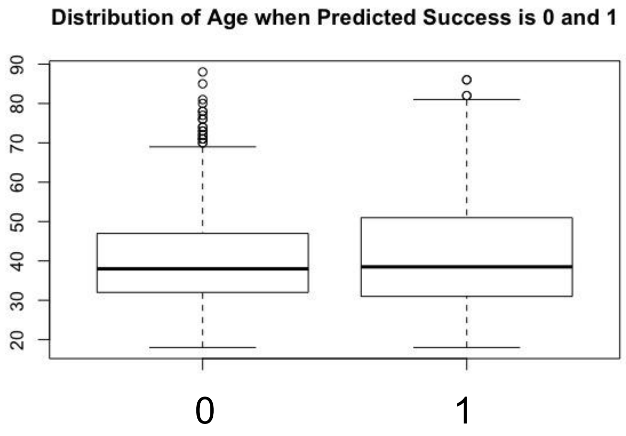
```
# Backward model
predicted.bwrd <- (predict(backward.model, type = 'response') >= 0.5)
actual.bwrd <- (marketing$made.account == 1)
sum(predicted.bwrd == actual.bwrd) / nrow(marketing)     # 0.9033

# Subset model
predicted.sub <- (predict(subset.model, type = 'response') >= 0.5)
actual.sub <- (marketing$made.account == 1)
sum(predicted.sub == actual.sub) / nrow(marketing)       # 0.9021
```

Graphs

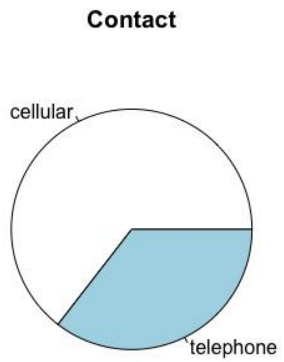
01

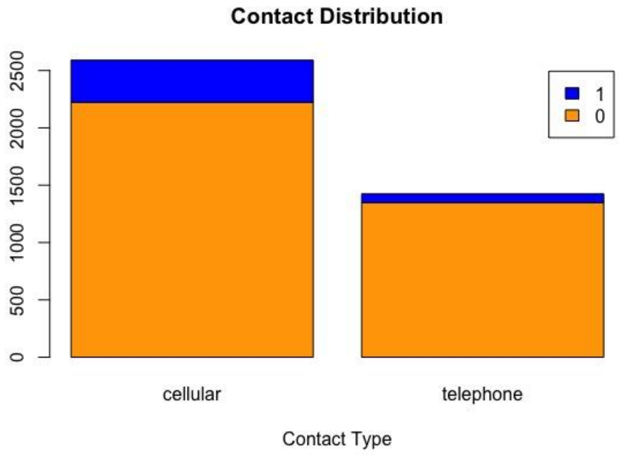

Age Frequency
Frequency
0
200
400
600
800
20
40
60
80
Age


Distribution of Age when Predicted Success is 0 and 1
20
30
40
50
60
70
80
90
0
1

Graphs

02


Contact
cellular
telephone


Contact Distribution
0
500
1000
1500
2000
2500
1
0
cellular
telephone
Contact Type

Testing the Model

- Let’s consider the case of Jorge (who is quite stubborn)

- Inputs:
  - Age: 30
  - Contact: Cellular
  - Month: August
  - Campaigns: 2
  - Previous Outcome: Nonexistent
  - Employee Variation Rate: 1.4
  - Consumer Price Index: 93.444
  - Consumer Confidence Index: -36.1

- Actual response: 0

```
> predict(backward.model, data.frame(age = 30,
contact = ‘cellular’, month = ‘aug’, campaign = 2,
poutcome=’nonexistent’, empr.var.rate=1.4,
cons.price.idx=93.444, cons.conf.idx=-36.1),
type=’response’)

0.04795391
```

Statistics from the Model

01

- **Eight variables** are statistically significant from the Backward Model
- Coefficients represent the predicted increase in log odds of predicted success assuming all other variables are held constant
- Example: if Consumer Price Index increases by 1, and all other variables are held constant, the log odds of predicted success will increase by 1.28

Variable	P-Value	Coefficient
campaign	0.024403	-7.741e-02
emp.var.rate	< 2e-16	-7.321e-01
cons.price.idx	< 2e-16	1.281e+00
cons.conf.idx	0.000866	5.103e-02
monthmar	4.32e-06	1.748e+00
contacttelephone	1.15e-05	-9.368e-05
poutcomenonexistent	0.011604	4.471e-01
poutcomesuccess	2.75e-12	1.742e+00

Statistics from the Model

02

- A confidence interval of the statistically significant variables is shown to the right

- Confidence interval shows we are 95% confident the coefficient will fall between the lower and upper ends

Variable	2.5%	97.5%
campaign	-1.488e-01	-0.014
emp.var.rate	-8.553e-01	-0.609
cons.price.idx	9.773e-01	1.585
cons.conf.idx	2.111e-02	0.081
monthmar	1.005e+00	2.502
contacttelephone	-1.366e+00	-0.528
poutcomenonexistent	1.058e-01	0.801
poutcomesuccess	1.258e+00	2.237

Statistics from the Model

03

A McFadden's pseudo-R^2 between 0.2-0.4 is optimal.

Since our calculated R^2 is **0.2138**, the model is said to have a **very good fit**.

Conclusion

01

Since the pseudo R^2s and predictive accuracy between models are about the same, we use AIC to judge.

With the highest R^2 and the lowest AIC, the backward.model clearly is the best at modeling our bank marketing information.

```
# Backward model
```

AIC	2224.604
McFadden's Pseudo-R^2	0.2138
Predictive Accuracy	0.9033

Answering the Question

The backward model most accurately predicts the success of the Portuguese Bank's marketing campaigns. Among the four of the eight statistically significant variables, we checked through the variables to determine which one has the highest practical significance.

Variable	**2.5%**	**97.5%**
campaign	-1.488e-01	-0.014
emp.var.rate	-8.553e-01	-0.609
cons.price.idx	9.773e-01	1.585
cons.conf.idx	2.111e-02	0.081
monthmar	1.005e+00	2.502
contacttelephone	-1.366e+00	-0.528
poutcomenonexistent	1.058e-01	0.801
poutcomesuccess	**1.258e+00**	**2.237**

Answering the Question

poutcomesucess is a categorical variable with a coefficient that is high relative to other dummy variables at 1.258, and it is an easily discovered piece of information as we market to customers, so we can consider that our most practically significant predictor.

Variable	**2.5%**	**97.5%**
campaign	-1.488e-01	-0.014
emp.var.rate	-8.553e-01	-0.609
cons.price.idx	9.773e-01	1.585
cons.conf.idx	2.111e-02	0.081
monthmar	1.005e+00	2.502
contacttelephone	-1.366e+00	-0.528
poutcomenonexistent	1.058e-01	0.801
poutcomesuccess	**1.258e+00**	**2.237**

Conclusion
02

Limitations:
- Lack of important predictor variables (e.g. income)
- Misinterpretation of research findings (e.g. poor data cleaning)

Future Extensions:
- Focus on certain successful predictor demographics before investing money in advertisements

Questions?