Business Statistics 41000

Tabloid data

Mladen Kolar

Problem

A large retailer wants to explore the predictability of response to a tabloid mailing.

If they mail a tabloid to a customer in their data-base, can they predict whether or not the customer will respond by making a purchase.

The dependent variable is 1 if they buy something, 0 if they do not.

They tried to come up with x's based on past purchasing behavior.

Data

The Predictive Analytics team builds a model for
the probability the customer responds given
information about the customer.

What information about a customer do they use?

- nTab: number of past orders.
- moCbook: months since last order.
- iRecMer1 : 1/months since last order in merchandise category 1.
- llDol: log of the dollar value of past purchases.

The data for these variables is obtained from the companies operational data base.

The retailer decided to perform an experiment by randomly picking 10,000 households to mail the tabloid to.

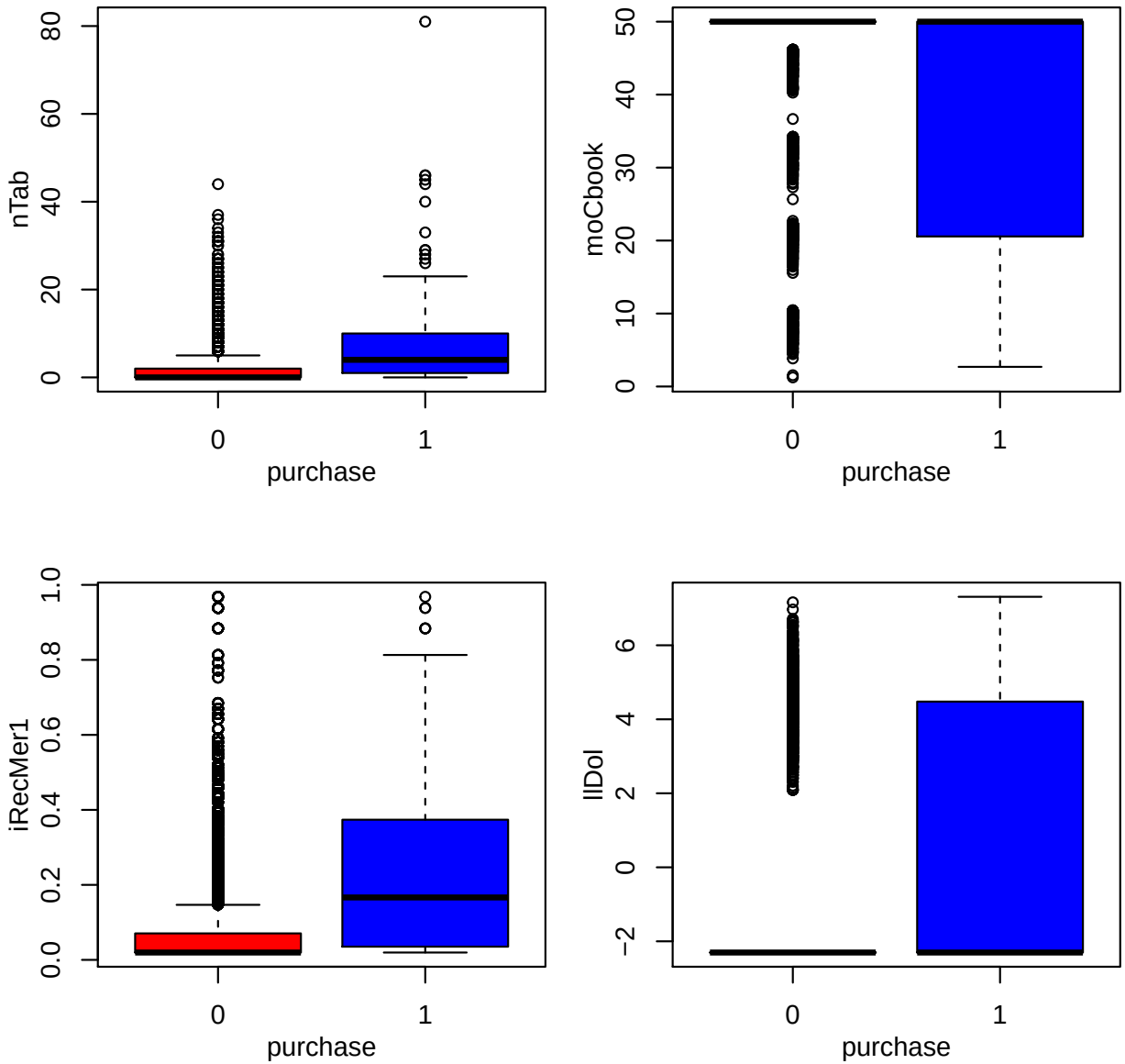
```
td = read.csv("tabloid.csv")
td$purchase = as.factor(td$purchase)
summary(td)
```

```
##  purchase      nTab           moCbook          iRecMer1
##  0:9742   Min.   : 0.000   Min.   : 1.248   Min.   :0.01961
##  1: 258   1st Qu.: 0.000   1st Qu.:50.000   1st Qu.:0.01961
##           Median : 0.000   Median :50.000   Median :0.01961
##           Mean   : 1.857   Mean   :47.597   Mean   :0.09362
##           3rd Qu.: 2.000   3rd Qu.:50.000   3rd Qu.:0.07398
##           Max.   :81.000   Max.   :50.000   Max.   :0.96819
##      llDol
##  Min.   :-2.303
##  1st Qu.:-2.303
##  Median :-2.303
##  Mean   :-1.387
##  3rd Qu.:-2.303
##  Max.   : 7.310
```

Notice that the percentage of households that make a purchase is pretty small!

$258/10000 = 0.0258$

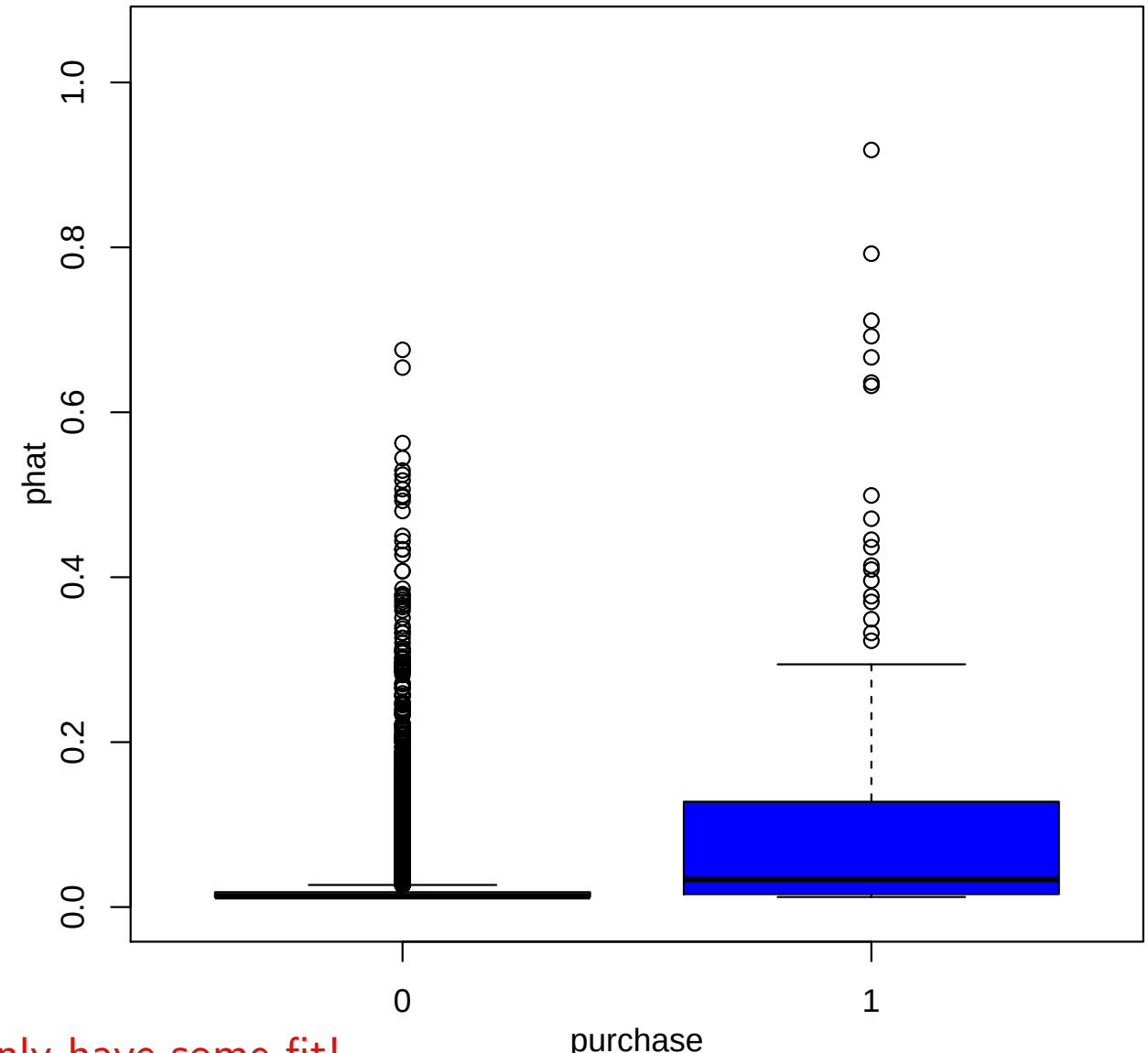
Here is `Y` plotted vs. each of the four `X`'s



```r
lgfit = glm(purchase~nTab+moCbook+iRecMer1+llDol,td,family=binomial)
summary(lgfit)
```

```
## Call:
## glm(formula = purchase ~ nTab + moCbook + iRecMer1 + llDol, family = binomial,
##     data = td)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.62131    0.25667  -10.21  < 2e-16 ***
## nTab         0.05530    0.01209    4.57  4.8e-06 ***
## moCbook     -0.03249    0.00527   -6.17  7.0e-10 ***
## iRecMer1     1.72688    0.31282    5.52  3.4e-08 ***
## llDol        0.07842    0.02630    2.98   0.0029 **
```

```
par(mar=c(3,3,3,1), mgp=c(2,1,0))
phat = predict(lgfit, newdata=td, type="response")
plot(phat~td$purchase, col=c("red","blue"),
     xlab="purchase", ylab="phat", ylim=c(0,1.05), cex.text=0.7)
```



We certainly have some fit!

The idea behind the tabloid example is that if we can predict who will buy we can target those customers and send them the tabloid.

To get an idea of how well our model is working, we can imagine choosing a customer from the data set to mail to first - did they buy?

We can look at the y value to see if they bought.

Whom would you mail to first?

You could mail the first 40 people in your database.

```
##    purchase   phat
## 1         0 0.0122
## 2         1 0.0670
## 3         0 0.0153
## 4         0 0.0129
## 5         0 0.0122
## 6         0 0.0429
## 7         0 0.0124
## 8         0 0.0122
## 9         0 0.0223
## 10        0 0.0122
## 11        0 0.0399
## 12        0 0.0122
## 13        0 0.0353
## 14        0 0.0163
## 15        0 0.0288
## 16        0 0.0125
## 17        0 0.0175
## 18        0 0.0122
## 19        0 0.0200
## 20        0 0.0122
```

```
##      purchase      phat
## 21        0 0.0184
## 22        0 0.0203
## 23        0 0.0122
## 24        0 0.0122
## 25        0 0.0144
## 26        0 0.0122
## 27        0 0.0122
## 28        0 0.0131
## 29        0 0.0160
## 30        0 0.0122
## 31        0 0.0122
## 32        0 0.0122
## 33        0 0.0265
## 34        0 0.0122
## 35        0 0.0122
## 36        0 0.0274
## 37        0 0.0122
## 38        0 0.0123
## 39        0 0.0122
## 40        0 0.0136
```

Out of the first 40, there is only one purchase.

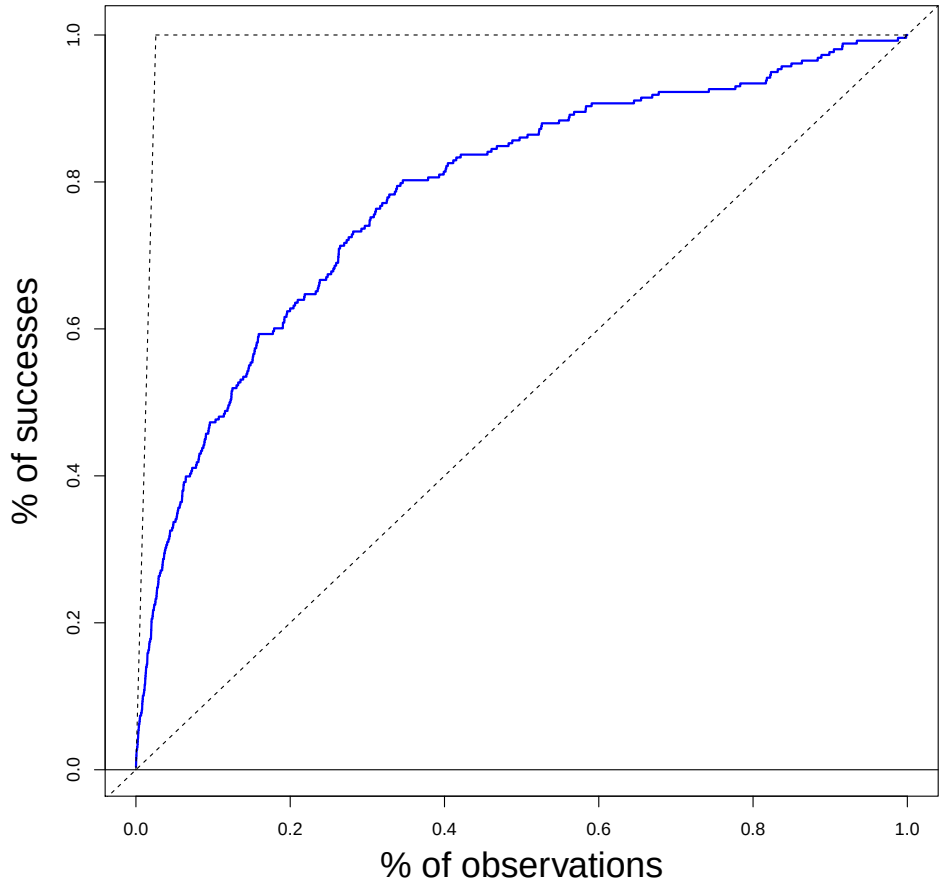
If you believe your model, you might mail to the household with the largest $\hat{p}$ (estimated prob of buying) first. Then you would mail to the household with the second largest $\hat{p}$ and so on.

```
##      purchase      phat
## 2000        1 0.9180914
## 2755        1 0.7923425
## 8862        1 0.7111080
## 3628        1 0.6922268
## 1284        0 0.6756920
## 529         1 0.6665625
## 8086        0 0.6542265
## 2072        1 0.6360684
## 1435        1 0.6320182
## 4524        0 0.5626667
## 4626        0 0.5444024
## 978         0 0.5293640
## 9351        0 0.5243046
## 7040        0 0.5172840
## 7424        0 0.5067277
## 6545        1 0.4990952
## 5716        0 0.4988779
## 1218        0 0.4973493
## 374         0 0.4926355
## 521         0 0.4802793
```

```
##      purchase      phat
## 1887        1 0.4709141
## 7703        0 0.4501004
## 789         1 0.4455207
## 8931        0 0.4438132
## 3853        1 0.4364154
## 5239        0 0.4338959
## 2999        0 0.4336161
## 6997        0 0.4271745
## 3526        1 0.4141329
## 8566        1 0.4092660
## 891         0 0.4074384
## 2417        0 0.4073038
## 5214        1 0.3958348
## 8490        0 0.3861329
## 6594        0 0.3795044
## 4548        0 0.3777539
## 6147        1 0.3770014
## 6548        0 0.3758676
## 1637        0 0.3727244
## 4748        1 0.3700390
```

You got 16 purchases out of the first 40 customers you targeted. Using only $40/10000 = 0.004$ of the data we got $16/258 = .062$ of the purchases!

The Lift Curve



Middle: using $\hat{p}$, after using 20% of the data, I have 60% of the purchases.

Bottom: guessing, after using 20% of the data, I have 20% of the purchases.

Top: perfect knowledge. .0258 of the data are purchases. Once I have this much data, I have all the purchases.