AUTHORS: KRYSTAL LY

Bank Telemarketing Data Set

Final Presentation

Presentation Outline

01 Hook

02 Problem Statement

03 Our Data

04 Data Wrangling

05 Modeling Process

06 Interpretation

07 Validation

08 Prediction

09 Future Thoughts

Hook

Typically feel annoyed by Telemarketing?

US TOO!

Telemarketing - a controversial approach

- Easy to reach out to customers
- Cheaper
- Damaging the company's image
- Startup costs are very expensive

What factors affect the outcome of Telemarketing campaigns?

- The outcome of telemarketing campaigns for a Portuguese retail bank
- Make predictions based on our model.

Problem Statement

What is likely to be the outcome of the telemarketing campaigns based on the characteristics of the clients and the calls?

Our Goal

Predict if the client will subscribe (yes/no) to a term deposit (variable y)

Our Data

Dataset	Bank Tele-Marketing Data Set
Data Source	**UCI Machine Learning Repository** [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

Data Background	• Collected from several telemarketing campaigns in which the Portuguese bank attempted to target customers through phone calls to sell long-term deposits
	• Includes both
	  ◦ the phone calls of which the bank executed
	  ◦ the phone calls of which clients contacted the help center.
	• Each observation includes
	  ◦ the outcome,
	  ◦ whether or not the target customers subscribed to the term deposit, and
	  ◦ the characteristics of the customers and the phone calls themselves.

Our Data

Group	Variable Names	Description of Variables
Group 1	• age • job • marital • education • default • housing • loan • contact	Characteristics of the client
Group 2	• month • day_of_week • campaign • pdays • previous • poutcome	Characteristics of the calls made to clients
Group 3	• emp.var.rate • cons.price.idx • cons.conf.idx • euribor3m • nr.employed • y	Characteristics of the economy

A quick look at our dataset

age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays
56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	261	1	999
57	services	married	high.school	unknown	no	no	telephone	may	mon	149	1	999
37	services	married	high.school	no	yes	no	telephone	may	mon	226	1	999
40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1	999
56	services	married	high.school	no	no	yes	telephone	may	mon	307	1	999
45	services	married	basic.9y	unknown	no	no	telephone	may	mon	198	1	999
59	admin.	married	professional.course	no	no	no	telephone	may	mon	139	1	999
41	blue-collar	married	unknown	unknown	no	no	telephone	may	mon	217	1	999
24	technician	single	professional.course	no	yes	no	telephone	may	mon	380	1	999
25	services	single	high.school	no	yes	no	telephone	may	mon	50	1	999
41	blue-collar	married	unknown	unknown	no	no	telephone	may	mon	55	1	999
25	services	single	high.school	no	yes	no	telephone	may	mon	222	1	999
29	blue-collar	single	high.school	no	no	yes	telephone	may	mon	137	1	999
57	housemaid	divorced	basic.4y	no	yes	no	telephone	may	mon	293	1	999
35	blue-collar	married	basic.6y	no	yes	no	telephone	may	mon	146	1	999
54	retired	married	basic.9y	unknown	yes	yes	telephone	may	mon	174	1	999
35	blue-collar	married	basic.6y	no	yes	no	telephone	may	mon	312	1	999
46	blue-collar	married	basic.6y	unknown	yes	yes	telephone	may	mon	440	1	999
50	blue-collar	married	basic.9y	no	yes	yes	telephone	may	mon	353	1	999
39	management	single	basic.9y	unknown	no	no	telephone	may	mon	195	1	999
30	unemployed	married	high.school	no	no	no	telephone	may	mon	38	1	999
55	blue-collar	married	basic.4y	unknown	yes	no	telephone	may	mon	262	1	999

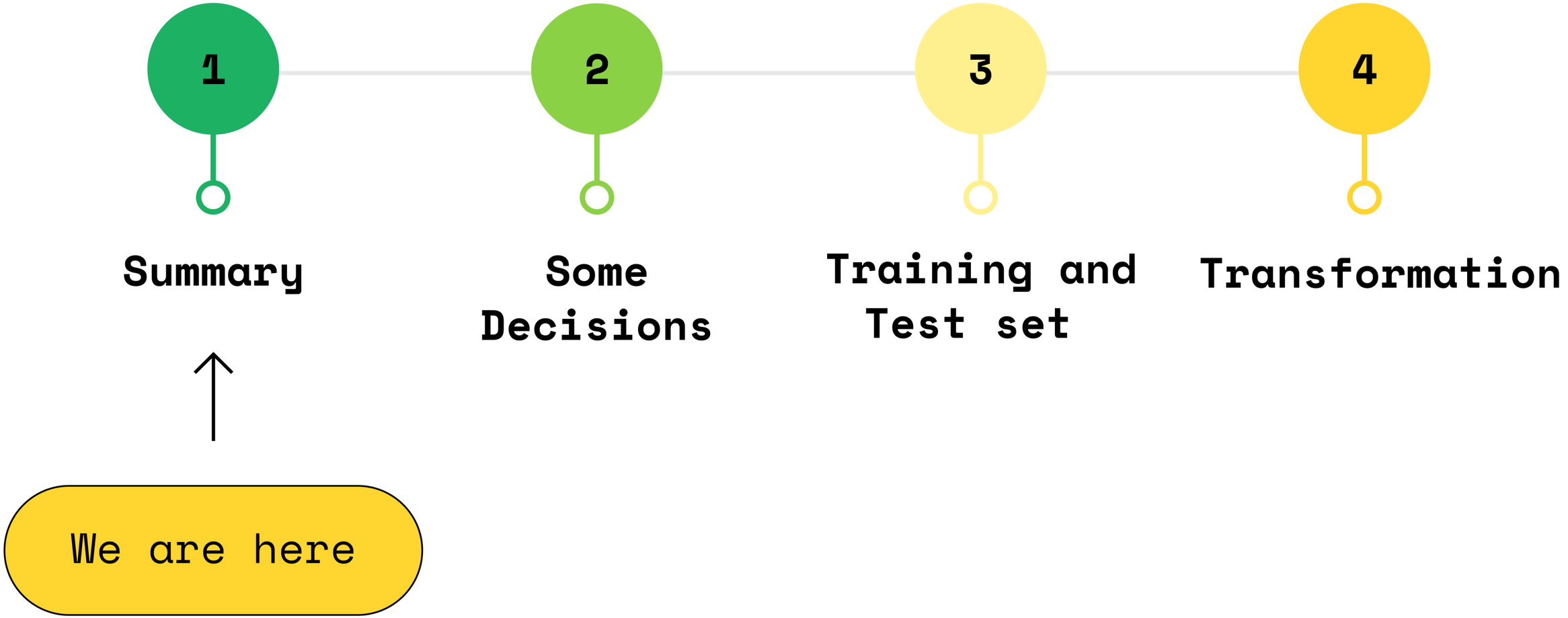
Data Wrangling

01 Turn all "unknown" string to NA values

02 Turn Bernoulli variables into 0 and 1 categories

03 Turn other variables into its suitable types of variables

Data Exploration



Summary

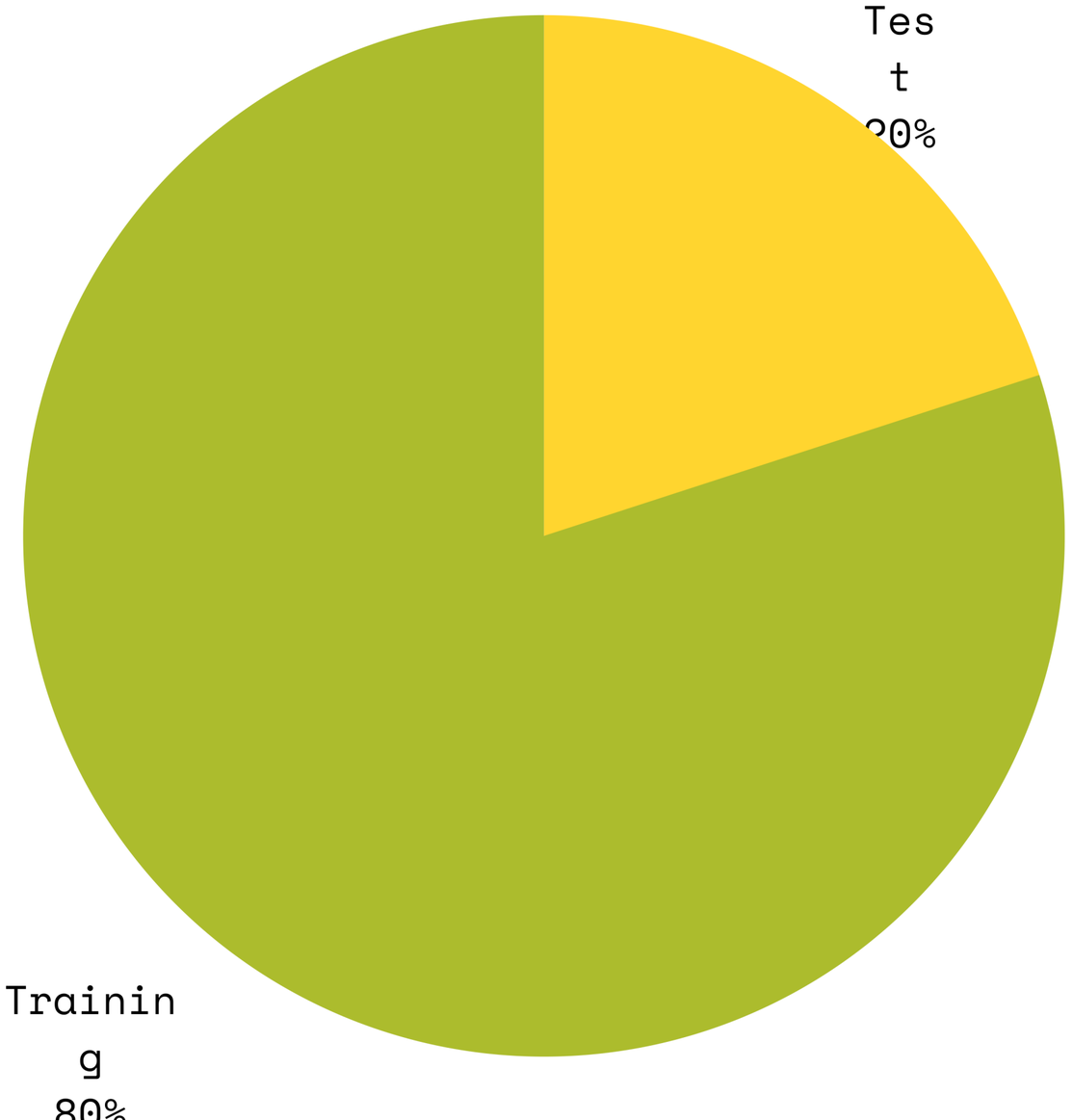
```
      age                  job           marital                    education     default      housing        loan
 Min.   :17.00   admin.     :10422   divorced: 4612   university.degree  :12168   0   :32588   0   :18622   0   :33950
 1st Qu.:32.00   blue-collar: 9254   married :24928   high.school        : 9515   1   :    3   1   :21576   1   : 6248
 Median :38.00   technician : 6743   single  :11568   basic.9y           : 6045   NA's: 8597   NA's:  990   NA's:  990
 Mean   :40.02   services   : 3969   NA's    :   80   professional.course: 5243
 3rd Qu.:47.00   management : 2924                     basic.4y           : 4176
 Max.   :98.00   (Other)    : 7546                     (Other)            : 2310
                 NA's       :  330                     NA's               : 1731
  telephone     month        day_of_week     duration          campaign          pdays          previous
 0:26144   may    :13769   fri:7827    Min.   :   0.0   Min.   : 1.000   999    :39673   Min.   :0.000
 1:15044   jul    : 7174   mon:8514    1st Qu.: 102.0   1st Qu.: 1.000   3      :  439   1st Qu.:0.000
           aug    : 6178   thu:8623    Median : 180.0   Median : 2.000   6      :  412   Median :0.000
           jun    : 5318   tue:8090    Mean   : 258.3   Mean   : 2.568   4      :  118   Mean   :0.173
           nov    : 4101   wed:8134    3rd Qu.: 319.0   3rd Qu.: 3.000   9      :   64   3rd Qu.:0.000
           apr    : 2632               Max.   :4918.0   Max.   :56.000   2      :   61   Max.   :7.000
           (Other): 2016                                                 (Other):  421
     poutcome      emp.var.rate      cons.price.idx  cons.conf.idx     euribor3m      nr.employed      deposit
 failure    : 4252   Min.   :-3.40000   Min.   :92.20   Min.   :-50.8   Min.   :0.634   Min.   :4964   Min.   :0.0000
 nonexistent:35563   1st Qu.:-1.80000   1st Qu.:93.08   1st Qu.:-42.7   1st Qu.:1.344   1st Qu.:5099   1st Qu.:0.0000
 success    : 1373   Median : 1.10000   Median :93.75   Median :-41.8   Median :4.857   Median :5191   Median :0.0000
                     Mean   : 0.08189   Mean   :93.58   Mean   :-40.5   Mean   :3.621   Mean   :5167   Mean   :0.1127
                     3rd Qu.: 1.40000   3rd Qu.:93.99   3rd Qu.:-36.4   3rd Qu.:4.961   3rd Qu.:5228   3rd Qu.:0.0000
                     Max.   : 1.40000   Max.   :94.77   Max.   :-26.9   Max.   :5.045   Max.   :5228   Max.   :1.0000
```

Make Some Decisions

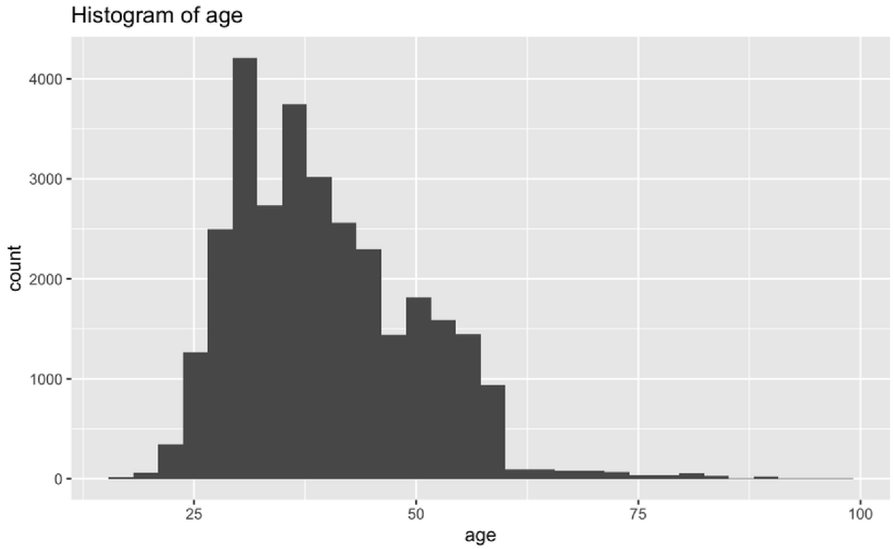
We decide to remove those columns below:

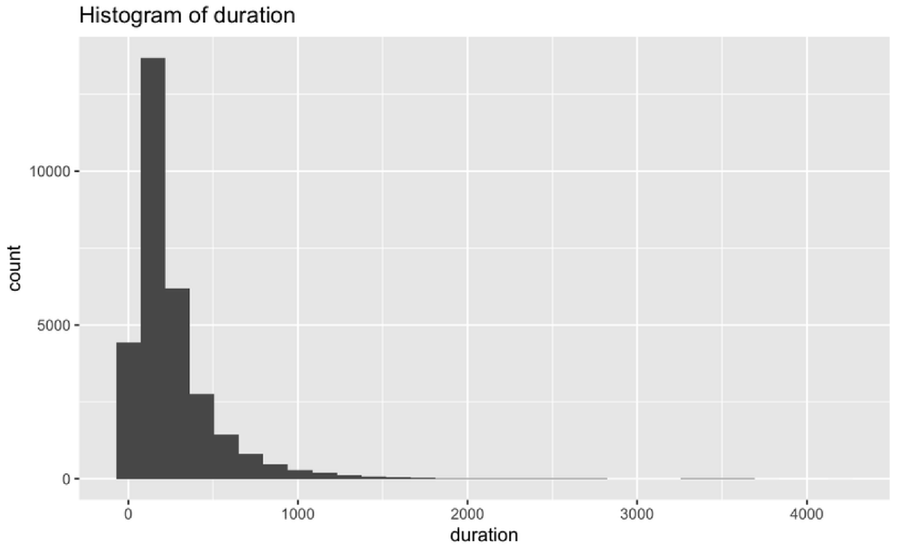
- `default` because it has too many NAs
- `poutcome` (previous campaign's outcome) because it has too many "nonexistent" result
- `pdays` (days after the customers were last contacted in the previous campaigns) because most of the customers had never been contacted before

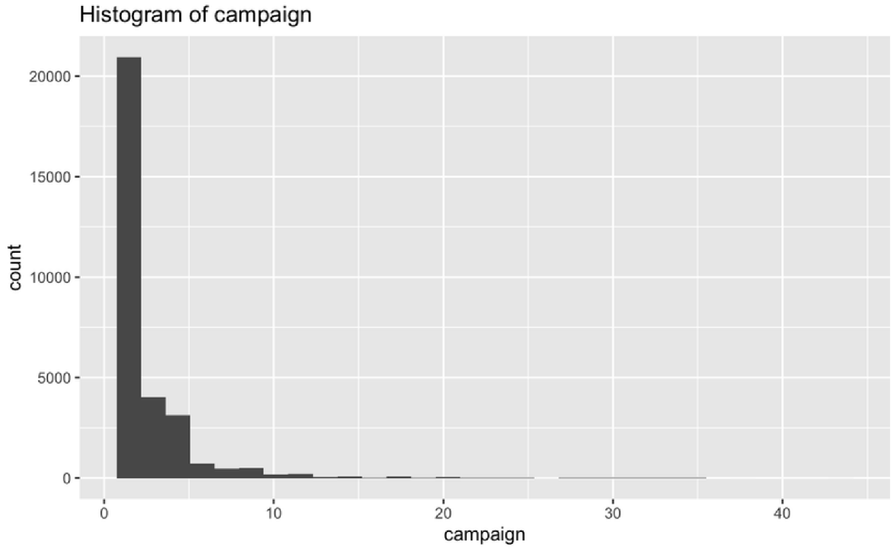
Training and Test Set

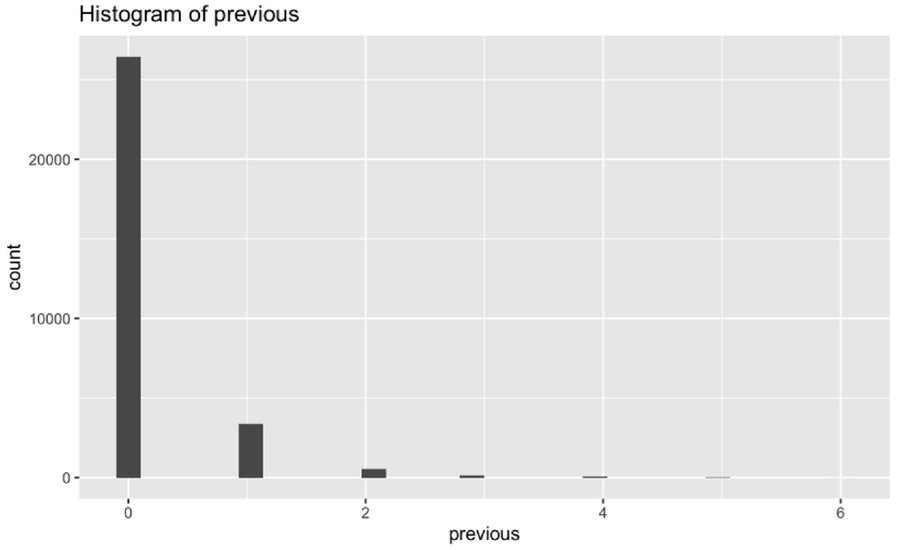


Histogram

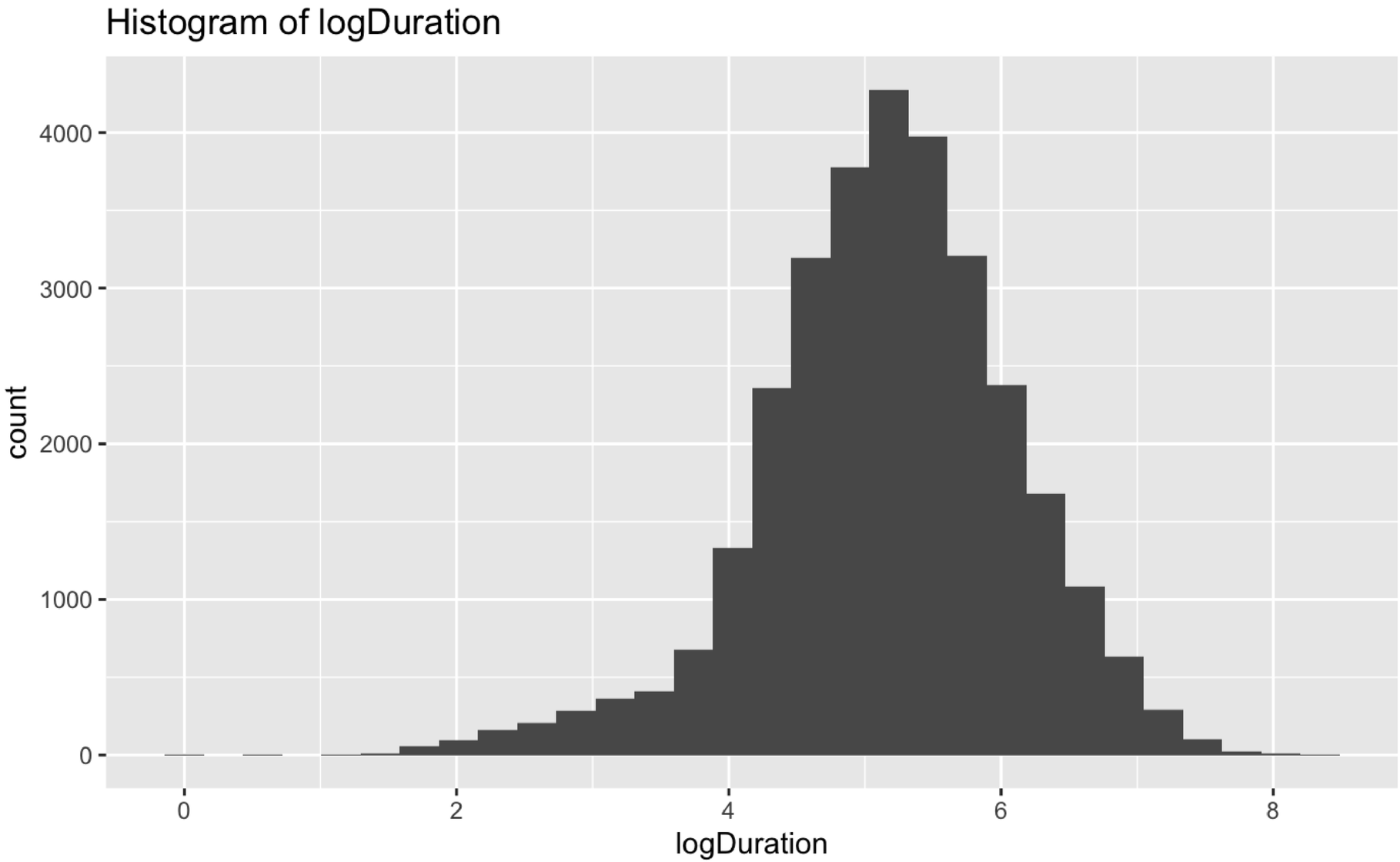

Histogram of age
count
4000
3000
2000
1000
0
25
50
75
100
age


Histogram of duration
count
10000
5000
0
0
1000
2000
3000
4000
duration


Histogram of campaign
count
20000
15000
10000
5000
0
0
10
20
30
40
campaign

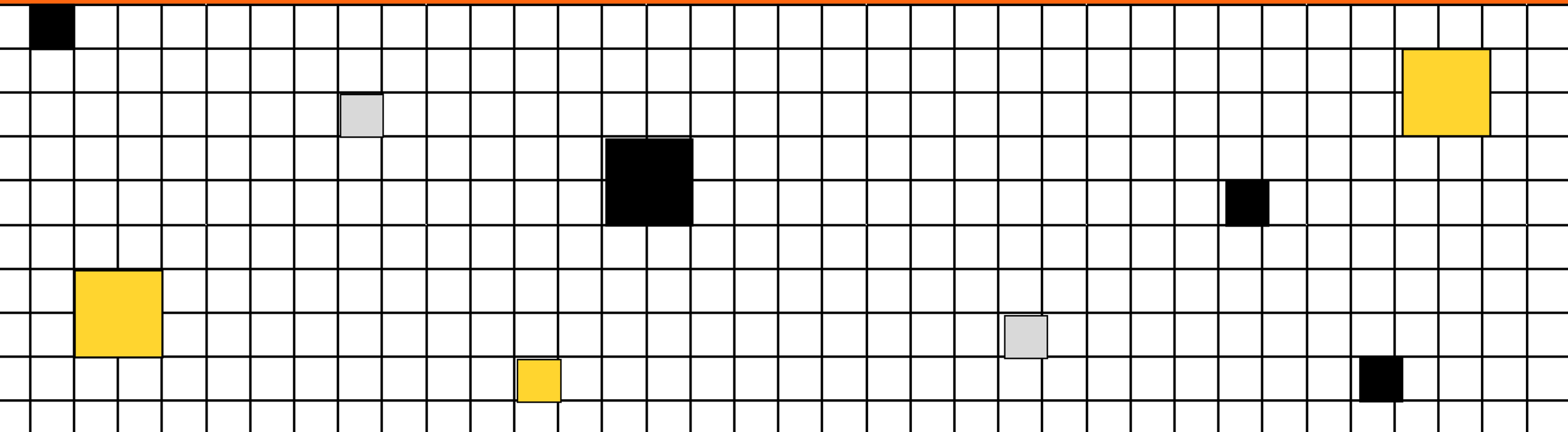

Histogram of previous
count
20000
10000
0
0
2
4
6
previous

Log Transformation



Modeling Process

Logistic Regression



Our process

All variables

Residual Deviance: 12182

Too many insignificant p-values

Remove age, job, marital, education, housing, loan, day_of_week, campaign nr.employed

Model 2

Residual Deviance: 12267

Dxy: 0.866

Some p-values of month categories not significant

Our process

Model 2

Residual Deviance: 12267

Dxy: 0.866

Some p-values of month categories not significant

Remove month

Model 3

Residual Deviance: 12803

Dxy: 0.843

P-value of euribor3m not significant (0.5496)

Our process

Model 3

Residual Deviance: 12803

Dxy: 0.843

P-value of euribor3m not significant

Remove euribor3m

Model 4

Residual Deviance: 12803

Dxy: 0.843

All p-values are significant

Drop-in deviance test

The drop-in deviance tests yield a p-value of 0 for all models, indicating that it's hard to have a larger or equal drop-in deviance

=> All 4 models are significantly better than the null model, explaining a larger amount of variation of `deposit`.

```
[1] 0
[1] 0
[1] 0
[1] 0
```

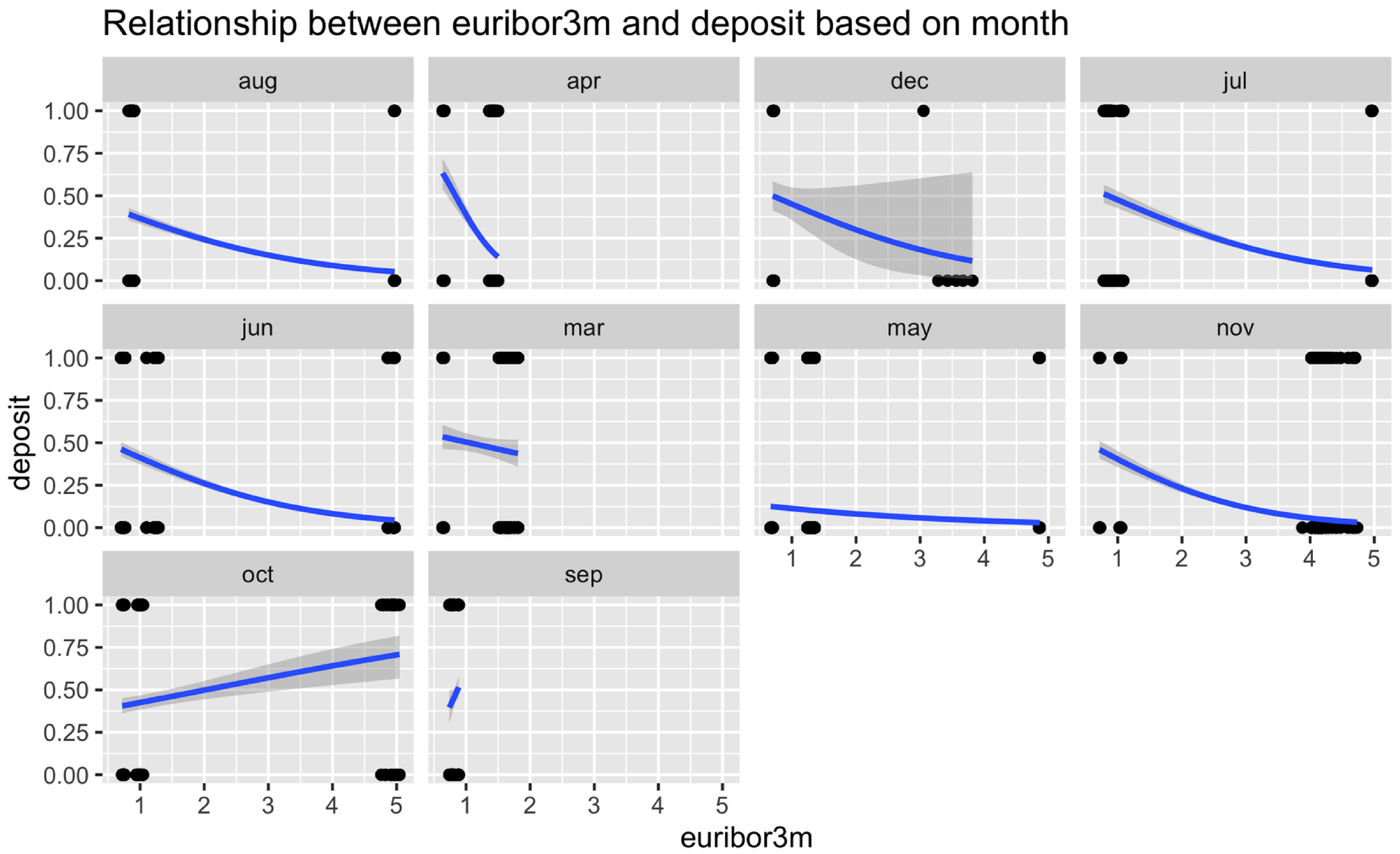
Model 2

01 Greatest Dxy (0.866)

02 P-values of most months are significant

03 P-value of drop-in deviance test = 0

Interaction plots



Our process

Model 2

Residual Deviance: 12267

Dxy: 0.866

Some p-values of month categories not significant

Add interaction term month:euribor3m

Model 5

Residual Deviance: 12031

Dxy: 0.872

Some p-values of month categories not significant

Model 1

```
educationbasic.6y                1.770e-01  1.346e-01   1.315 0.188473
educationbasic.9y                8.958e-02  1.067e-01   0.840 0.400958
educationhigh.school             1.414e-01  1.051e-01   1.345 0.178665
educationilliterate              1.181e+00  8.270e-01   1.428 0.153161
educationprofessional.course     2.212e-01  1.150e-01   1.923 0.054458 .
educationuniversity.degree       2.847e-01  1.060e-01   2.686 0.007236 **
```

Our process

Model 5

Residual Deviance: 12267

Dxy: 0.872

Some p-values of month categories not significant

Add higherEd

Model 6

Residual Deviance: 12016

Dxy: 0.873

P-value of higherEd: 9.20e-05

Model 6

01 Greatest Dxy (0.873)

02 Smallest Residual Deviance

03 P-values of most month:euribor3m are significant

Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.926e+02	1.942e+01	-9.914	< 2e-16	***
telephone1	-4.998e-01	1.006e-01	-4.968	6.78e-07	***
previous	1.606e-01	3.971e-02	4.045	5.24e-05	***
emp.var.rate	-1.506e+00	2.334e-01	-6.451	1.11e-10	***
cons.price.idx	1.920e+00	2.016e-01	9.522	< 2e-16	***
cons.conf.idx	7.659e-02	1.184e-02	6.471	9.74e-11	***
logDuration	2.278e+00	4.103e-02	55.523	< 2e-16	***
monthapr	9.520e-01	4.959e-01	1.920	0.054884	.
monthdec	-1.909e+00	4.776e-01	-3.997	6.42e-05	***
monthjul	-4.581e-01	2.331e-01	-1.966	0.049351	*
monthjun	-2.874e-01	2.547e-01	-1.128	0.259281	
monthmar	-7.295e-01	4.060e-01	-1.797	0.072347	.
monthmay	-5.148e-01	2.286e-01	-2.252	0.024325	*
monthnov	-8.559e-01	2.489e-01	-3.439	0.000585	***

Interpretation

```
monthoct                -1.697e+00  2.101e-01  -8.076 6.70e-16 ***
monthsep                -5.287e+00  2.519e+00  -2.099 0.035806 *
euribor3m                4.183e-01  1.925e-01   2.173 0.029756 *
higherEd                 1.903e-01  4.867e-02   3.911 9.20e-05 ***
monthapr:euribor3m      -1.067e+00  4.019e-01  -2.655 0.007932 **
monthdec:euribor3m       1.899e+00  5.104e-01   3.721 0.000199 ***
monthjul:euribor3m      -4.836e-03  5.116e-02  -0.095 0.924686
monthjun:euribor3m      -2.111e-01  5.493e-02  -3.843 0.000122 ***
monthmar:euribor3m       2.174e+00  3.502e-01   6.208 5.35e-10 ***
monthmay:euribor3m      -2.673e-01  7.488e-02  -3.570 0.000357 ***
monthnov:euribor3m      -6.019e-02  6.483e-02  -0.928 0.353176
monthoct:euribor3m       1.112e+00  1.461e-01   7.610 2.75e-14 ***
monthsep:euribor3m       6.002e+00  3.005e+00   1.997 0.045817 *
```

Interpretation

```
    Null deviance: 21272  on 30593  degrees of freedom
Residual deviance: 12016  on 30567  degrees of freedom
```

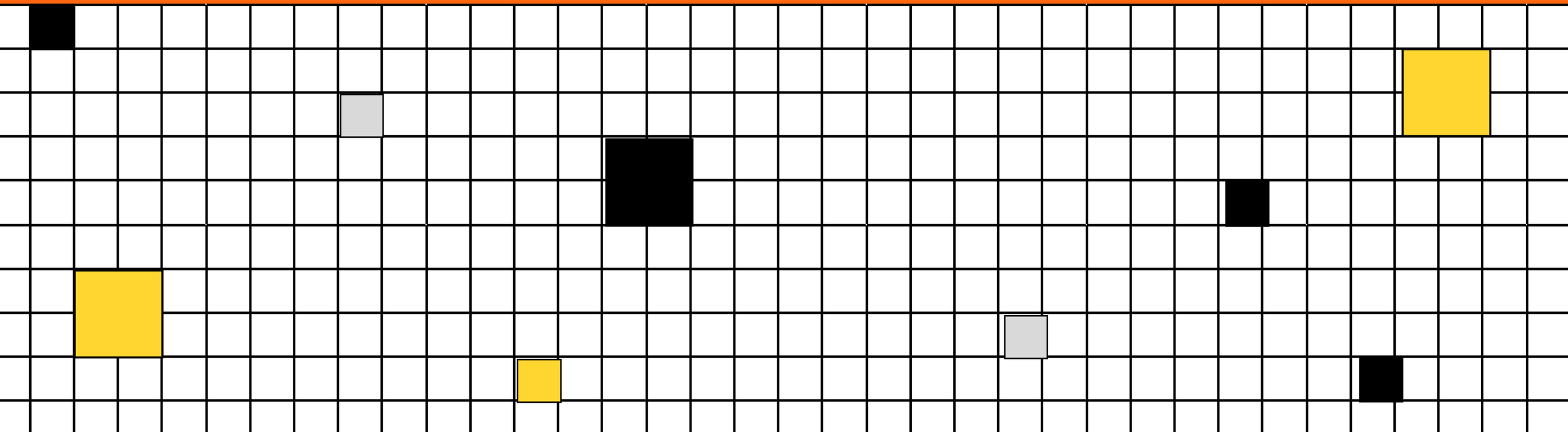
Rank Discrim. Indexes	
C	0.936
Dxy	0.873
gamma	0.873
tau-a	0.172

Interpretation

- Bank should look at the economy of the country before launching a telemarketing campaign
- They should also train their telemarketing team to handle different type of calls appropriately
- Avoid the end of the year for telemarketing campaigns
- They might also aim their campaign at customers who have higher education levels

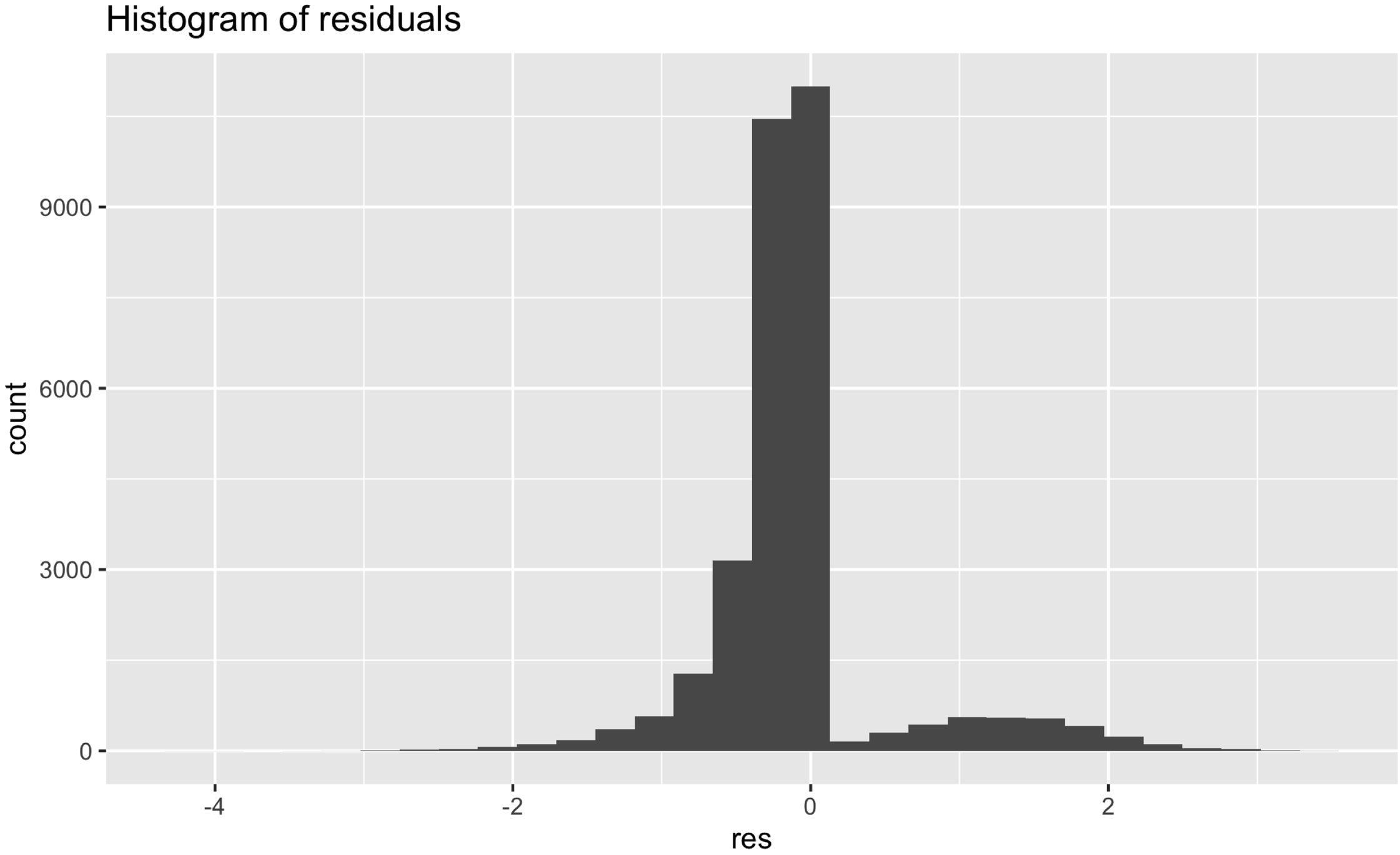
Validation

Logistic Regression



Drop-in deviance test

```
[1] 0
```

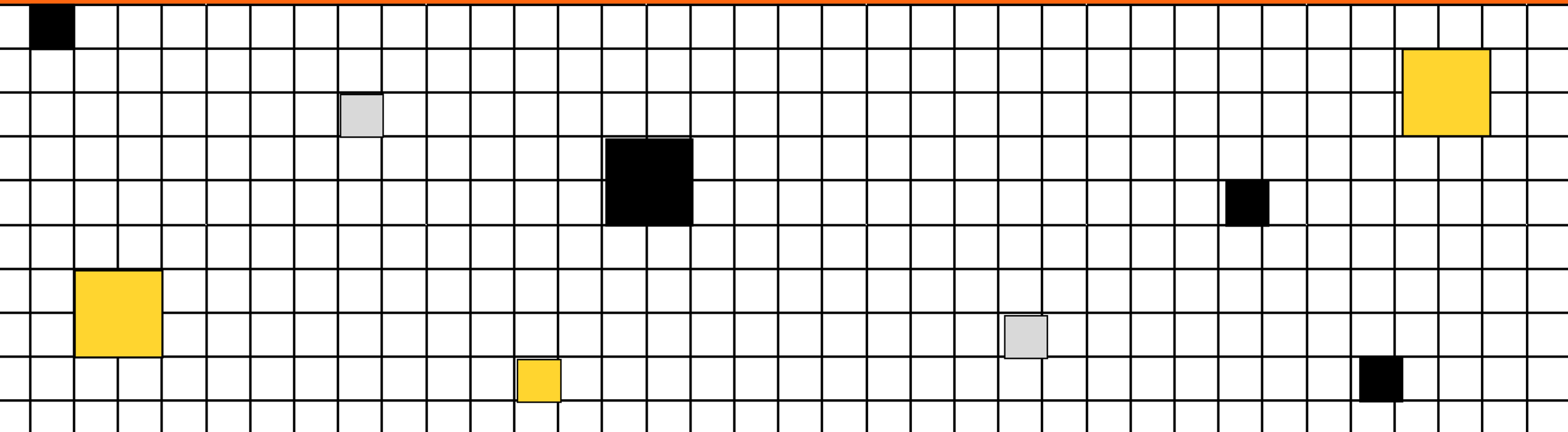

Histogram of residuals
count
9000
6000
3000
0
-4
-2
0
2
res

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  bankTrain2$deposit, bankTrain2$fit
X-squared = 2891.9, df = 2998, p-value = 0.9159
```

Prediction

Logistic Regression



Random Prediction

We randomly choose three data points in the test set

deposit <dbl>	result <dbl>
1	0.27135977
0	0.01087991
0	0.06210997

Recall, Precision, Accuracy

Recall:
TP/(TP+FN)

43.8%

Precision:
TP/(TP+FP)

67.5%

Accuracy:
(TP+TN)/Total

91.1%

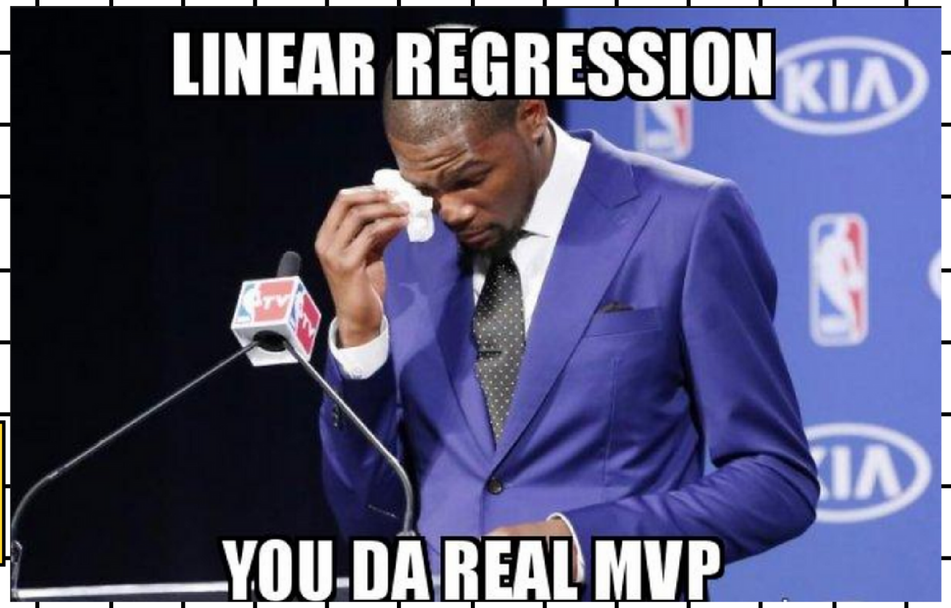
Future thoughts

- We have an unbalanced dataset => oversampling undersampling methods to get more reliable results

- Analyze whether or not we can apply this model to other similiar-size banks in Portugal

AUTHORS: KRYSTAL LY & AMNA KHALID

Thank you for your time!



Email or message for any questions or clarifications