MARKETING ANALYTICS IN R: STATISTICAL MODELING

Welcome to this Chapter! Churn Prevention in Online Marketing

Verena Pflieger
Data Scientist at INWT Statistics

Churn Prevention

Binary Logistic Regression

`1) Probability to churn`

$$P(Y=1)$$

`2) log Odds`

$$\log\frac{P(Y=1)}{P(Y=0)}=\beta_0+\sum_{p=1}^{P}\beta_p x_p$$

`3) Odds`

$$\frac{P(Y=1)}{P(Y=0)}=e^Z,\ \text{with}\quad Z=\beta_0+\sum_{p=1}^{P}\beta_p x_p$$

`4) Probability to churn`

$$P(Y=1)=\frac{e^Z}{1+e^Z}$$

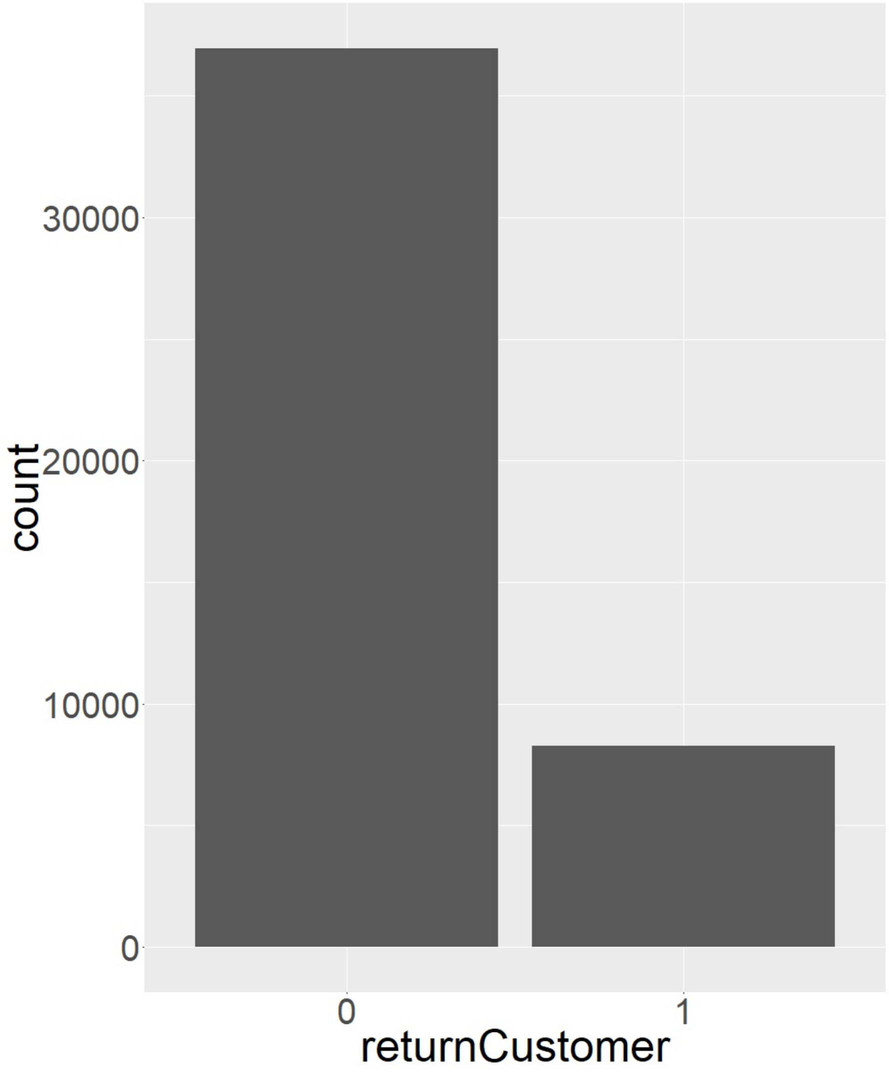
Data Discovery I

```
## 'data.frame': 45236 obs. of 21 variables:

## $ ID             : Factor w/ 45236 levels "1","3","5","7",..
## $ orderDate      : Date, format: "2014-12-23" "2014-09-10" ....
## $ title          : Factor w/ 4 levels "Mr","Company",..: 1 1 1 ...
## $ newsletter     : Factor w/ 2 levels "No","Yes": 0 0 0 1 ...
## $ websiteDesign  : Factor w/ 3 levels "1","2","3": 2 1 1 3 ...
## $ paymentMethod  : Factor w/ 4 levels "Cash","Credit Card",..: 3 4 ...
## $ couponDiscount : Factor w/ 2 levels "No","Yes": 1 0 0 0 0 1 0 0 ...
...
## $ returnCustomer : Factor w/ 2 levels "No","Yes": 0 0 0 0 ...
```

Data Discovery II

```
ggplot(churnData, aes(x = returnCustomer)) +
    geom_histogram(stat = "count")
```



MARKETING ANALYTICS IN R: STATISTICAL MODELING

Let's start analyzing!

MARKETING ANALYTICS IN R: STATISTICAL MODELING

Modeling & Model Selection

Verena Pflieger
Data Scientist at INWT Statistics

Model Specification

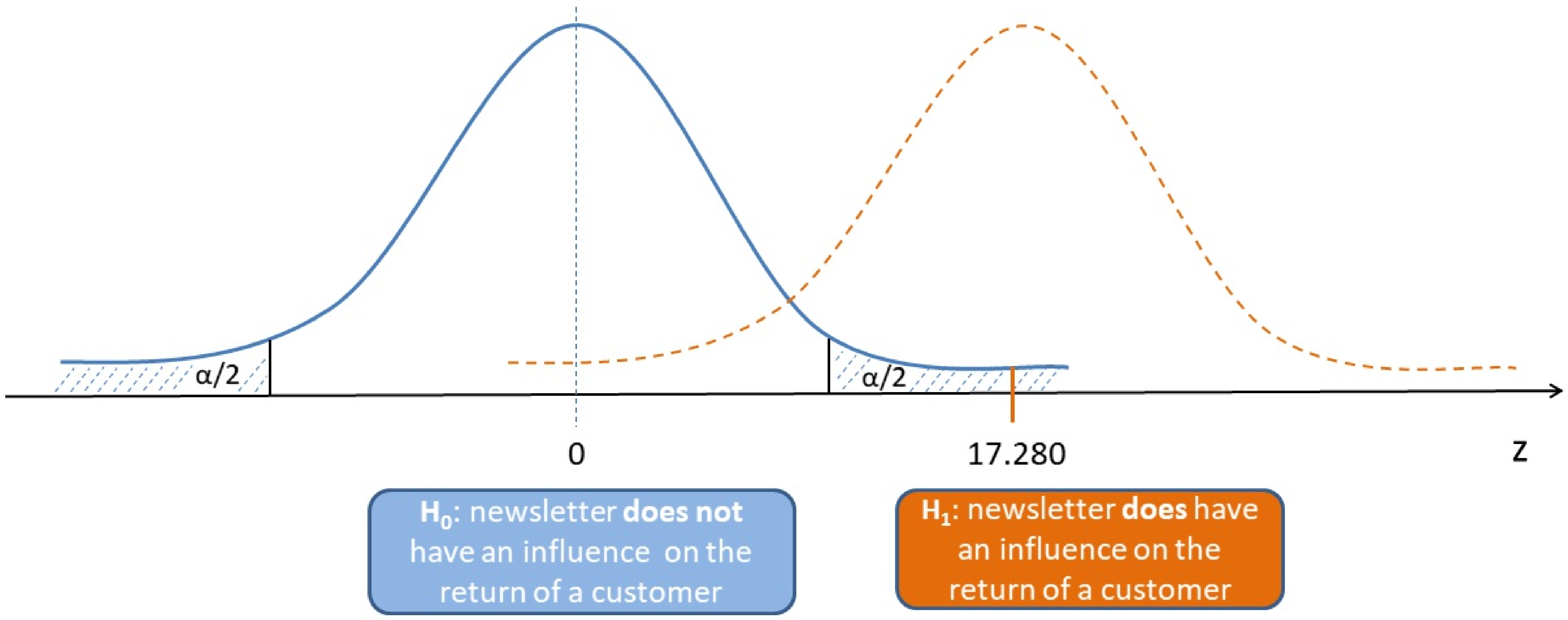
```
logitModelFull <- glm(returnCustomer ~ title + newsletter + websiteDesign +
    ..., family = binomial, churnData)

summary(logitModelFull)
```

```
## Coefficients:
##                            Estimate  Std.Error  z value  Pr(>|z|)
## (Intercept)                -1.49074  0.04930    -30.239  < 2e-16   ***
## titleCompany               -0.21215  0.05286    -4.013   5.99e-05  ***
## titleMrs                    0.03086  0.02953     1.045   0.29586
## newsletter1                 0.52373  0.03031     17.280  < 2e-16   ***
## websiteDesign2             -0.45679  0.16267    -2.808   0.00498    **
## websiteDesign3             -0.28800  0.15899    -1.811   0.07007     .
## paymentMethodCredidCard    -0.24192  0.04843    -4.995   5.89e-07  ***
## tvEquipment                 -0.51475  1.08141    -0.476   0.63408
...
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
...
## AIC: 41762
```

Statistical Significance

```
## Coefficients:
##                         Estimate  Std.Error  z value  Pr(>|z|)
## ...
## newsletter1             0.52373  0.03031     17.280  < 2e-16   ***
## ...
```



Coefficient Interpretation

Log odds equation:

$$\log \frac{P(returnCustomer=1)}{P("returnCustomer"=0)} = -1.49 - 0.21 \cdot titleCompany + 0.52 \cdot newsletter1 + \ldots$$

Transformation to odds:

```
coefsExp <- coef(logitModelFull) %>% exp() %>% round(2)
coefsExp
```

```
## (Intercept)  titleCompany    titleMrs    titleOthers
## 0.23         0.81            1.03        1.77

## newsletter1  websiteDesign2  ...
## 1.69         0.63            ...
```

Model Selection

```
library(MASS)

logitModelNew <- stepAIC(logitModelFull, trace = 0)

summary(logitModelNew)
```

```
## Coefficients:
##                          Estimate  Std.Error  z value  Pr(>|z|)
## (Intercept)              -1.49130  0.04928    -30.260  < 2e-16   ***
## titleCompany             -0.21131  0.05285    -3.998   6.38e-05  ***
## titleMrs                  0.03159  0.02951     1.071   0.28432
## newsletter1               0.52332  0.03030     17.269  < 2e-16   ***
...
## videogameDownload         0.26474  0.05256     5.037   4.74e-07  ***
## prodRemitted              0.89528  0.07619     11.751  < 2e-16   ***

## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
...
## AIC: 41756
```

Results of the Step-AIC Function

Removed Variables	Remaining Variables
tvEquipment	newsletter
prodOthers	paymentMethod
	dvd
	blueray
	...

MARKETING ANALYTICS IN R: STATISTICAL MODELING

Let's apply what I have shown you!

MARKETING ANALYTICS IN R: STATISTICAL MODELING

In-Sample Model Fit & Thresholding

Verena Pflieger
Data Scientist at INWT Statistics

Pseudo R^2 Statistics I

McFadden: $$R^2 = 1 - \frac{L_{\text{null}}}{L_{\text{full}}}$$

Cox & Snell: $$R^2 = 1 - \left(\frac{L_{\text{null}}}{L_{\text{full}}}\right)^{\frac{2}{n}}$$

Nagelkerke: $$R^2 = \frac{1 - \left(\frac{L_{\text{null}}}{L_{\text{full}}}\right)^{\frac{2}{n}}}{1 - (L_{\text{null}})^{\frac{2}{n}}}$$

```
Interpretation:

Reasonable if > 0.2

Good if  > 0.4

Very Good if > 0.5
```

Pseudo R^2 Statistics II

```
library(descr)

LogRegR2(logitModelNew)
```

```
## Chi2                          1321.717
## Df                            19
## Sig.                          0
## Cox and Snell Index           0.02879553
## Nagelkerke Index              0.0469131
## McFadden's R2                 0.03071032
```

Predict Probabilities

```
library(SDMTools)
churnData$predNew <- predict(logitModelNew, type = "response",
                     na.action = na.exclude)
data %>% select(returnCustomer, predNew) %>% tail()
```

```
      returnCustomer   predNew
45231              0 0.2843944
45232              0 0.1552756
45233              1 0.2522597
45234              1 0.1454276
45235              0 0.2698819
45236              0 0.2886988
```

Confusion Matrix

Prediction \ Truth	negative	positive
negative	true-negative	false-negative
positive	false-positive	true-positive

```
confMatrixNew <- confusion.matrix(churnData$returnCustomer,
                    churnData$predNew, threshold = 0.5)
confMatrixNew
```

```
##      obs
## pred 0     1
##    0 36921 8242
##    1 43    30
```

Accuracy

```
accuracyNew <- sum(diag(confMatrixNew)) / sum(confMatrixNew)
accuracyNew
```

```
## [1] 0.8168494
```

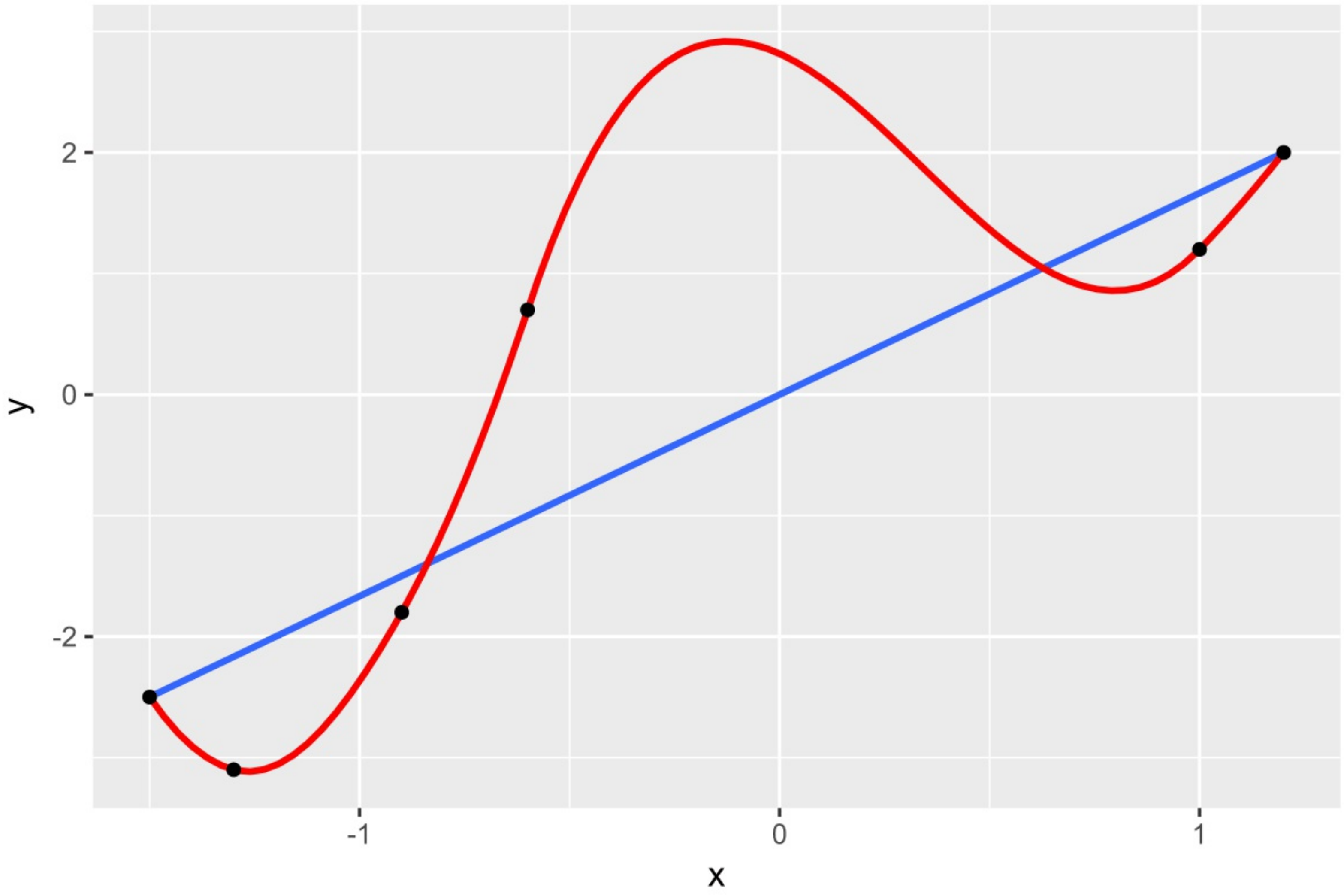
Finding the Optimal Threshold

Prediction \ Truth	returnCustomer = 0	returnCustomer = 1
returnCustomer = 0	5	-15
returnCustomer = 1	0	0

payoff = 5 * true negative - 15 * false negative

Threshold	Accuracy	Payoff
0.5	0.817	60975
0.4	0.815	62180
[0.3]	[0.794]	[65740]
0.2	0.668	65670
0.1	0.241	10550

Overfitting



MARKETING ANALYTICS IN R: STATISTICAL MODELING

Let's try it out!

MARKETING ANALYTICS IN R: STATISTICAL MODELING

Out-of-Sample Validation and Cross-Validation

Verena Pflieger
Data Scientist at INWT Statistics

Out-of-Sample Fit: Training and Test Data

1) Divide the dataset in training and test data

```
# Generating random index for training and test set
# set.seed ensures reproducibility of random components
set.seed(534381)

churnData$isTrain <- rbinom(nrow(churnData), 1, 0.66)
train <- subset(churnData, churnData$isTrain == 1)
test <- subset(churnData, churnData$isTrain == 0)
```

Out-of-Sample Fit: Building Model

2) Build a model based on training data

```
# Modeling logitTrainNew
logitTrainNew <- glm( returnCustomer ~ title + newsletter + websiteDesign +
        paymentMethod + couponDiscount + purchaseValue + throughAffiliate +
        shippingFees + dvd + blueray + vinyl + videogameDownload +
        prodOthers + prodRemitted, family = binomial, data = train)

# Out-of-sample prediction for logitTrainNew
test$predNew <- predict(logitTrainNew, type = "response", newdata = test)
```

Out-of-Sample Accuracy

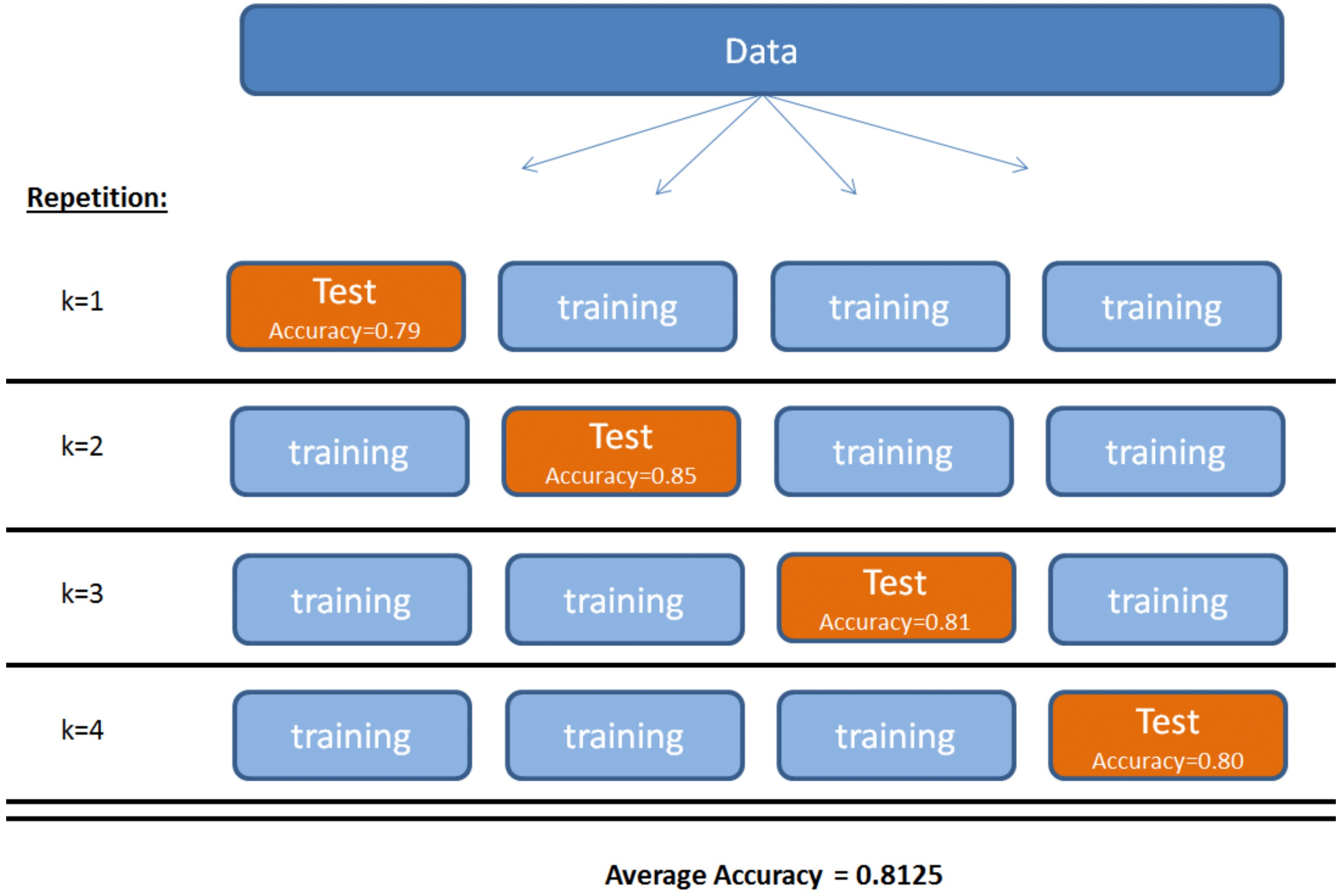
```
#calculating the confusion matrix
confMatrixNew <- confusion.matrix(test$returnCustomer, test$predNew,
                                  threshold = 0.3)
confMatrixNew

#calculating the accuracy
accuracyNew <- sum(diag(confMatrixNew)) / sum(confMatrixNew)
accuracyNew
```

```
    obs
pred     0    1
   0 11939 2449
   1   716  350


[1] 0.7951987
```

Cross-Validation: Set-up



Cross-Validation: Accuracy

Calculation of cross-validated accuracy

```
library(boot)
# Accuracy function with threshold = 0.3
Acc03 <- function(r, pi = 0) {
  cm <- confusion.matrix(r, pi, threshold = 0.3)
  acc <- sum(diag(cm)) / sum(cm)
  return(acc)
}

# Accuracy
set.seed(534381)
cv.glm(churnData, logitModelNew, cost = Acc03, K = 6)$delta
```

```
[1] 0.7943894
```

Learnings and Relevance

	Learnings Logistic Regression
You have learned...	how to predict customers of an online shop that are likely to churn
	to use a binary logistic regression to calculate probabilities
	that the choice of the threshold is crucial

	Learnings from the Model
You have learned...	that customers, signing up for a newsletter are more likely to return
	that customers, using a coupon are less likely to return
	that customers, without shipping fees are more likely to return

MARKETING ANALYTICS IN R: STATISTICAL MODELING

Last Exercise!