

Pilgrim Bank

Luca Bajardi e Francesca Collini

31/07/2020

Carichiamo i dati leggendo il file csv e settiamo il seme del generatore pseudo-casuale così da avere i risultati sempre uguali.

```
rm(list =ls())
set.seed(1)
Pilgrim = read.csv(file = "PilgrimABC.csv", header=T)
attach(Pilgrim)
```

Osserviamo che il nostro dataset contiene 31634 osservazioni e 11 variabili (le ultime due non sono interessanti ai fini della nostra analisi). I dati sono relativi a due anni, 1999 e 2000, per ciascuno abbiamo a disposizione un'informazione relativa al profitto (o alla perdita) di un determinato cliente in quell'anno e un'informazione che ci dice se quel cliente in quell'anno ha utilizzato l'opzione di online banking oppure no: la variabile Online quindi è una variabile binaria. Abbiamo poi a disposizione altre informazioni relative al cliente:

- AGE: variabile categorica che indica a che fascia di età appartiene il cliente (1 = less than 15 years; 2 = 15-24 years; 3 = 25-34 years; 4 = 35-44 years; 5 = 45-54 years; 6 = 55-64 years; 7 = 65 years and older.)
- INCOME: variabile categorica che indica il range del reddito di ciascun cliente (1 = less than \$15,000; 2 = \$15,000 – \$19,999; 3 = \$20,000 – \$29,999; 4 = \$30,000 – \$39,999; 5 = \$40,000 – \$49,999; 6 = \$50,000 – \$74,999; 7 = \$75,000 – \$99,999; 8 = \$100,000 – \$124,999; 9 = \$125,000 and more.)
- TENURE: rappresenta l'età di servizio
- DISTRICT: variabile categorica che indica uno delle tre regioni geografiche in cui si trova il cliente.

```
dim(Pilgrim)
```

```
## [1] 31634      11
names(Pilgrim[, 1:9])
```



```
## [1] "ID"          "X9Profit"     "X9Online"     "X9Age"        "X9Inc"
## [6] "X9Tenure"    "X9District"   "X0Profit"     "X0Online"
```

Rinominiamo le colonne per evitare problemi con i nomi:

```
Profit=Pilgrim$X9Profit
Online=Pilgrim$X9Online
Age=Pilgrim$X9Age
Income=Pilgrim$X9Inc
Tenure=Pilgrim$X9Tenure
District=Pilgrim$X9District
```

```
head(Pilgrim)
```

```
##   ID X9Profit X9Online X9Age X9Inc X9Tenure X9District X0Profit X0Online
## 1  1       21       0    NA    NA     6.33      1200      NA      NA
## 2  2      -6       0     6     3    29.50      1200     -32      0
```

```

## 3 3 -49 1 5 5 26.41 1100 -22 1
## 4 4 -4 0 NA NA 2.25 1200 NA NA
## 5 5 -61 0 2 9 9.91 1200 -4 0
## 6 6 -38 0 NA 3 2.33 1300 14 0
## X9Billpay X0Billpay
## 1 0 NA
## 2 0 0
## 3 0 0
## 4 0 NA
## 5 0 0
## 6 0 0

```

Dal summary del profitto possiamo facilmente osservare che la distribuzione è molto asimmetrica perché media e mediana sono molto diverse tra loro. Inoltre notiamo che tutto il primo quartile è negativo.

```
summary(Profit)
```

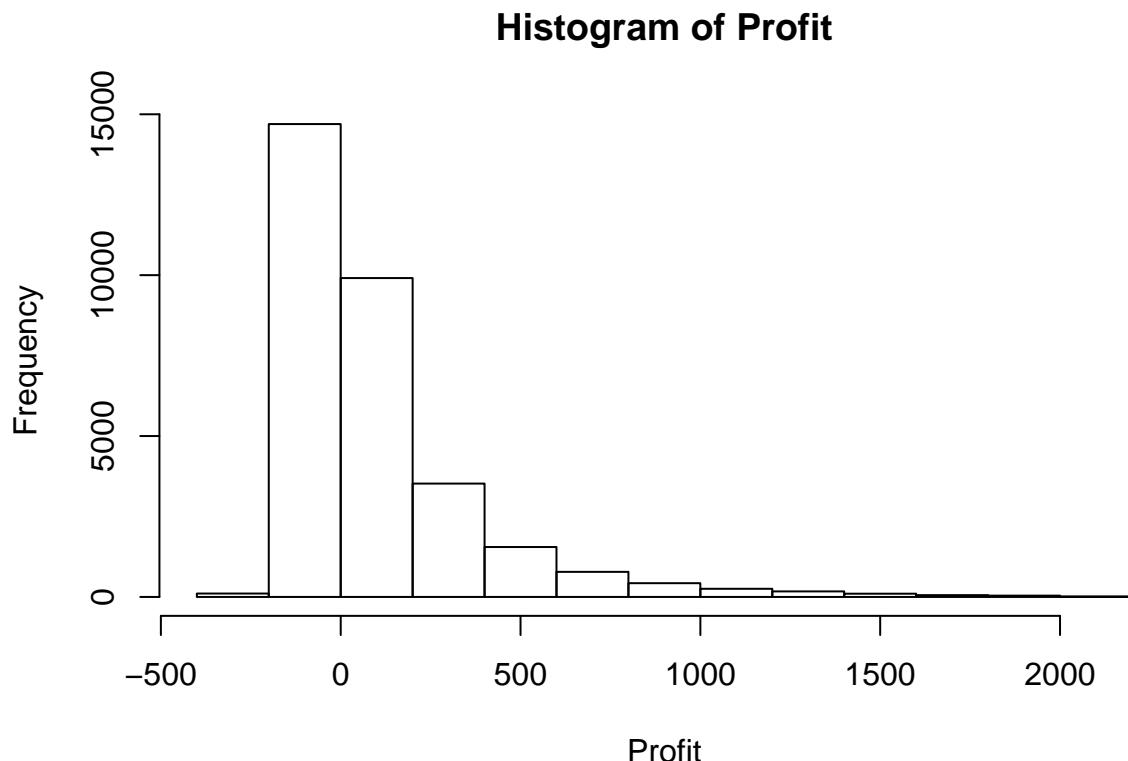
```

##   Min. 1st Qu. Median Mean 3rd Qu. Max.
## -221.0 -34.0 9.0 111.5 164.0 2071.0

```

Dall'istogramma sul profitto nel 1999 possiamo notare che c'è un discreto numero di clienti che mi fa perdere e che l'istogramma è molto scodato a destra quindi ci sono pochissimi clienti che mi fanno guadagnare molto.

```
hist(Profit)
```



Possiamo notare che ci sono 16832 clienti che generano profitto sulle 31634 presenti nel dataset.

```

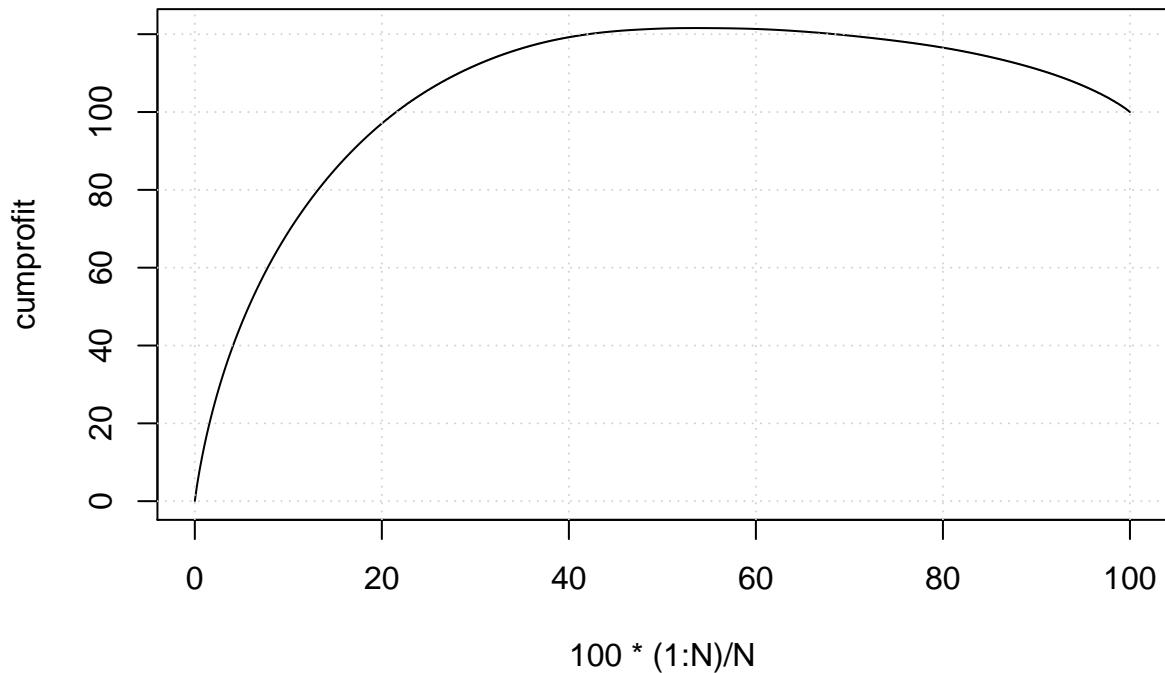
N=length(Profit)
Nprofitable = sum(Profit>0)
cat('profitable = ', Nprofitable, ' out of ', N, '\n')

```

```
## profitable = 16832 out of 31634
```

Tramite la curva di Pareto possiamo confermare che poche persone fanno guadagnare tanto, mentre la maggior parte delle persone fa guadagnare poco o addirittura fa perdere guadagno.

```
cumprofit=cumsum(sort(Profit,decreasing=TRUE))*100/sum(Profit)
plot(100*(1:N)/N,cumprofit,type='l')
grid()
```



Dall'analisi dei profitti medi possiamo notare che il profitto generato da chi utilizza i servizi online è maggiore rispetto a chi li utilizza offline.

```
cat('average profit ', mean(Profit), '\n')

## average profit 111.5027

ProfitOnline = Profit[Online==1]
cat('average profit ON', mean(ProfitOnline), '\n')

## average profit ON 116.6668

ProfitOffline = Profit[Online==0]
cat('average profit OFF', mean(ProfitOffline), '\n')

## average profit OFF 110.7862
```

Quello che vorremmo capire è se questa differenza è stocasticamente significativa, quindi se davvero coloro che utilizzano i servizi online mi permettono di guadagnare di più. Per questo motivo possiamo eseguire un `t.test` sulle due popolazioni, i clienti che utilizzano il servizio online e coloro che non lo usano, e vado a vedere quanto vale il p-value. L'ipotesi nulla in questo caso è che le medie siano uguali e che quindi la

differenza tra le medie delle due popolazioni non sia significativa. Analizziamo inoltre l'intervallo di confidenza per vedere se i dati sono sufficienti o no. Infatti se intervallo fosse troppo grande vorrebbe dire che avrei bisogno di più clienti per formulare una teoria generale.

```
cat('Conf int profit ', t.test(Profit)$conf.int, '\n')

## Conf int profit 108.496 114.5094

cat('p-value difference ', t.test(ProfitOnline, ProfitOffline)$p.value, '\n')

## p-value difference 0.2254368
```

Risulta esserci una differenza di circa 6 dollari tra le due popolazioni, quindi non risulta più di tanto significativa dal lato business, ma anche il p-value è maggiore di 0.05 e quindi possiamo concludere che questa differenza non è significativa nemmeno dal punto di vista statistico, e quindi non rifiutiamo l'ipotesi nulla sulla differenza tra le medie.

Possiamo ottenere lo stesso risultato impostando il modello di regressione classico:

```
mod = lm(Profit ~ Online)
summary(mod)

##
## Call:
## lm(formula = Profit ~ Online)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -337.67 -144.79 -101.79   52.21 1960.21 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 110.786     1.637   67.678 <2e-16 ***
## Online       5.881      4.690    1.254    0.21    
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 272.8 on 31632 degrees of freedom
## Multiple R-squared:  4.97e-05, Adjusted R-squared:  1.809e-05 
## F-statistic: 1.572 on 1 and 31632 DF,  p-value: 0.2099
```

Dal summary del modello, possiamo osservare che il valore dell'intercetta è la media del profitto tra coloro che operano offline e il valore aggiunto di quelli che operano online è di quasi 6 dollari come la differenza tra la media di quelli che operano online e la media di quelli che operano offline. Anche in questo caso però, il p-value è superiore a 0.05 come in precedenza. Questo significa che la variabile Online non è significativa. Tuttavia, questo potrebbe essere dovuto al fatto che stiamo mettendo insieme clienti molto diversi tra loro e quindi considerando un unico modello non riusciamo a distinguere bene i singoli effetti. Per questo introduciamo l'età nel modello di regressione.

La variabile **Age** è riconosciuta numerica anche se in realtà è categorica. Nell'analisi dovremo trasformarla in **factor** perché altrimenti vi è troppa influenza dell'ordinamento delle variabili categoriali.

```
Age1 = as.factor(Age)
summary(Age1)

## 1 2 3 4 5 6 7 NA's
## 710 3650 5390 5376 3236 2290 2693 8289
```

Possiamo notare che ci sono 8289 osservazioni in cui non è presente l'età, infatti nel **summary** sottostante

possiamo notare che nonostante ci siano 32 mila osservazioni ci sono solo 23 mila gradi di libertà in quanto quelle con i missing values sono eliminate automaticamente.

```
mod = lm(Profit ~ Online+Age1)
summary(mod)

##
## Call:
## lm(formula = Profit ~ Online + Age1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -404.52 -162.90  -84.62   68.80 1952.10 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -1.802     10.485  -0.172   0.864    
## Online       27.246     5.519   4.937 8.01e-07 *** 
## Age12        54.425    11.401   4.774 1.82e-06 *** 
## Age13        112.699   11.098   10.155 < 2e-16 *** 
## Age14        133.820   11.103   12.053 < 2e-16 *** 
## Age15        144.986   11.531   12.574 < 2e-16 *** 
## Age16        160.844   11.965   13.443 < 2e-16 *** 
## Age17        193.072   11.757   16.422 < 2e-16 *** 
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
##
## Residual standard error: 277.9 on 23337 degrees of freedom
##   (8289 observations deleted due to missingness)
## Multiple R-squared:  0.02494,   Adjusted R-squared:  0.02464 
## F-statistic: 85.26 on 7 and 23337 DF,  p-value: < 2.2e-16
```

Ora sulle fasce d'età vi è una significatività sia statistica che di business, infatti le diverse età potrebbero implicare diversi redditi, in linea generale un giovane tende ad usare di più l'online banking ma ha un reddito minore e quindi fa guadagnare di meno.

Posso notare che a prescindere dall'utilizzo dell'online o dell'offline i giovani sono meno profittevoli degli anziani:

```
lapply(split(Profit,as.factor(Age)),mean)
```

```
## $`1` 
## [1] 3.45493
## 
## $`2` 
## [1] 58.48959
## 
## $`3` 
## [1] 115.1122
## 
## $`4` 
## [1] 135.6618
## 
## $`5` 
## [1] 145.7596
## 
```

```

## $`6`
## [1] 160.41
##
## $`7`
## [1] 192.2614

```

Inoltre il 20% dei giovani usa l'online, questa percentuale scende per le fasce di età successive fino ad arrivare a poco più del 3%.

```
lapply(split(Online,as.factor(Age)),mean)
```

```

## $`1`
## [1] 0.1929577
##
## $`2`
## [1] 0.2153425
##
## $`3`
## [1] 0.154731
##
## $`4`
## [1] 0.1337426
##
## $`5`
## [1] 0.09456119
##
## $`6`
## [1] 0.05021834
##
## $`7`
## [1] 0.03639064

```

Ci potrebbe essere un bias dovuto ai dati mancanti. Infatti ci sono molte osservazioni in cui non abbiamo l'informazione sull'età. Non sappiamo perché questi dati sono mancanti, anche se è difficile pensare che sia semplicemente dovuto a delle dimenticanze o sviste in fase di raccolta dei dati, forse avrebbe più senso pensare ai conti contestati oppure al fatto che il conto venga intestato ad una società.

```
sum(is.na(Age))
```

```
## [1] 8289
```

```
AgeGiven = ifelse(is.na(Age),0,1) # 0 dove c'è NA, 1 se c'è l'età
mod = lm(Profit ~ AgeGiven)
summary(mod)
```

```

##
## Call:
## lm(formula = Profit ~ AgeGiven)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -346.19  -150.19   -90.96   50.81 1961.04
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 72.962     2.986   24.43   <2e-16 ***
## AgeGiven    52.224     3.476   15.02   <2e-16 ***

```

```

## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 271.9 on 31632 degrees of freedom
## Multiple R-squared: 0.007085, Adjusted R-squared: 0.007054
## F-statistic: 225.7 on 1 and 31632 DF, p-value: < 2.2e-16

```

C'è una differenza statisticamente significativa sul profitto tra le osservazioni in cui è presente l'informazione sull'età e dove non c'è. La variabile AgeGiven infatti risulta significativa, la distribuzione tra le due popolazioni di conseguenza non risulta omogenea. Quindi, eliminando i dati dove manca l'età, stiamo distorcendo l'analisi, perché queste medie sono "sporcate". Infatti c'è una profitabilità media più alta tra chi mi ha dato l'età rispetto a chi non me l'ha data.

Possiamo provare diversi metodi per cercare di recuperare quelle osservazioni dove l'età non è presente, in modo da includerle comunque nell'analisi. Se avessimo delle righe complete ed alcune con un solo campo mancante, potremmo costruire un modello di regressione in modo da prevedere quel valore. Una possibile alternativa, più immediata, è quella di sostituire i valori mancanti con i valori nulli. Questa strada sembra abbastanza convincente, quando l'età è una variabile categorica e possiamo quindi scegliere un valore arbitrario. Quando invece la variabile è numerica non ha molto senso, perché equivale a dire che coloro che non hanno fornito l'età, risultano neonati.

```

AgeZero = ifelse(is.na(Age), 0, Age)
table(AgeZero)

## AgeZero
##   0   1   2   3   4   5   6   7
## 8289  710 3650 5390 5376 3236 2290 2693

mod = lm(Profit ~ Online+AgeZero)
summary(mod)

## 
## Call:
## lm(formula = Profit ~ Online + AgeZero)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -393.91 -147.07  -82.03   49.97 1976.97 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 57.0311    2.6014  21.923 < 2e-16 ***
## Online      13.7925    4.6487   2.967  0.00301 **  
## AgeZero     17.6803    0.6697  26.402 < 2e-16 *** 
## ---      
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 269.9 on 31631 degrees of freedom
## Multiple R-squared: 0.02161, Adjusted R-squared: 0.02155 
## F-statistic: 349.3 on 2 and 31631 DF, p-value: < 2.2e-16

```

I valori dei coefficienti sono dovuti al fatto che in modo arbitrario abbiamo scelto di sostituire i valori mancanti con il valore nullo.

Replace missing with mean

Una possibile alternativa è quella di sostituire i valori mancanti con la media dell'età calcolata sui valori presenti.

```
mm = mean(Age, na.rm=TRUE) #non consideriamo i valori mancanti
AgeAverage = ifelse(is.na(Age), mm, Age)
table(AgeAverage)
```

```
## AgeAverage
##          1          2          3          4
##      710      3650      5390      5376
## 4.04604840436924
##          5          6          7
##      8289      3236      2290      2693
```

```
mod = lm(Profit ~ Online+AgeAverage)
summary(mod)
```

```
##
## Call:
## lm(formula = Profit ~ Online + AgeAverage)
##
## Residuals:
##   Min     1Q Median     3Q    Max 
## -398.01 -144.99 -91.28  55.00 1981.04 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  4.911     4.776   1.028   0.304    
## Online       22.005     4.699   4.683 2.84e-06 ***
## AgeAverage   25.682     1.090  23.572 < 2e-16 ***
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 270.5 on 31631 degrees of freedom
## Multiple R-squared:  0.01731, Adjusted R-squared:  0.01725 
## F-statistic: 278.6 on 2 and 31631 DF, p-value: < 2.2e-16
```

Ovviamente osserviamo che i valori dei coefficienti cambiano, ed in particolare cambia il contributo della variabile Online, quindi a seconda di come tappo il buco, di come scelgo di inserire i valori mancanti nel modello, otteniamo un risultato diverso. Questo ci suggerisce che probabilmente, non è questa la strada giusta per procedere.

```
mod = lm(Profit ~ Online+AgeZero+AgeGiven)
summary(mod)
```

```
##
## Call:
## lm(formula = Profit ~ Online + AgeZero + AgeGiven)
##
## Residuals:
##   Min     1Q Median     3Q    Max 
## -409.34 -144.14 -82.69  52.07 1967.12 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  70.926     3.000  23.643 < 2e-16 ***
## 
```

```

## Online      19.649     4.685    4.194 2.75e-05 ***
## AgeZero    25.603     1.086   23.582 < 2e-16 ***
## AgeGiven   -51.849     5.598   -9.263 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 269.5 on 31630 degrees of freedom
## Multiple R-squared:  0.02426, Adjusted R-squared:  0.02417
## F-statistic: 262.1 on 3 and 31630 DF, p-value: < 2.2e-16

```

Per tentare di evitare questo problema potremmo considerare i modelli ottenuti tenendo conto sia della variabile Age opportunamente modificata, sia del fatto che la variabile Age venga fornita oppure no. Questo ci permette di considerare una variabile categorica che indica la presenza o meno dell'informazione sull'età, quindi non importa più di tanto come vado a rimepire il buco dell'informazione mancante.

```

mod = lm(Profit ~ Online+AgeAverage+AgeGiven)
summary(mod)

```

```

##
## Call:
## lm(formula = Profit ~ Online + AgeAverage + AgeGiven)
##
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -409.34 -144.14  -82.69   52.07 1967.12 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -32.663     5.377  -6.074 1.26e-09 ***
## Online       19.649     4.685    4.194 2.75e-05 ***
## AgeAverage   25.603     1.086   23.582 < 2e-16 ***
## AgeGiven     51.740     3.448   15.006 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 269.5 on 31630 degrees of freedom
## Multiple R-squared:  0.02426, Adjusted R-squared:  0.02417
## F-statistic: 262.1 on 3 and 31630 DF, p-value: < 2.2e-16

```

Osserviamo che in entrambi i casi, tutte le variabili risultano significative perché il p-value corrispondente è sufficientemente piccolo. Inoltre il valore della variabile Online è lo stesso del caso precedente anche se abbiamo scelto di tappare i buchi scegliendo dei valori arbitrari. Tuttavia il valore dell' R^2 risulta essere molto molto piccolo, quindi questo modello povero arriva a spiegare circa il 2.5% della variabilità.

Per cercare un modello più valido, possiamo considerare un'altra variabile, ad esempio il reddito di ciascun cliente. Anche in questo caso possiamo scegliere come includere i valori NA, scegliamo ad esempio i valori nulli.

```

IncomeZero = ifelse(is.na(Income), 0, Income)
IncomeGiven = ifelse(is.na(Income), 0, 1)
mod = lm(Profit ~ Online+AgeAverage+AgeGiven+IncomeZero+IncomeGiven)
summary(mod)

```

```

##
## Call:
## lm(formula = Profit ~ Online + AgeAverage + AgeGiven + IncomeZero +
##     IncomeGiven)

```

```

##
## Residuals:
##      Min     1Q Median     3Q    Max
## -459.03 -144.61  -74.61   50.17 1963.39
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -38.191    5.336  -7.158 8.38e-13 ***
## Online       11.867    4.650   2.552   0.0107 *
## AgeAverage   26.891    1.078  24.941 < 2e-16 ***
## AgeGiven     14.490    8.272   1.752   0.0799 .
## IncomeZero   18.771    0.748  25.094 < 2e-16 ***
## IncomeGiven  -63.553    9.047  -7.024 2.20e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 266.8 on 31628 degrees of freedom
## Multiple R-squared:  0.04365, Adjusted R-squared:  0.0435
## F-statistic: 288.7 on 5 and 31628 DF, p-value: < 2.2e-16

```

Possiamo notare, che il valore della variabile Online scende un po'e che l' R^2 è quasi raddoppiato, ma rimane comunque piccolo.

Notiamo che non ci sono valori mancanti nella variabile District, tuttavia viene considerata numerica, quindi vorremmo trasformarla in categorica, in modo che ogni valore venga associato ad uno dei tre distretti. Quello che possiamo fare è introdurre due variabili binarie che rappresentano i distretti. Notiamo inoltre che non ci sono valori mancanti neanche nella variabile Tenure.

```

# Control for Tenure and district
any(is.na(District))

## [1] FALSE
table(District)

## District
## 1100 1200 1300
## 3142 24342 4150

District1100 = ifelse(District==1100,1,0)
District1200 = ifelse(District==1200,1,0)
any(is.na(Tenure))

```

```
## [1] FALSE
```

Andando ad effettuare la regressione con tutte queste variabili, notiamo che l' R^2 (e anche l' R^2_{adj}) sta aumentando anche se sempre di poco, mentre il contributo della variabile Online adesso vale circa 13.

```

mod = lm(Profit ~ Online+AgeAverage+AgeGiven+IncomeZero+IncomeGiven+Tenure+District1100+District1200)
summary(mod)

```

```

##
## Call:
## lm(formula = Profit ~ Online + AgeAverage + AgeGiven + IncomeZero +
##     IncomeGiven + Tenure + District1100 + District1200)
##
## Residuals:
##      Min     1Q Median     3Q    Max
## -487.17 -141.21  -65.88   48.87 1993.27

```

```

## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -44.2382   6.5180  -6.787 1.16e-11 *** 
## Online       13.8233   4.6091   2.999  0.00271 **  
## AgeAverage   16.6701   1.1482  14.519 < 2e-16 *** 
## AgeGiven     4.3913    8.2017   0.535  0.59237    
## IncomeZero   16.8530   0.7554  22.310 < 2e-16 *** 
## IncomeGiven  -57.1191   8.9956  -6.350 2.19e-10 *** 
## Tenure        4.7464    0.1918  24.742 < 2e-16 *** 
## District1100 -7.9955   6.2582  -1.278  0.20140    
## District1200 13.1986   4.4734   2.950  0.00318 ** 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 264.2 on 31625 degrees of freedom
## Multiple R-squared:  0.06234, Adjusted R-squared:  0.0621 
## F-statistic: 262.8 on 8 and 31625 DF, p-value: < 2.2e-16

#potremmo considerare Age come categorico

AgeCat = ifelse(is.na(Age)==TRUE,0,Age)
Age1=as.factor(AgeCat)
levels(Age1)

## [1] "0" "1" "2" "3" "4" "5" "6" "7"

mod = lm(Profit ~ Online+Age1+IncomeZero+IncomeGiven+Tenure+District1100+District1200)
summary(mod)

## 
## Call:
## lm(formula = Profit ~ Online + Age1 + IncomeZero + IncomeGiven +
##     Tenure + District1100 + District1200)
## 
## Residuals:
##      Min      1Q      Median      3Q      Max      
## -498.67 -141.61   -65.93    48.75 1993.55 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 22.8351   5.0941   4.483 7.40e-06 *** 
## Online       13.8916   4.6093   3.014 0.002582 **  
## Age11       -63.1264  12.3213  -5.123 3.02e-07 *** 
## Age12       -34.6976   9.0630  -3.829 0.000129 *** 
## Age13        1.5890   8.8141   0.180 0.856937    
## Age14        4.4084   8.9231   0.494 0.621280    
## Age15        7.3930   9.4162   0.785 0.432382    
## Age16       26.5698   9.8813   2.689 0.007173 ** 
## Age17       64.8400   9.6484   6.720 1.84e-11 *** 
## IncomeZero  16.7425   0.7762  21.570 < 2e-16 *** 
## IncomeGiven -57.4804  9.0048  -6.383 1.76e-10 *** 
## Tenure       4.7925    0.1920  24.965 < 2e-16 *** 
## District1100 -7.9082   6.2555  -1.264 0.206171    
## District1200 13.2550   4.4722   2.964 0.003040 ** 
## --- 

```

```

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 264.1 on 31620 degrees of freedom
## Multiple R-squared:  0.06342,   Adjusted R-squared:  0.06304
## F-statistic: 164.7 on 13 and 31620 DF,  p-value: < 2.2e-16

#potremmo considerare il rapporto tra Online e Age

```

PARTE B

In conclusione, i modelli trovati sono poco utili nella previsione del profitto. Possiamo provare a migliorare la situazione, utilizzando la storia della banca, e quindi anche le informazioni passate dei clienti.

```

#Andiamo a rinominare le variabili
Profit9=X9Profit
Online9=X9Online
Profit0=X0Profit
Online0=X0Online

```

Modello base

Proviamo a prevedere il profitto di ciascun cliente nel 2000, andando ad utilizzare l'informazione binaria riguardo all'uso dell'online banking. L' R^2 è troppo basso, quindi vorremmo migliorare questo modello.

```

mod1 = lm(Profit0 ~ Online9)
summary(mod1)

##
## Call:
## lm(formula = Profit0 ~ Online9)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -5784.9  -172.9  -120.9    62.1 26944.1 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 141.900    2.565  55.319 < 2e-16 ***
## Online9      23.490    7.267   3.232  0.00123 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 389.9 on 26394 degrees of freedom
##   (5238 observations deleted due to missingness)
## Multiple R-squared:  0.0003957, Adjusted R-squared:  0.0003578 
## F-statistic: 10.45 on 1 and 26394 DF,  p-value: 0.001229

```

Per provare a far aumentare il valore dell' R^2 , possiamo utilizzare le informazioni anagrafiche a nostra disposizione relative al singolo cliente, quelle che abbiamo utilizzato nei modelli precedenti.

```

mod2 = lm(Profit0 ~ Online9+AgeZero+AgeGiven+IncomeZero+IncomeGiven+Tenure+District1100+District1200)
summary(mod2)

##
## Call:
## lm(formula = Profit0 ~ Online9 + AgeZero + AgeGiven + IncomeZero +
## 
```

```

##      IncomeGiven + Tenure + District1100 + District1200)
##
## Residuals:
##   Min     1Q  Median     3Q    Max
## -5933.2 -168.7  -85.0    56.8 26797.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 58.7253   8.8328   6.649 3.02e-11 ***
## Online9     28.5864   7.2548   3.940 8.16e-05 ***
## AgeZero     13.4625   1.7582   7.657 1.97e-14 ***
## AgeGiven    -57.8432  14.3949  -4.018 5.88e-05 ***
## IncomeZero   21.5758   1.1483  18.789 < 2e-16 ***
## IncomeGiven -85.7359  14.0981  -6.081 1.21e-09 ***
## Tenure       4.7547   0.3007  15.809 < 2e-16 ***
## District1100 -13.5127 10.0338  -1.347   0.178
## District1200 10.9915   7.1485   1.538   0.124
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 383.3 on 26387 degrees of freedom
##   (5238 observations deleted due to missingness)
## Multiple R-squared:  0.03419,   Adjusted R-squared:  0.0339
## F-statistic: 116.8 on 8 and 26387 DF,  p-value: < 2.2e-16

```

Questo migliora abbastanza le prestazioni, perché così riusciamo a spiegare molta più della variabilità presente. Inoltre, abbiamo un'ulteriore informazione sui singoli clienti: il profitto nel 1999. Inserendo questa informazione nel modello è come se dicesse che il profitto nell'anno successivo dipende dal profitto dell'anno in corso e dalle caratteristiche del singolo cliente. Quindi in linea di principio, coloro che l'anno precedente mi hanno fatto guadagnare, continueranno a farmi guadagnare anche nell'anno successivo.

```
mod3 = lm(Profit0 ~ Profit9+Online9+AgeZero+AgeGiven+IncomeZero+IncomeGiven+Tenure+District1100+District1200)
summary(mod3)
```

```

##
## Call:
## lm(formula = Profit0 ~ Profit9 + Online9 + AgeZero + AgeGiven +
##     IncomeZero + IncomeGiven + Tenure + District1100 + District1200)
##
## Residuals:
##   Min     1Q  Median     3Q    Max
## -6948.0 -72.6  -33.2    28.6 26901.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.070248   7.185594   4.881 1.06e-06 ***
## Profit9      0.825353   0.007099  116.264 < 2e-16 ***
## Online9     15.063749   5.900660   2.553 0.010689 *
## AgeZero     -0.783422   1.434981  -0.546 0.585108
## AgeGiven    -2.298837  11.715494  -0.196 0.844438
## IncomeZero   7.123508   0.942052   7.562 4.11e-14 ***
## IncomeGiven -32.355443  11.473583  -2.820 0.004806 **
## Tenure       0.922225   0.246777   3.737 0.000187 ***
## District1100 -8.012553   8.159471  -0.982 0.326112
## District1200 -1.517990   5.814085  -0.261 0.794026

```

```

## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 311.7 on 26386 degrees of freedom
##   (5238 observations deleted due to missingness)
## Multiple R-squared: 0.3614, Adjusted R-squared: 0.3611
## F-statistic: 1659 on 9 and 26386 DF, p-value: < 2.2e-16

```

Adesso l' R^2 risulta notevolmente migliorato: riusciamo a spiegare oltre il 36% della variabilità.

Possiamo anche supporre che i dati anagrafici non siano rilevanti, ma magari non li abbiamo utilizzati nel modo giusto nel modello. Infatti provando a tenere la variabile del profitto precedente come regressore, ma eliminando le variabili `Age` e `Income`, otteniamo un modello che spiega all'incirca la stessa variabilità.

```

mod4 = lm(Profit0 ~ Profit9+Online9+Tenure+District1100+District1200)
summary(mod4)

```

```

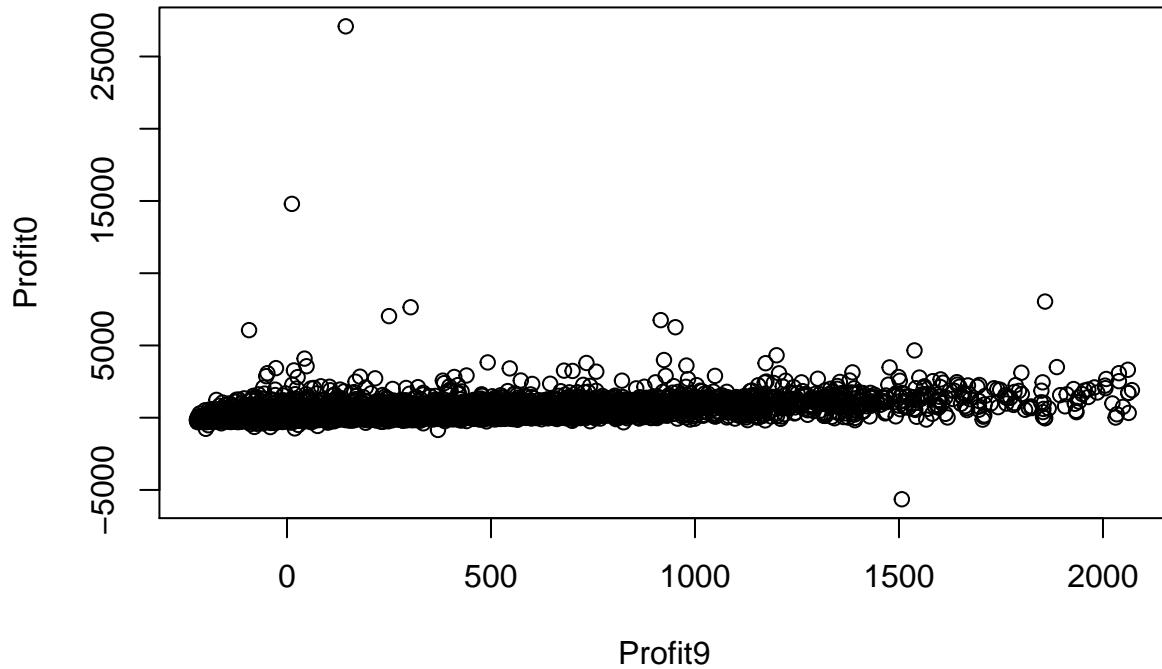
##
## Call:
## lm(formula = Profit0 ~ Profit9 + Online9 + Tenure + District1100 +
##     District1200)
##
## Residuals:
##    Min      1Q  Median      3Q      Max
## -6965.3   -70.3   -35.7    27.6  26908.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 31.023428  5.932798  5.229 1.72e-07 ***
## Profit9       0.831875  0.007017 118.547 < 2e-16 ***
## Online9      18.774205  5.841623  3.214  0.00131 **
## Tenure        0.919726  0.228984  4.017 5.92e-05 ***
## District1100 -11.205954  8.153961 -1.374  0.16936
## District1200  3.889211  5.776169  0.673  0.50075
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 312 on 26390 degrees of freedom
##   (5238 observations deleted due to missingness)
## Multiple R-squared: 0.3599, Adjusted R-squared: 0.3598
## F-statistic: 2968 on 5 and 26390 DF, p-value: < 2.2e-16

```

Profitto nel tempo

Quello che possiamo osservare è come le due variabili legate al profitto del singolo cliente nei diversi anni, siano correlate tra loro. Osserviamo infatti che c'è una correlazione di circa il 60%. Sono presenti inoltre solamente 8 osservazioni in cui il profitto nel 2000 risulta essere maggiore di 5000 e una osservazione con profitto nel 2000 minore di -5000, quindi possiamo pensare che quelle osservazioni siano degli outliers.

```
plot(Profit9,Profit0)
```



```
cor(Profit9,Profit0,use="complete.obs")
```

```
## [1] 0.5993369
```

```
sum(Profit0>5000,na.rm=TRUE)
```

```
## [1] 8
```

```
sum(Profit0<(-5000),na.rm=TRUE)
```

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

Quello che possiamo fare è andare ad osservare il valore dell' R^2 nel caso in cui questi valori vengano eliminati, per capire se il modello risulta migliore.

```
c=which(abs(Profit0)>5000)
detach(Pilgrim)
data=Pilgrim[-c,]
attach(data)
Profit9=X9Profit
Online9=X9Online
Age=X9Age
Income=X9Inc
Tenure=X9Tenure
District=X9District
Profit0=X0Profit
Online0=X0Online
District1100 = ifelse(District==1100,1,0)
District1200 = ifelse(District==1200,1,0)
```

```

AgeGiven = ifelse(is.na(Age),0,1)
AgeZero = ifelse(is.na(Age),0,1)
IncomeZero = ifelse(is.na(Income),0,Income)
IncomeGiven = ifelse(is.na(Income),0,1)
mod3 = lm(Profit0 ~ Profit9+Online9+AgeZero+AgeGiven+IncomeZero+IncomeGiven+Tenure+District1100+District1200)
summary(mod3)

##
## Call:
## lm(formula = Profit0 ~ Profit9 + Online9 + AgeZero + AgeGiven +
##     IncomeZero + IncomeGiven + Tenure + District1100 + District1200)
##
## Residuals:
##    Min      1Q  Median      3Q      Max
## -1711.1   -69.9   -31.1    30.8  4010.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.234524  5.178529  6.997 2.68e-12 ***
## Profit9       0.821781  0.005123 160.402 < 2e-16 ***
## Online9      18.161148  4.252329  4.271 1.95e-05 ***
## AgeZero      -1.760224  1.034234 -1.702 0.088775 .
## AgeGiven      2.183430  8.442798  0.259 0.795935
## IncomeZero    6.252322  0.678954  9.209 < 2e-16 ***
## IncomeGiven   -27.438630  8.268377 -3.319 0.000906 ***
## Tenure        0.917593  0.177904  5.158 2.52e-07 ***
## District1100 -13.050532  5.881041 -2.219 0.026489 *
## District1200  -6.504494  4.190384 -1.552 0.120616
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 224.6 on 26377 degrees of freedom
## (5238 observations deleted due to missingness)
## Multiple R-squared:  0.5174, Adjusted R-squared:  0.5173
## F-statistic:  3143 on 9 and 26377 DF,  p-value: < 2.2e-16
mod4 = lm(Profit0 ~ Profit9+Online9+Tenure+District1100+District1200)
summary(mod4)

##
## Call:
## lm(formula = Profit0 ~ Profit9 + Online9 + Tenure + District1100 +
##     District1200)
##
## Residuals:
##    Min      1Q  Median      3Q      Max
## -1712.1   -67.9   -33.5    30.1  4009.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.956128  4.278295  7.937 2.16e-15 ***
## Profit9       0.827164  0.005068 163.228 < 2e-16 ***
## Online9      22.032087  4.212378  5.230 1.71e-07 ***
## Tenure        0.852151  0.165189  5.159 2.51e-07 ***

```

```

## District1100 -16.081030  5.880673 -2.735  0.00625 **
## District1200  -1.713170  4.165725 -0.411  0.68089
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 225 on 26381 degrees of freedom
##   (5238 observations deleted due to missingness)
## Multiple R-squared:  0.5158, Adjusted R-squared:  0.5157
## F-statistic:  5620 on 5 and 26381 DF,  p-value: < 2.2e-16

```

Quindi andando a ripetere il codice precedente, quando però eliminiamo gli outliers, miglioriamo l' R^2 che negli ultimi due modelli raggiunge circa il 50%. Per questo motivo nelle successive analisi considereremo il dataset senza outliers.

Osserviamo inoltre che abbiamo un numero diverso di valori mancanti: ci sono 5219 osservazioni in cui non abbiamo l'informazione sulla online banking nel 2000 e 5238 osservazioni in cui non abbiamo l'informazione sul profitto nel 2000, questo vuol dire che il cliente non è rimasto nella banca nell'anno successivo. Quindi logicamente, non ci sono osservazioni in cui abbiamo l'informazione di Profit0 ma non di Online0. Ci sono invece 19 osservazioni in cui al contrario abbiamo l'informazione sul profitto, ma non sappiamo se il cliente ha utilizzato la banca online.

```

sum(is.na(Online0))

## [1] 5219

sum(is.na(P0))

## [1] 5238

which(is.na(P0) != is.na(Online0))

## [1] 2309 4127 4190 5129 5681 8610 9308 11028 11778 12735 18731
## [12] 21060 23432 25439 26308 27972 30210 30952 31535

which(is.na(P0) < is.na(Online0))

## integer(0)

which(is.na(P0) > is.na(Online0))

## [1] 2309 4127 4190 5129 5681 8610 9308 11028 11778 12735 18731
## [12] 21060 23432 25439 26308 27972 30210 30952 31535

```

ADD retain variable

Quello che possiamo fare è aggiungere una variabile aggiuntiva `Retain` che ci dice se il cliente rimane nella banca anche nell'anno successivo. A questo punto vorremmo capire come prevedere al meglio se un cliente rimane con la banca nell'anno successivo oppure no, questo potrebbe dipendere da tutte le caratteristiche del cliente.

```

Retain = ifelse(is.na(P0), 0, 1)
mod1 = lm(Retain ~ Online9)
summary(mod1)

##
## Call:
## lm(formula = Retain ~ Online9)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.5000 -0.2500  0.0000  0.2500  0.5000
## 
```

```

## -0.8534  0.1466  0.1683  0.1683  0.1683
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.831731   0.002230 372.904 < 2e-16 ***
## Online9      0.021668   0.006389   3.391 0.000696 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3717 on 31623 degrees of freedom
## Multiple R-squared:  0.0003636, Adjusted R-squared:  0.000332
## F-statistic:  11.5 on 1 and 31623 DF, p-value: 0.0006963
mod2 = lm(Retain ~ Online9+AgeZero+AgeGiven+IncomeZero+IncomeGiven+Tenure+District1100+District1200)
summary(mod2)

##
## Call:
## lm(formula = Retain ~ Online9 + AgeZero + AgeGiven + IncomeZero +
##     IncomeGiven + Tenure + District1100 + District1200)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04745  0.02557  0.07969  0.10979  0.46685
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.5356312  0.0065091 82.289 < 2e-16 ***
## Online9      0.0107948  0.0058871  1.834 0.066719 .
## AgeZero      0.0014012  0.0014667  0.955 0.339418
## AgeGiven     0.1664643  0.0117311 14.190 < 2e-16 ***
## IncomeZero   0.0035132  0.0009650  3.641 0.000272 ***
## IncomeGiven  0.1501841  0.0114900 13.071 < 2e-16 ***
## Tenure        0.0039240  0.0002451 16.009 < 2e-16 ***
## District1100 -0.0031135  0.0079946 -0.389 0.696949
## District1200  0.0074261  0.0057144  1.300 0.193766
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3375 on 31616 degrees of freedom
## Multiple R-squared:  0.176, Adjusted R-squared:  0.1758
## F-statistic: 844.3 on 8 and 31616 DF, p-value: < 2.2e-16
mod3 = lm(Retain ~ Profit9+Online9+AgeZero+AgeGiven+IncomeZero+IncomeGiven+Tenure+District1100+District1200)
summary(mod3)

##
## Call:
## lm(formula = Retain ~ Profit9 + Online9 + AgeZero + AgeGiven +
##     IncomeZero + IncomeGiven + Tenure + District1100 + District1200)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.05798  0.02531  0.07933  0.11024  0.46756
##

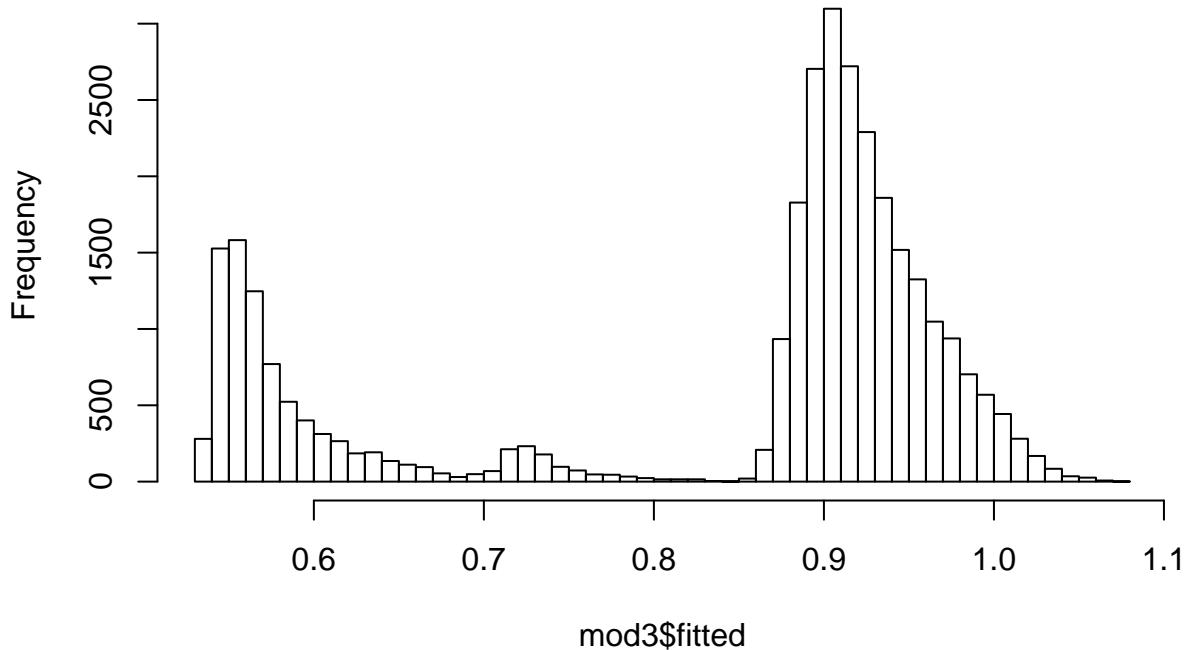
```

```

## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.352e-01  6.511e-03 82.203 < 2e-16 ***
## Profit9         1.895e-05  7.191e-06  2.635  0.00842 **
## Online9          1.053e-02  5.887e-03  1.789  0.07369 .
## AgeZero          1.085e-03  1.471e-03  0.738  0.46073
## AgeGiven         1.677e-01  1.174e-02 14.283 < 2e-16 ***
## IncomeZero       3.195e-03  9.724e-04  3.285  0.00102 **
## IncomeGiven      1.513e-01  1.150e-02 13.158 < 2e-16 ***
## Tenure           3.834e-03  2.474e-04 15.496 < 2e-16 ***
## District1100   -2.966e-03  7.994e-03 -0.371  0.71061
## District1200    7.172e-03  5.715e-03  1.255  0.20947
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3375 on 31615 degrees of freedom
## Multiple R-squared:  0.1762, Adjusted R-squared:  0.176
## F-statistic: 751.4 on 9 and 31615 DF,  p-value: < 2.2e-16
hist(mod3$fitted, nclass=50, main ="Histogram of mod3")

```

Histogram of mod3



Andare a considerare il valore dell' R^2 per valutare la bontà del modello non è molto opportuno perché stiamo prevedendo se il cliente rimane oppure no, quindi è una variabile binaria. Dall'istogramma possiamo facilmente dedurre ci sono due gruppi abbastanza distinti, quindi quello che possiamo fare è fissare una certa soglia al di sopra della quale possiamo concludere che il cliente è sicuro, mentre al di sotto si trovano quei clienti che probabilmente il prossimo anno cambieranno la banca. Infatti ci sono due mode molto alte e sufficientemente lontane, mentre tra i valori di 0.7 e 0.8, troviamo una "zona grigia", dove la moda è molto

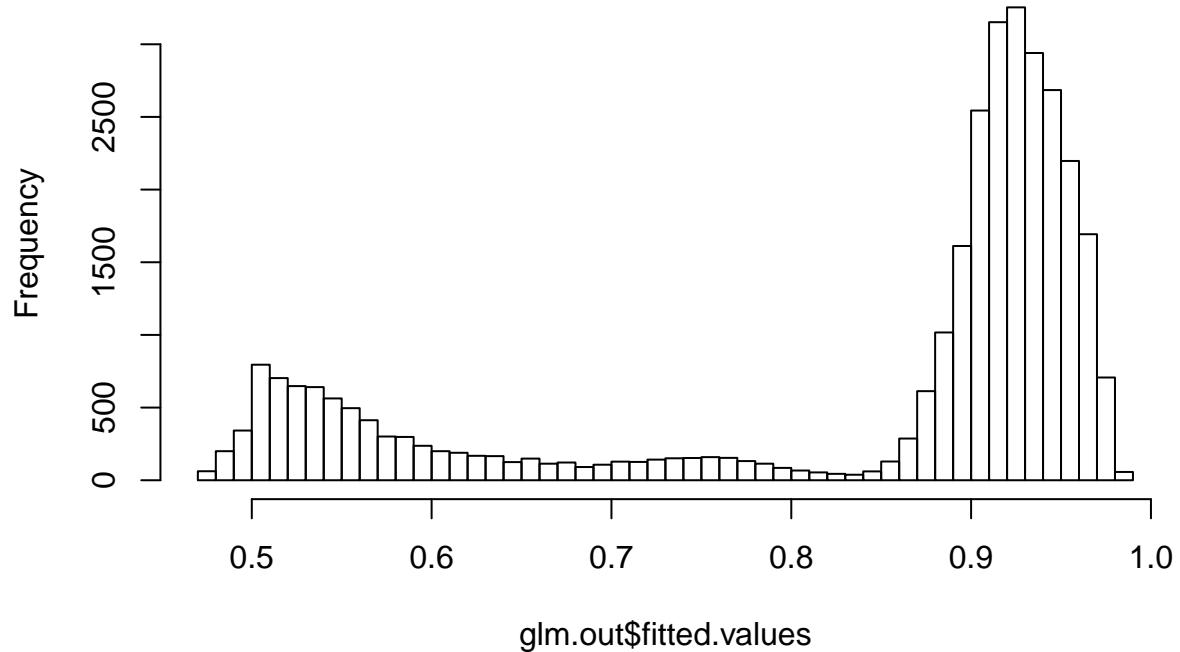
bassa e i clienti sono abbastanza incerti. Inoltre saremmo interessati ad interpretare i valori dell'istogramma come delle probabilità, anche se ci sono dei valori più grandi di 1. In questo modo sarei in grado di capire sotto quale valore dovrei preoccuparmi del cliente.

#— PART 3 - ANALYZE retention with Logistic regression Proviamo a prevedere il comportamento futuro di un cliente, utilizzando la regressione logistica.

```
glm.out = glm(Retain ~ Profit9+Online9+AgeZero+AgeGiven+IncomeZero+
  IncomeGiven+Tenure+District1100+District1200,family=binomial(logit))
summary(glm.out)

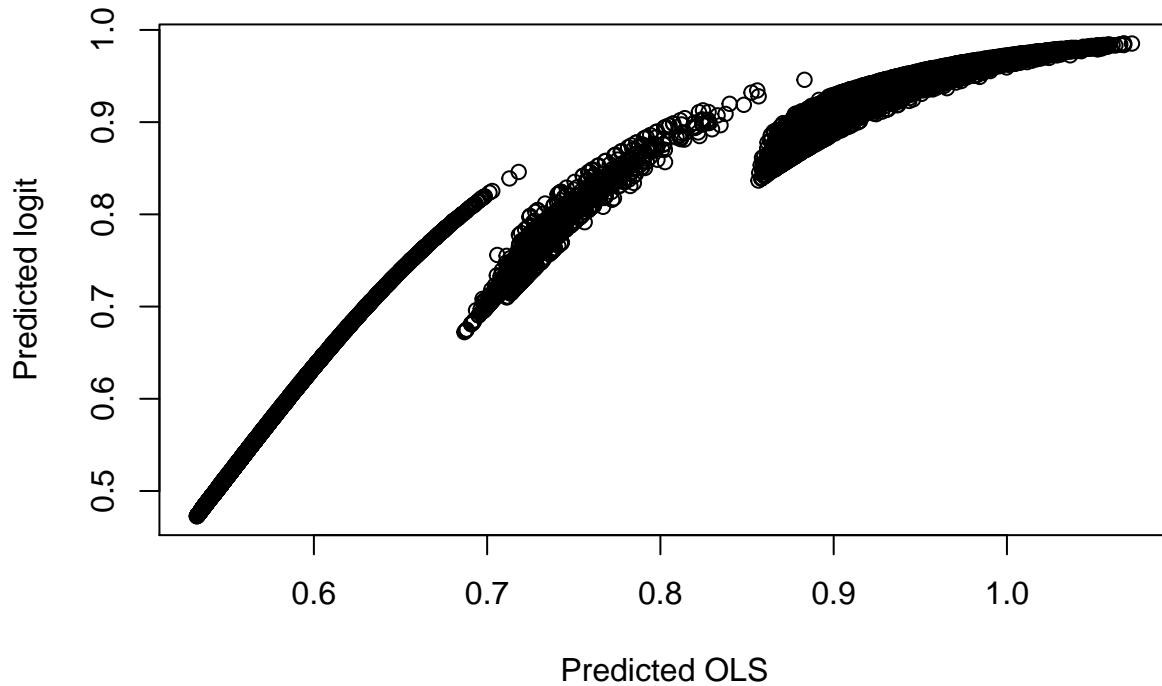
##
## Call:
## glm(formula = Retain ~ Profit9 + Online9 + AgeZero + AgeGiven +
##       IncomeZero + IncomeGiven + Tenure + District1100 + District1200,
##       family = binomial(logit))
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max 
## -2.8377  0.2917  0.3884  0.4667  1.2243 
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -7.981e-02 5.145e-02 -1.551   0.1209    
## Profit9      1.832e-04 7.182e-05  2.550   0.0108 *  
## Online9      1.058e-01 5.333e-02  1.984   0.0472 *  
## AgeZero      6.841e-02 1.596e-02  4.287  1.81e-05 *** 
## AgeGiven     8.372e-01 9.307e-02  8.996 < 2e-16 *** 
## IncomeZero   5.320e-02 1.058e-02  5.030  4.89e-07 *** 
## IncomeGiven  7.752e-01 8.999e-02  8.615 < 2e-16 *** 
## Tenure       3.748e-02 2.413e-03 15.530 < 2e-16 *** 
## District1100 -3.193e-02 6.798e-02 -0.470   0.6386    
## District1200  6.678e-02 4.915e-02  1.359   0.1742    
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 28392  on 31624  degrees of freedom
## Residual deviance: 23297  on 31615  degrees of freedom
## AIC: 23317
##
## Number of Fisher Scoring iterations: 5
hist(glm.out$fitted.values,nclass=50, main="Histogram of Logistic Regression")
```

Histogram of Logistic Regression



Questo metodo risulta molto veloce,in quanto si ferma dopo solo 5 iterazioni. Inoltre le diagnostiche del modello sono un po' diverse, notiamo che non abbiamo più a disposizione l' R^2 ,ma non cambia il senso generale, come osserviamo dall'istogramma.

```
plot(mod3$fitted, glm.out$fitted, xlab="Predicted OLS", ylab="Predicted logit")
```



```
# Regressione logistica andando ad eliminare gli outliers
glm.out = glm(Retain ~ Profit9+Online9+AgeZero+AgeGiven+IncomeZero+
  IncomeGiven+Tenure+District1100+District1200,family=binomial(logit))
summary(glm.out)
```

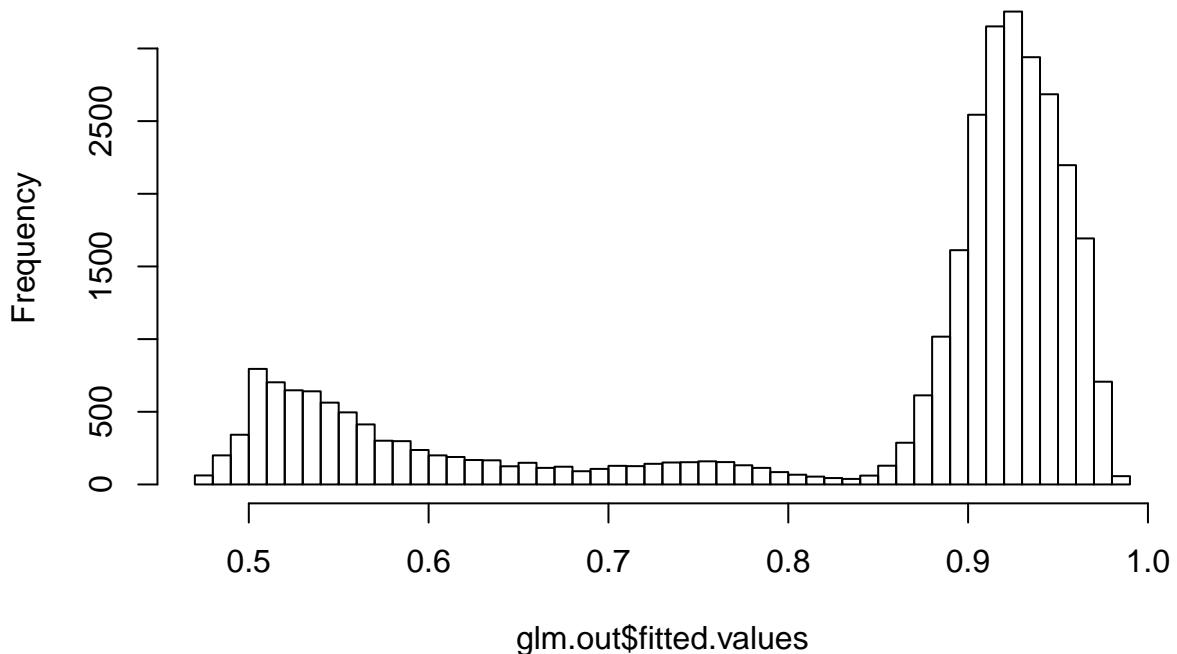
```
##
## Call:
## glm(formula = Retain ~ Profit9 + Online9 + AgeZero + AgeGiven +
##       IncomeZero + IncomeGiven + Tenure + District1100 + District1200,
##       family = binomial(logit))
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -2.8377    0.2917    0.3884    0.4667   1.2243
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.981e-02  5.145e-02 -1.551  0.1209
## Profit9      1.832e-04  7.182e-05  2.550  0.0108 *
## Online9      1.058e-01  5.333e-02  1.984  0.0472 *
## AgeZero      6.841e-02  1.596e-02  4.287 1.81e-05 ***
## AgeGiven     8.372e-01  9.307e-02  8.996 < 2e-16 ***
## IncomeZero   5.320e-02  1.058e-02  5.030 4.89e-07 ***
## IncomeGiven  7.752e-01  8.999e-02  8.615 < 2e-16 ***
## Tenure       3.748e-02  2.413e-03 15.530 < 2e-16 ***
## District1100 -3.193e-02  6.798e-02 -0.470  0.6386
```

```

## District1200  6.678e-02  4.915e-02   1.359   0.1742
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 28392  on 31624  degrees of freedom
## Residual deviance: 23297  on 31615  degrees of freedom
## AIC: 23317
##
## Number of Fisher Scoring iterations: 5
hist(glm.out$fitted.values,nclass=50, main="Histogram of Logistic Regression")

```

Histogram of Logistic Regression



Bilanciamento dataset

```

Retain = as.factor(Retain)
summary(Retain)

##      0      1
## 5238 26387

```

Possiamo notare che i dati sono sbilanciati, quindi dobbiamo andare a bilanciare i dati mediante un oversampling:

```

library(ROSE)

## Loaded ROSE 0.0-3
dim_data_balanced = max(sum(Retain==1),sum(Retain==0))*2
data_balanced_over <- ovun.sample(as.factor(Retain) ~ Profit9+Online9+AgeZero+AgeGiven+IncomeZero+Income

```

Per analizzare meglio i vari algoritmi di classificazione dividiamo in due parti il dataset bilanciato, una parte per il training e una parte per la validazione del modello.

```

set.seed(1)
train = sample(1:nrow(data_balanced_over), nrow(data_balanced_over)*0.7)
df.train = data_balanced_over[train,]
df.test = data_balanced_over[-train,]
target.train = df.train$Retain
target.test = df.test$Retain

```

Decision Tree

```

library(tree)
setup<-tree.control(nrow(df.train),mincut = 2, minsize = 6, mindev = 0.001)
tree.Pilgrim = tree(Retain ~ Profit9+Online9+AgeZero+AgeGiven+IncomeZero+IncomeGiven+Tenure+District1100
tree.pred = predict(tree.Pilgrim, df.test, type="class", probability=TRUE)
library(caret)

## Loading required package: lattice
## Loading required package: ggplot2
cf = confusionMatrix(data = tree.pred, reference = target.test, mode = "prec_recall")
cf

## Confusion Matrix and Statistics
##
##          Reference
## Prediction 1 0
##           1 6606 2873
##           0 1273 5081
##
##          Accuracy : 0.7381
##             95% CI : (0.7312, 0.745)
##    No Information Rate : 0.5024
##    P-Value [Acc > NIR] : < 2.2e-16
##
##          Kappa : 0.4768
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##          Precision : 0.6969
##          Recall : 0.8384
##          F1 : 0.7611
##          Prevalence : 0.4976
##    Detection Rate : 0.4172
##    Detection Prevalence : 0.5987
##    Balanced Accuracy : 0.7386
##

```

```

##      'Positive' Class : 1
##
```

Logistic Regression

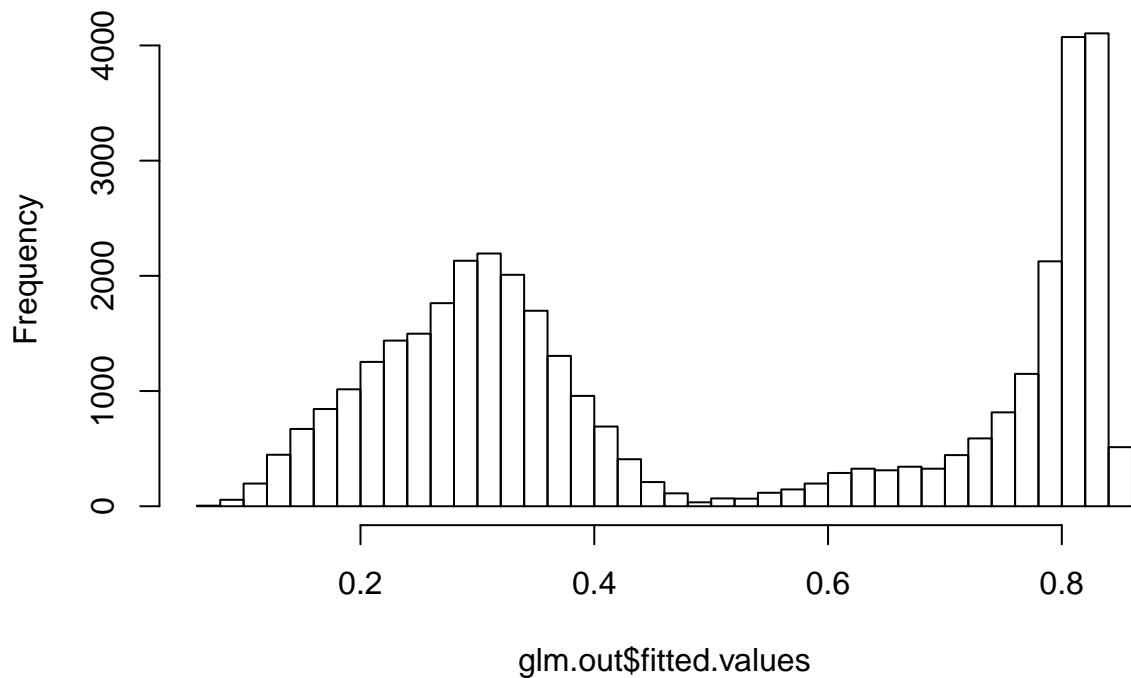
```

glm.out = glm(Retain ~ Profit9+Online9+AgeZero+AgeGiven+IncomeZero+
  IncomeGiven+Tenure+District1100+District1200,family=binomial(logit), data=df.train)
summary(glm.out)

##
## Call:
## glm(formula = Retain ~ Profit9 + Online9 + AgeZero + AgeGiven +
##       IncomeZero + IncomeGiven + Tenure + District1100 + District1200,
##       family = binomial(logit), data = df.train)
##
## Deviance Residuals:
##    Min      1Q   Median      3Q      Max
## -1.9623 -0.8565 -0.4466  0.7157  2.1846
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.646e+00 3.890e-02 42.325 < 2e-16 ***
## Profit9     -1.709e-04 4.948e-05 -3.454 0.000552 ***
## Online9     -1.456e-02 3.751e-02 -0.388 0.697850
## AgeZero     -6.681e-02 9.847e-03 -6.785 1.16e-11 ***
## AgeGiven    -8.429e-01 7.006e-02 -12.031 < 2e-16 ***
## IncomeZero   -5.580e-02 6.642e-03 -8.401 < 2e-16 ***
## IncomeGiven  -7.540e-01 6.807e-02 -11.077 < 2e-16 ***
## Tenure       -3.657e-02 1.651e-03 -22.151 < 2e-16 ***
## District1100 1.236e-01 4.957e-02  2.493 0.012663 *
## District1200 -5.624e-02 3.576e-02 -1.573 0.115817
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 51211  on 36940  degrees of freedom
## Residual deviance: 41028  on 36931  degrees of freedom
## AIC: 41048
##
## Number of Fisher Scoring iterations: 4
hist(glm.out$fitted.values,nclass=50, main="Histogram of Logistic Regression")

```

Histogram of Logistic Regression



```
pred=predict(glm.out, df.test, type="response")
logit.pred = rep(0, dim(df.test)[1])
logit.pred[pred > .5] = 1
logit.pred=factor(logit.pred,labels=c("1","0"))
cf = confusionMatrix(data = logit.pred, reference = target.test, mode = "prec_recall", positive="1")
cf

## Confusion Matrix and Statistics
##
##          Reference
## Prediction      1      0
##           1 6347 2666
##           0 1532 5288
##
##          Accuracy : 0.7349
##             95% CI : (0.7279, 0.7417)
##    No Information Rate : 0.5024
##    P-Value [Acc > NIR] : < 2.2e-16
##
##          Kappa : 0.4701
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##          Precision : 0.7042
##          Recall : 0.8056
##          F1 : 0.7515
##  Prevalence : 0.4976
```

```

##           Detection Rate : 0.4009
##     Detection Prevalence : 0.5693
##     Balanced Accuracy : 0.7352
##
##           'Positive' Class : 1
##
specificity=specificity(logit.pred,target.test, positive="1")
cat('Specificity:',specificity, '\n')

## Specificity: 0.6648227

```

KNN

```

set.seed(1)
train.control=trainControl(method="repeatedcv", number=3, repeats=1)
fit=train(Retain ~ Profit9+Online9+AgeZero+AgeGiven+IncomeZero+IncomeGiven+Tenure+District1100+District
          method="knn",
          tuneGrid=expand.grid(k=c(4,8,10,15)),
          trControl=train.control,
          metric="Accuracy",
          data=df.train)
fit

## k-Nearest Neighbors
##
## 36941 samples
##      9 predictor
##      2 classes: '1', '0'
##
## No pre-processing
## Resampling: Cross-Validated (3 fold, repeated 1 times)
## Summary of sample sizes: 24627, 24628, 24627
## Resampling results across tuning parameters:
##
##     k     Accuracy   Kappa
##     4    0.7439701  0.4881729
##     8    0.7311117  0.4623654
##    10   0.7285130  0.4571430
##    15   0.7300559  0.4601636
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 4.

library(class)
knn_pred <- knn(df.train,df.test,cl=target.train,k=4)
cf = confusionMatrix(data = knn_pred, reference = target.test, mode = "prec_recall")
cf

## Confusion Matrix and Statistics
##
##           Reference
## Prediction    1     0
##           1 5880  513
##           0 1999 7441

```

```

##                               Accuracy : 0.8413
##                               95% CI : (0.8356, 0.847)
## No Information Rate : 0.5024
## P-Value [Acc > NIR] : < 2.2e-16
##
##                               Kappa : 0.6824
##
## Mcnemar's Test P-Value : < 2.2e-16
##
##                               Precision : 0.9198
##                               Recall : 0.7463
##                               F1 : 0.8240
##                               Prevalence : 0.4976
## Detection Rate : 0.3714
## Detection Prevalence : 0.4038
## Balanced Accuracy : 0.8409
##
## 'Positive' Class : 1
##
specificity=specificity(knn_pred,target.test, positive="1")
cat('Specificity:',specificity, '\n')

## Specificity: 0.9355041

```

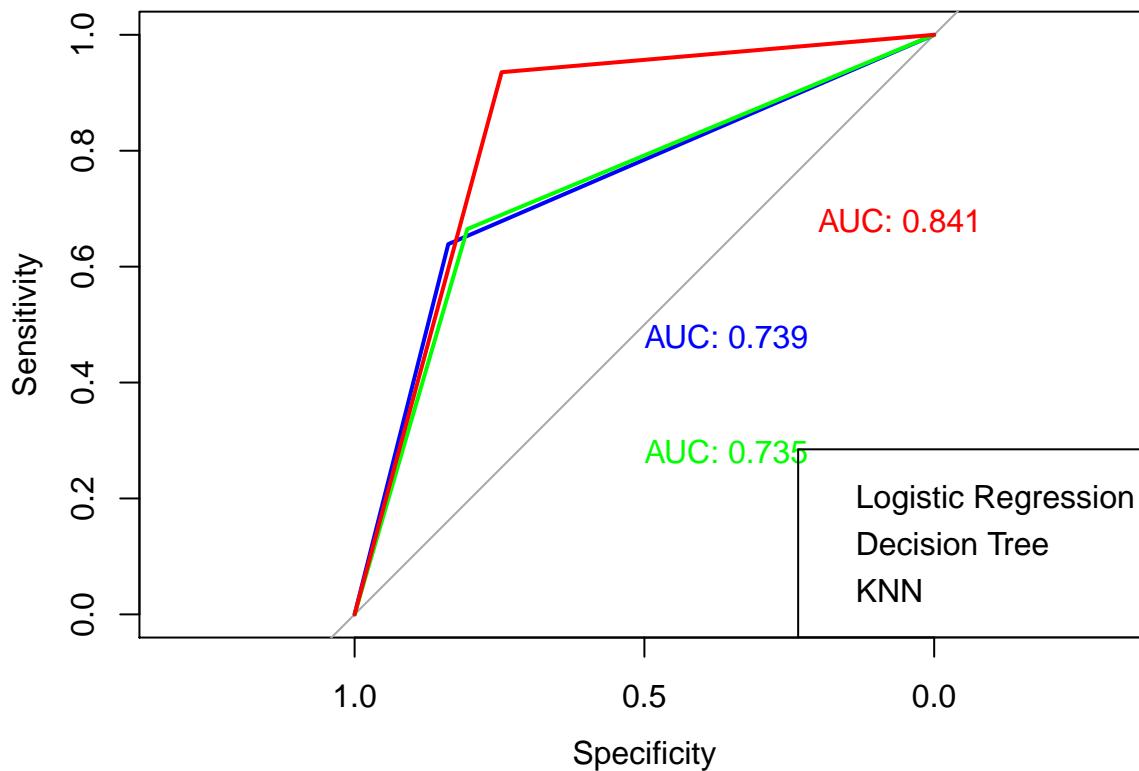
Confronto tra modelli

```

library(pROC)
treeROC = roc(target.test,factor(tree.pred, ordered = TRUE), plot=TRUE, print.auc=TRUE, col="blue")
logit_roc=plot.roc(target.test,factor(logit.pred, ordered = TRUE), add=TRUE, col="green", print.auc=TRUE)
knn_roc = plot.roc(target.test,factor(knn_pred, ordered = TRUE), add=TRUE, col="red", print.auc=TRUE, p

legend("bottomright", legend = c("Logistic Regression", "Decision Tree", "KNN"), col=c("green", "blue",

```



— PART 4 - Demographics vs. Past profit to analyze profitability Possiamo considerare le variabili Age e Income effettivamente come categoriche.

```
mod1 = lm(Profit0 ~
  Profit9+Online9+Tenure+District1100+District1200+factor(Age)+factor(Income))
summary(mod1)
```

```
##
## Call:
## lm(formula = Profit0 ~ Profit9 + Online9 + Tenure + District1100 +
##     District1200 + factor(Age) + factor(Income))
##
## Residuals:
##    Min      1Q  Median      3Q     Max 
## -1712.8   -72.8   -29.9    35.6  3442.5 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 20.773203  11.557292   1.797 0.072285 .
## Profit9       0.826567   0.005612 147.297 < 2e-16 ***
## Online9      10.383005   4.702057   2.208 0.027242 *  
## Tenure        0.717158   0.200863   3.570 0.000357 *** 
## District1100 -13.690068   6.643184  -2.061 0.039337 *  
## District1200 -6.147888   4.716427  -1.304 0.192416  
## factor(Age)2  5.870936  10.745660   0.546 0.584828  
## factor(Age)3 17.685914  10.545089   1.677 0.093524 .  
## factor(Age)4  6.515502  10.611035   0.614 0.539202
```

```

## factor(Age)5      -6.109311 10.981095 -0.556 0.577979
## factor(Age)6      2.621036 11.354052  0.231 0.817437
## factor(Age)7      5.245526 11.215677  0.468 0.640007
## factor(Income)2   -17.195745 10.024557 -1.715 0.086294 .
## factor(Income)3   -0.235820  7.238542 -0.033 0.974011
## factor(Income)4    6.336258  7.356581  0.861 0.389080
## factor(Income)5   -3.419286  7.335712 -0.466 0.641138
## factor(Income)6   12.747035  6.426292  1.984 0.047316 *
## factor(Income)7   14.878733  7.018788  2.120 0.034031 *
## factor(Income)8   37.963230  8.016829  4.735 2.20e-06 ***
## factor(Income)9   46.868290  7.210886  6.500 8.23e-11 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 224.2 on 21055 degrees of freedom
##   (10550 observations deleted due to missingness)
## Multiple R-squared:  0.5337, Adjusted R-squared:  0.5333
## F-statistic:  1268 on 19 and 21055 DF,  p-value: < 2.2e-16
mod2 = lm(Profit0 ~ Profit9+Online9+Tenure)
summary(mod2)

##
## Call:
## lm(formula = Profit0 ~ Profit9 + Online9 + Tenure)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1711.9   -67.8   -33.8    30.5  4010.5 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 31.092530  2.330534 13.341 < 2e-16 ***
## Profit9      0.827704  0.005061 163.532 < 2e-16 ***
## Online9      22.499007  4.206967  5.348 8.97e-08 ***
## Tenure       0.838474  0.165117  5.078 3.84e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 225 on 26383 degrees of freedom
##   (5238 observations deleted due to missingness)
## Multiple R-squared:  0.5156, Adjusted R-squared:  0.5155
## F-statistic:  9360 on 3 and 26383 DF,  p-value: < 2.2e-16
mod3 = lm(Profit0 ~ District1100+District1200+factor(Age)+factor(Income))
summary(mod3)

##
## Call:
## lm(formula = Profit0 ~ District1100 + District1200 + factor(Age) +
##   factor(Income))
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1009.8  -173.8  -88.1    73.0  4384.2 

```

```

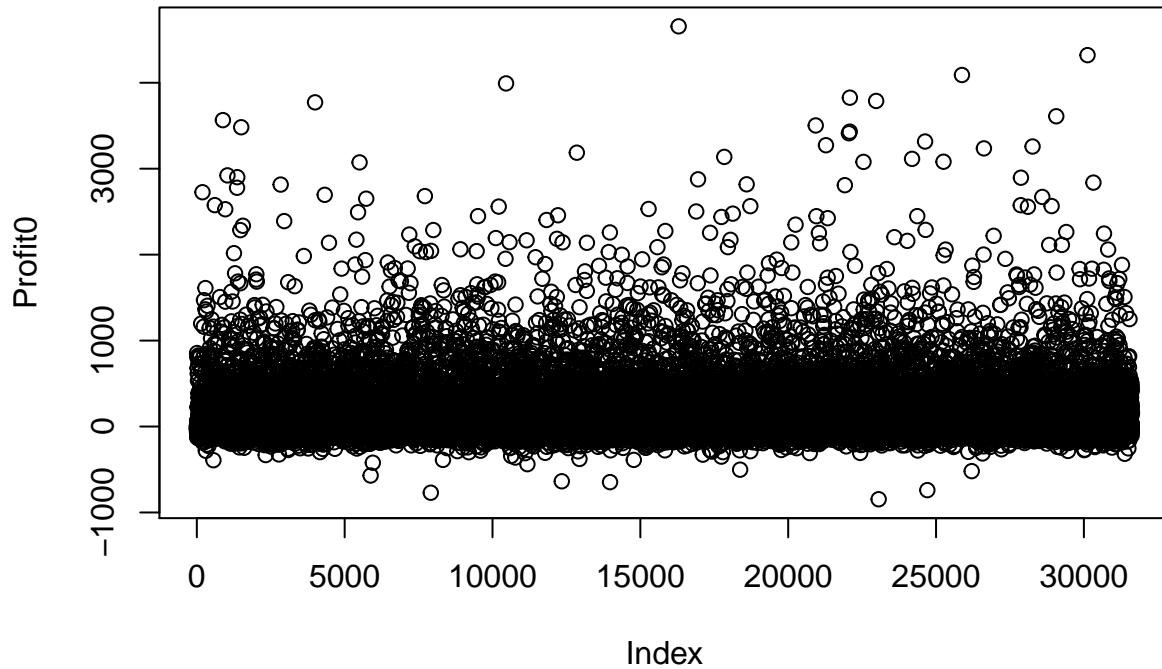
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 0.2372    16.4794   0.014   0.989    
## District1100 -19.4450   9.5145  -2.044   0.041 *  
## District1200  3.9288   6.7527   0.582   0.561    
## factor(Age)2  31.9241  15.3864   2.075   0.038 *  
## factor(Age)3  83.7490  15.0601   5.561  2.71e-08 *** 
## factor(Age)4  88.4328  15.0783   5.865  4.56e-09 *** 
## factor(Age)5  88.6467  15.5040   5.718  1.09e-08 *** 
## factor(Age)6 124.0141  15.9086   7.795  6.72e-15 *** 
## factor(Age)7 159.6297  15.6390  10.207 < 2e-16 *** 
## factor(Income)2 -9.1954  14.3563  -0.641   0.522    
## factor(Income)3 12.4778  10.3633   1.204   0.229    
## factor(Income)4 23.5950  10.5241   2.242   0.025 *  
## factor(Income)5 16.6441  10.4934   1.586   0.113    
## factor(Income)6 55.4366  9.1833   6.037  1.60e-09 *** 
## factor(Income)7 76.3413  10.0196   7.619  2.66e-14 *** 
## factor(Income)8 114.8142  11.4435  10.033 < 2e-16 *** 
## factor(Income)9 183.8698  10.2152  18.000 < 2e-16 *** 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 321.2 on 21058 degrees of freedom
##   (10550 observations deleted due to missingness)
## Multiple R-squared:  0.04327,   Adjusted R-squared:  0.04255 
## F-statistic: 59.53 on 16 and 21058 DF,  p-value: < 2.2e-16

summary(Profit0)

##      Min. 1st Qu. Median  Mean 3rd Qu.  Max.  NA's
## -846.0 -30.0   23.0 141.9 206.0 4656.0 5238

plot(Profit0)

```



```

m=which(Profit0<=-5000)
M=which(Profit0>=5000)
d=Pilgrim[-m,]
dtab=d[-M, ]
summary(dtab)

##          ID       X9Profit      X9Online      X9Age       X9Inc
##  Min.   : NA   Min.   : NA   Min.   : NA   Min.   : NA   Min.   : NA
##  1st Qu.: NA  1st Qu.: NA  1st Qu.: NA  1st Qu.: NA  1st Qu.: NA
##  Median : NA  Median : NA  Median : NA  Median : NA  Median : NA
##  Mean   :NaN  Mean   :NaN  Mean   :NaN  Mean   :NaN  Mean   :NaN
##  3rd Qu.: NA  3rd Qu.: NA  3rd Qu.: NA  3rd Qu.: NA  3rd Qu.: NA
##  Max.   : NA  Max.   : NA  Max.   : NA  Max.   : NA  Max.   : NA
##          X9Tenure     X9District    X0Profit      X0Online      X9Billpay
##  Min.   : NA   Min.   : NA   Min.   : NA   Min.   : NA   Min.   : NA
##  1st Qu.: NA  1st Qu.: NA  1st Qu.: NA  1st Qu.: NA  1st Qu.: NA
##  Median : NA  Median : NA  Median : NA  Median : NA  Median : NA
##  Mean   :NaN  Mean   :NaN  Mean   :NaN  Mean   :NaN  Mean   :NaN
##  3rd Qu.: NA  3rd Qu.: NA  3rd Qu.: NA  3rd Qu.: NA  3rd Qu.: NA
##  Max.   : NA  Max.   : NA  Max.   : NA  Max.   : NA  Max.   : NA
##          X0Billpay
##  Min.   : NA
##  1st Qu.: NA
##  Median : NA
##  Mean   :NaN
##  3rd Qu.: NA

```

```
## Max. : NA  
#detach(Pilgrim)
```