Alessia Poli, 922096, EMA
a. a. 2019 2020

ESERCIZIO IN R
MODELLO DI REGRESSIONE LOGISTICA E ALBERO DI CLASSIFICAZIONE: DATASET SocTelePag.csv
Problema da risolvere:
I dati contenuti nel dataset SocTelePag.csv riguardano le informazioni relative ad una societa` televisiva a pagamento in virtu` delle quali sviluppare strumenti, analisi e metodologie per lo studio del rischio e delle previsioni degli abbandoni aziendali.

I dati a disposizione:
Le variabili presenti nel dataset sono state estratte da diversi database in uso all'interno dell'azienda. In particolare, la variabile risposta comprende due differenti tipologie di clienti: coloro che al momento dell'osservazione sono attivi e quanti, invece, hanno disdetto l'abbonamento. Le altre variabili in input si riferiscono agli aspetti socio-demografici, comportamentali e contrattuali dei clienti. Le variabili piu` importanti ai fini dell'analisi sono le seguenti:

- PACCHETTO: tipologia del pacchetto acquistato;
- AREA NIELSEN: area geografica relativa alla residenza del cliente;
- FASCIA REDDITO: fascia di reddito;
- FASCIA ETA: fascia di età;
- ESG ATTIVAZIONI: chiamate call center per attivazioni (1: s`ı; 0: no);
- ESG PROBLEMI TECNICI: chiamate call center per problemi tecnici (1: s`ı; 0: no);
- ESG PROMOZIONI: chiamate call center per promozioni (1: s`ı; 0: no);
- ESG VARIAZIONI CONTRATTUALI: chiamate call center per variazioni del contratto (1: s`ı; 0: no);
- DURATA: durata del contratto in mesi;
- STATO: stato attivo (0: non abbandono), stato inattivo (1: abbandono);
- CANALE VENDITA SKY CENTER: canale di vendita Sky Center (1: s`ı; 0: no);
- CANALE VENDITA SKY SERVICE: canale di vendita Sky Service (1: s`ı; 0: no);
- CANALE VENDITA TELESELLING: canale di vendita Teleselling (1: s`ı; 0: no);
- MOP PO: modalità di pagamento con bollettino postale (1: s`ı; 0: no);
- MOP CC: modalità di pagamento con carta di credito (1: s`ı; 0: no);
- OFFERTA ONLY DECODER: tipo di offerta solo decoder (1: s`ı; 0: no); • OFFERTA PRONTO SKY: tipo di offerta pronto Sky (1: s`ı; 0: no).

Obiettivi:

1. DEFINIRE UNA NUOVA VARIABILE CHIAMATA durata_gruppo OTTENUTA SUDDIVIDENDO LA VARIABILE DURATA NELLE SEGUENTI 8 CLASSI: 0-3 Mesi, 3-6 Mesi, 6-12 Mesi, 12-18 Mesi, 18-24 Mesi, 24-30 Mesi, 30-36 Mesi, > 36 Mesi. ELIMINARE QUINDI DAL DATASET LA VARIABILE DURATA.

```
> summary(SocTelePag$durata) # analisi della variabile durata (espressa in mesi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   12.00   18.00   26.34   39.00  167.00
```

Poiché osservo che il valore minimo della variabile "durata" è pari a 0 e il suo livello massimo è pari a 167, si ritiene opportuna una suddivisione della variabile in 8 classi (il numero di classi è stabilito arbitrariamente). Basandosi sui livelli dei quartili e della mediana vengono stabiliti gli estremi di ciascuna delle classi: 0-3 mesi, 3-6 mesi, 6-12 mesi, 12-18 mesi, 18-24 mesi, 24-30

mesi, 30-36 mesi, più di 36 mesi.

Dopo aver indicato la funzione che permette la suddivisione:

```
durata_gruppo<-function(durata){
if (durata>=0 & durata<=3){return('0-3 Mesi')} # se durata>=0 e durata <=3,
restituisci come categoria 0-3 Mesi
else if (durata>3 & durata<=6){return('3-6 Mesi')} # altrimenti se durata>3 e
durata<=6, restituisci come categoria 3-6 Mesi
else if (durata>6 & durata<=12){return('6-12 Mesi')} # altrimenti se durata>6 e
durata<=12, restituisci come categoria 6-12 Mesi
else if (durata>12 & durata<=18){return('12-18 Mesi')} # altrimenti se durata>12
e durata<=18, restituisci come categoria 12-18 Mesi
else if (durata>18 & durata<=24){return('18-24 Mesi')} # altrimenti se durata>18
e durata<=24, restituisci come categoria 18-24 Mesi
else if (durata>24 & durata<=30){return('24-30 Mesi')} # altrimenti se durata>24
e durata<=30, restituisci come categoria 24-30 Mesi
else if (durata>30 & durata<=36){return('30-36 Mesi')} # altrimenti se durata>30
e durata <=36, restituisci come categoria 30-36 Mesi
else if (durata>36){return('> 36 Mesi') # altrimenti se durata>36, restituisci
come categoria >36 Mesi
}
}
```

Dopo aver introdotto la nuova variabile nel dataset (`SocTelePag$durata_gruppo<-sapply(SocTelePag$durata,durata_gruppo)`) è necessario trasformare la variabile in variabile factor, ossia in valori non numerici (`SocTelePag$durata_gruppo<-as.factor(SocTelePag$durata_gruppo)`): la variabile durata_gruppo viene in questo modo binarizzata. È infine possibile eliminare la variabile durata (`SocTelePag$durata<-NULL`).

2. IMPLEMENTARE IL MODELLO DI REGRESSIONE LOGISTICA: ELENCARE LE VARIABILI CHE RISULTANO SIGNIFICATIVE AL LIVELLO α=0.01; SULLA BASE DELL'EFFETTIVA SIGNIFICATIVITÀ DELLE VARIABILI DI CUI AL PUNTO PRECEDENTE, COMMENTARE I VALORI DEI COEFFICIENTI DI REGRESSIONE ASSOCIATI A: ESG ATTIVAZIONI, DURATA GRUPPO 0-3 MESI, FASCIA REDDITO BASSO; DETERMINARE GLI ODDS-RATIOS E I RELATIVI P-VALUES, ELENCARE GLI ODDS-RATIOS SIGNIFICATIVI AL 10%, COMMENTARE GLI ODDS-RATIOS ASSOCIATI A DURATA GRUPO 3-6 MESI E OFFERTA ONLY DECODER; DETERMINARE GLI EFFETTI MARGINALI E COMMENTARNE I VALORI ASSOCIATI A DURATA GRUPPO 24-30 MESI E MOP PO; CALCOLARE LO PSEUDO-R2 E COMMENTARNE IL VALORE.

Il modello di regressione logistica viene implementato attraverso il comando "glm" (Generalised Linear Models), in particolare riportando la variabile target, ossia la variabile di riferimento, "stato", che indica lo stato dell'abbonamento (attivo=0, ossia "non abbandono", inattivo=1, ossia "abbandono") (`Logit<-glm(stato~.,data=SocTelePag,family=binomial)`).

```
> summary(Logit) # risultati

Call:
glm(formula = stato ~ ., family = binomial, data = SocTelePag)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1695  -0.4866  -0.3036  -0.2166   3.2393

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -1.738367   0.176229  -9.864  < 2e-16 ***
PACCHETTOR02-CINEMA            -0.740386   0.121642  -6.087 1.15e-09 ***
PACCHETTOR03-SPORT             -0.749194   0.167260  -4.479 7.49e-06 ***
PACCHETTOR04-CINEMA + SPORT    -0.990784   0.121657  -8.144 3.82e-16 ***
PACCHETTOR05-CINEMA + CALCIO   -0.742219   0.137952  -5.380 7.44e-08 ***
PACCHETTOR06-SPORT + CALCIO    -1.075912   0.159208  -6.758 1.40e-11 ***
PACCHETTOR07-TUTTO SKY         -0.718044   0.115846  -6.198 5.71e-10 ***
AREA_NIELSENNORD EST           -0.161017   0.083674  -1.924 0.054313 .
AREA_NIELSENNORD OVEST         -0.127050   0.068030  -1.868 0.061822 .
AREA_NIELSENSUD ISOLE           0.328954   0.062446   5.268 1.38e-07 ***
FASCIA_REDDITOBASSO             0.143152   0.101217   1.414 0.157272
FASCIA_REDDITOMEDIO            -0.007421   0.074428  -0.100 0.920577
FASCIA_REDDITOMEDIO ALTO       -0.051973   0.085577  -0.607 0.543635
FASCIA_REDDITOMEDIO BASSO       0.081330   0.081097   1.003 0.315922
FASCIA_ETAda 102 a 112         -7.702406 119.468126  -0.064 0.948594
FASCIA_ETAda 25 a 35 anni      -0.307088   0.103825  -2.958 0.003099 **
FASCIA_ETAda 36 a 46 anni      -0.405121   0.103719  -3.906 9.39e-05 ***
FASCIA_ETAda 47 a 57 anni      -0.516272   0.109576  -4.712 2.46e-06 ***
FASCIA_ETAda 58 a 68 anni      -0.402651   0.118413  -3.400 0.000673 ***
FASCIA_ETAda 69 a 79 anni      -0.129431   0.138440  -0.935 0.349826
FASCIA_ETAda 80 a 90 anni       0.135654   0.203694   0.666 0.505433
FASCIA_ETAda 91 a 101 anni      0.204158   0.753364   0.271 0.786395
ESG_ATTIVAZIONI                -1.628587   0.131573 -12.378  < 2e-16 ***
ESG_PROBLEMI_TECNICI           -0.253076   0.052769  -4.796 1.62e-06 ***
ESG_PROMOZIONI                 -1.126972   0.122473  -9.202  < 2e-16 ***
ESG_VARIAZIONI_CONTRATTUALI     0.134908   0.065802   2.050 0.040343 *
CANALE_VENDITA_SKY_CENTER      -0.133801   0.076949  -1.739 0.082063 .
CANALE_VENDITA_SKY_SERVICE     -0.298293   0.086294  -3.457 0.000547 ***
CANALE_VENDITA_TELESELLING     -0.135633   0.102061  -1.329 0.183866
MOP_PO                          1.293568   0.053239  24.297  < 2e-16 ***
MOP_CC                         -0.218588   0.074770  -2.923 0.003462 **
OFFERTA_ONLY_DECODER           -1.308865   0.129426 -10.113  < 2e-16 ***
OFFERTA_PRONTO_SKY             -1.153014   0.124587  -9.255  < 2e-16 ***
durata_gruppo0-3 Mesi           4.378829   0.188980  23.171  < 2e-16 ***
durata_gruppo12-18 Mesi         0.262246   0.086584   3.029 0.002455 **
durata_gruppo18-24 Mesi         0.244260   0.126038   1.938 0.052624 .
durata_gruppo24-30 Mesi        -0.237730   0.102763  -2.313 0.020702 *
durata_gruppo3-6 Mesi           2.275431   0.159746  14.244  < 2e-16 ***
durata_gruppo30-36 Mesi         0.399500   0.129756   3.079 0.002078 **
durata_gruppo6-12 Mesi          2.360059   0.089739  26.299  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16357  on 22253  degrees of freedom
Residual deviance: 12988  on 22214  degrees of freedom
AIC: 13068

Number of Fisher Scoring iterations: 9
```

Il livello di significatività α, relativo a ciascuno dei coefficienti di regressione, è indicato dai simboli specificati nell'output: in particolare le modalità variabili significative ad un livello di α pari 0.01 sono contrassegnate da un asterisco (*): ESG_VARIAZIONI_CONTRATTUALI, durata_gruppo24-30 Mesi. Risultano significativi ad un livello di α non solo pari a 0.01, ma anche per valori più bassi anche le seguenti modalità: FASCIA_ETAda 25 a 35 anni, durata_gruppo12-18 Mesi, durata_gruppo30-36 Mesi (per livello di α pari a 0.001 (**)); PACCHETTOR02-CINEMA, PACCHETTOR03-SPORT, PACCHETTOR04-CINEMA + SPORT, PACCHETTOR05-CINEMA + CALCIO,

PACCHETTOR06-SPORT + CALCIO, PACCHETTOR07-TUTTO SKY, AREA_NIELSENSUD ISOLE, FASCIA_ETAda 36 a 46 anni, FASCIA_ETAda 47 a 57 anni, FASCIA_ETAda 58 a 68 anni, ESG_ATTIVAZIONI, ESG_PROBLEMI_TECNICI, ESG_PROMOZIONI, CANALE_VENDITA_SKY_SERVICE, MOP_PO, OFFERTA_ONLY_DECODER, OFFERTA_PRONTO_SKY, durata_gruppo0-3 Mesi, durata_gruppo3-6 Mesi, durata_gruppo6-12 Mesi ((per livello di α pari a 0 (***)).

Commento dei valori dei coefficienti di regressione associati a:

- ESG_ATTIVAZIONI: valore stimato del coefficiente di regressione pari a -1.628587, tutti i clienti che hanno attivato l'abbonamento via chiamata al call center presentano una riduzione nel logit della probabilità di abbandono pari a 1.628587 rispetto ai clienti che non attivano l'abbonamento via call center, per questo motivo β stimato ha segno negativo (livello di significatività di α pari a 0.001 con p-value molto piccolo, minore di *10^(-16)).
- durata_gruppo0-3 Mesi: valore stimato del coefficiente di regressione pari a 4.378829, i clienti la cui durata del contratto è compresa tra gli 0 e i 3 mesi presentano un incremento nel logit della probabilità di abbandono pari a 4.378829 rispetto ai clienti che hanno durata del contratto diversa da 0-3 mesi, per questo motivo β stimato ha segno positivo (livello di significatività di α pari a 0.001 con p-value molto piccolo, minore di 2*10^(-16)).
- FASCIA_REDDITO_BASSO: valore stimato del coefficiente di regressione pari a 0.143152, i clienti che appartengono alla categoria della fascia di reddito basso hanno un incremento nel logit della probabilità di abbandono pari a 0.143152 rispetto ai clienti che rientrano nelle categorie delle fasce di reddito medio, medio-alto e medio basso, per questo motivo β stimato ha segno negativo. Tuttavia, dal momento che il livello di significatività è pari al 100%, ossia per α pari a 1, la variabile non è considerata rilevante ai fini del modello predittivo Logit.

Dopo aver caricato il pacchetto mfx, di supporto per il calcolo degli odds ratios e degli effetti marginali, è possibile determinare gli odds-ratios, grazie al comando `logitor(stato~.,data=SocTelePag)`:

```
> logitor(stato~.,data=SocTelePag)
Call:
logitor(formula = stato ~ ., data = SocTelePag)

Odds Ratio:
                            OddsRatio  Std. Err.       z     P>|z|
PACCHETTOR02-CINEMA         4.7693e-01 5.8014e-02  -6.0866 1.153e-09 ***
PACCHETTOR03-SPORT          4.7275e-01 7.9072e-02  -4.4792 7.491e-06 ***
PACCHETTOR04-CINEMA + SPORT 3.7129e-01 4.5170e-02  -8.1441 3.822e-16 ***
PACCHETTOR05-CINEMA + CALCIO 4.7606e-01 6.5673e-02  -5.3803 7.438e-08 ***
PACCHETTOR06-SPORT + CALCIO 3.4099e-01 5.4288e-02  -6.7579 1.400e-11 ***
PACCHETTOR07-TUTTO SKY      4.8771e-01 5.6499e-02  -6.1983 5.709e-10 ***
AREA_NIELSENNORD EST        8.5128e-01 7.1230e-02  -1.9243 0.0543134 .
AREA_NIELSENNORD OVEST      8.8069e-01 5.9913e-02  -1.8676 0.0618224 .
AREA_NIELSENSUD ISOLE       1.3895e+00 8.6770e-02   5.2678 1.381e-07 ***
FASCIA_REDDITOBASSO         1.1539e+00 1.1680e-01   1.4143 0.1572719
FASCIA_REDDITOMEDIO         9.9261e-01 7.3878e-02  -0.0997 0.9205774
FASCIA_REDDITOMEDIO ALTO    9.4935e-01 8.1242e-02  -0.6073 0.5436349
FASCIA_REDDITOMEDIO BASSO   1.0847e+00 8.7969e-02   1.0029 0.3159215
FASCIA_ETAda 102 a 112      4.5174e-04 5.3968e-02  -0.0645 0.9485940
FASCIA_ETAda 25 a 35 anni   7.3559e-01 7.6372e-02  -2.9578 0.0030989 **
FASCIA_ETAda 36 a 46 anni   6.6690e-01 6.9170e-02  -3.9059 9.386e-05 ***
FASCIA_ETAda 47 a 57 anni   5.9674e-01 6.5388e-02  -4.7115 2.458e-06 ***
FASCIA_ETAda 58 a 68 anni   6.6855e-01 7.9164e-02  -3.4004 0.0006729 ***
FASCIA_ETAda 69 a 79 anni   8.7860e-01 1.2163e-01  -0.9349 0.3498263
FASCIA_ETAda 80 a 90 anni   1.1453e+00 2.3329e-01   0.6660 0.5054326
FASCIA_ETAda 91 a 101 anni  1.2265e+00 9.2399e-01   0.2710 0.7863946
ESG_ATTIVAZIONI             1.9621e-01 2.5815e-02 -12.3778 < 2.2e-16 ***
ESG_PROBLEMI_TECNICI        7.7641e-01 4.0971e-02  -4.7959 1.620e-06 ***
```

```
ESG_PROMOZIONI               3.2401e-01 3.9683e-02  -9.2018 < 2.2e-16 ***
ESG_VARIAZIONI_CONTRATTUALI  1.1444e+00 7.5306e-02   2.0502 0.0403434 *
CANALE_VENDITA_SKY_CENTER    8.7476e-01 6.7312e-02  -1.7388 0.0820630 .
CANALE_VENDITA_SKY_SERVICE   7.4208e-01 6.4038e-02  -3.4567 0.0005469 ***
CANALE_VENDITA_TELESELLING   8.7316e-01 8.9116e-02  -1.3289 0.1838661
MOP_PO                       3.6458e+00 1.9410e-01  24.2974 < 2.2e-16 ***
MOP_CC                       8.0365e-01 6.0089e-02  -2.9235 0.0034617 **
OFFERTA_ONLY_DECODER         2.7013e-01 3.4962e-02 -10.1128 < 2.2e-16 ***
OFFERTA_PRONTO_SKY           3.1568e-01 3.9330e-02  -9.2547 < 2.2e-16 ***
durata_gruppo0-3 Mesi        7.9745e+01 1.5070e+01  23.1708 < 2.2e-16 ***
durata_gruppo12-18 Mesi      1.2998e+00 1.1255e-01   3.0288 0.0024553 **
durata_gruppo18-24 Mesi      1.2767e+00 1.6091e-01   1.9380 0.0526244 .
durata_gruppo24-30 Mesi      7.8842e-01 8.1020e-02  -2.3134 0.0207017 *
durata_gruppo3-6 Mesi        9.7321e+00 1.5547e+00  14.2441 < 2.2e-16 ***
durata_gruppo30-36 Mesi      1.4911e+00 1.9348e-01   3.0789 0.0020780 **
durata_gruppo6-12 Mesi       1.0592e+01 9.5048e-01  26.2991 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Risultano significativi al livello di α pari a 0.1 (ossia al 10%), che non vengono contrassegnati dal alcun segno: FASCIA_REDDITOBASSO, FASCIA_REDDITOMEDIO, FASCIA_REDDITOMEDIO ALTO, FASCIA_REDDITOMEDIO BASSO, FASCIA_ETAda 102 a 112, FASCIA_ETAda 69 a 79 anni, FASCIA_ETAda 80 a 90 anni, FASCIA_ETAda 91 a 101 anni, CANALE_VENDITA_TELESELLING.
Risultano significativi ad un livello di α non solo pari a 0.1, ma anche per valori più bassi anche le seguenti modalità: AREA_NIELSENNORD, EST AREA_NIELSENNORD OVEST, CANALE_VENDITA_SKY_CENTER, durata_gruppo18-24 Mesi (per livello di α pari a 0.05 (.)); ESG_VARIAZIONI_CONTRATTUALI, durata_gruppo24-30 Mesi (per livello di α pari a 0.01 (*)); FASCIA_ETAda 25 a 35 anni, MOP_CC, durata_gruppo30-36 Mesi (per livello di α pari a 0.001 (**)); PACCHETTOR02-CINEMA, PACCHETTOR03-SPORT, PACCHETTOR04-CINEMA + SPORT, PACCHETTOR05-CINEMA + CALCIO, PACCHETTOR06-SPORT + CALCIO, PACCHETTOR07-TUTTO SKY, AREA_NIELSENSUD ISOLE, FASCIA_ETAda 36 a 46 anni, FASCIA_ETAda 47 a 57 anni, FASCIA_ETAda 58 a 68 anni, ESG_ATTIVAZIONI, ESG_PROBLEMI_TECNICI, ESG_PROMOZIONI, CANALE_VENDITA_SKY_SERVICE, MOP_PO, OFFERTA_ONLY_DECODER, OFFERTA_PRONTO_SKY , durata_gruppo0-3 Mesi , durata_gruppo3-6 Mesi, durata_gruppo6-12 Mesi (per livello di α pari a 0 (***)).

Commentare gli odds-ratios associati a:

- durata_gruppo3-6 Mesi: odds-ratio pari a 9.7321. Poiché assume valore maggiore di 1, esso rappresenta l'incremento di probabilità di abbandono dei clienti con contratto di durata tra i 3 e i 6 mesi rispetto ai clienti con contratto di durata minore di 3 mesi o maggiore di 6 mesi. Per ogni cliente con durata di contratto tra i 3 e i 6 mesi si registra un aumento della propensione all'abbandono pari a 9.7321 rispetto agli a quella degli altri clienti (livello di significatività di α pari a 0.001 con p-value pari molto piccolo, pari a 2.2*10^(-16)).
- OFFERTA_ONLY_DECODER: odds-ratio pari a 2.7013*10^(-1), ossia a 0.27013. Il valore dell'odds ratio è compreso tra 0 e 1, c'è quindi relazione negativa tra il tipo di offerta solo decoder e la propensione all'abbandono, rispetto a coloro che non l'attivano, vi è quindi una minor propensione all'abbandono da parte dei clienti che attivano l'offerta only decoder rispetto a chi non la attiva. La minor propensione all'abbandono ammonta a 1/0.27013=3.7019, quindi i clienti che attivano l'offerta solo decoder presentano una propensione all'abbandono che risulta essere minore di 3.7019 volte di quella dei clienti che non attivano l'offerta only decoder (livello di significatività di α pari a 0.001 con p-value pari molto piccolo, pari a 2.2*10^(-16)).

È necessario ora calcolare gli effetti marginali con il comando `logitmfx(stato~.,data=SocTelePag)`:

```
> logitmfx(stato~.,data=SocTelePag) # effetti marginali
Call:
logitmfx(formula = stato ~ ., data = SocTelePag)

Marginal Effects:
                                dF/dx   Std. Err.       z     P>|z|
PACCHETTOR02-CINEMA          -0.04189067  0.00564679  -7.4185 1.185e-13 ***
PACCHETTOR03-SPORT           -0.03868904  0.00626711  -6.1733 6.686e-10 ***
PACCHETTOR04-CINEMA + SPORT  -0.05669134  0.00588959  -9.6257 < 2.2e-16 ***
PACCHETTOR05-CINEMA + CALCIO -0.03919017  0.00546358  -7.1730 7.338e-13 ***
PACCHETTOR06-SPORT + CALCIO  -0.04978302  0.00471648 -10.5551 < 2.2e-16 ***
PACCHETTOR07-TUTTO SKY       -0.04783688  0.00756748  -6.3214 2.592e-10 ***
AREA_NIELSENNORD EST         -0.01060847  0.00525708  -2.0179 0.0435977 *
AREA_NIELSENNORD OVEST       -0.00860453  0.00451860  -1.9042 0.0568784 .
AREA_NIELSENSUD ISOLE         0.02416028  0.00489137   4.9394 7.838e-07 ***
FASCIA_REDDITOBASSO           0.01041543  0.00775458   1.3431 0.1792294
FASCIA_REDDITOMEDIO          -0.00051184  0.00512978  -0.0998 0.9205197
FASCIA_REDDITOMEDIO ALTO     -0.00353853  0.00574708  -0.6157 0.5380860
FASCIA_REDDITOMEDIO BASSO     0.00573013  0.00583120   0.9827 0.3257713
FASCIA_ETAda 102 a 112       -0.07457429  0.00482401 -15.4590 < 2.2e-16 ***
FASCIA_ETAda 25 a 35 anni    -0.01985022  0.00630486  -3.1484 0.0016417 **
FASCIA_ETAda 36 a 46 anni    -0.02640425  0.00642614  -4.1089 3.976e-05 ***
FASCIA_ETAda 47 a 57 anni    -0.03164791  0.00599102  -5.2826 1.274e-07 ***
FASCIA_ETAda 58 a 68 anni    -0.02447059  0.00632466  -3.8691 0.0001092 ***
FASCIA_ETAda 69 a 79 anni    -0.00850304  0.00865389  -0.9826 0.3258194
FASCIA_ETAda 80 a 90 anni     0.00990647  0.01571057   0.6306 0.5283280
FASCIA_ETAda 91 a 101 anni    0.01537182  0.06166271   0.2493 0.8031374
ESG_ATTIVAZIONI              -0.07110167  0.00374807 -18.9702 < 2.2e-16 ***
ESG_PROBLEMI_TECNICI         -0.01685750  0.00338983  -4.9730 6.594e-07 ***
ESG_PROMOZIONI               -0.05520179  0.00398721 -13.8447 < 2.2e-16 ***
ESG_VARIAZIONI_CONTRATTUALI   0.00970081  0.00492575   1.9694 0.0489062 *
CANALE_VENDITA_SKY_CENTER    -0.00920763  0.00527506  -1.7455 0.0808972 .
CANALE_VENDITA_SKY_SERVICE   -0.01925903  0.00519877  -3.7045 0.0002118 ***
CANALE_VENDITA_TELESELLING   -0.00895495  0.00643819  -1.3909 0.1642527
MOP_PO                        0.11805810  0.00610458  19.3393 < 2.2e-16 ***
MOP_CC                       -0.01436245  0.00465523  -3.0852 0.0020340 **
OFFERTA_ONLY_DECODER         -0.05958417  0.00394944 -15.0867 < 2.2e-16 ***
OFFERTA_PRONTO_SKY           -0.05501703  0.00415348 -13.2460 < 2.2e-16 ***
durata_gruppo0-3 Mesi         0.77217441  0.02516066  30.6897 < 2.2e-16 ***
durata_gruppo12-18 Mesi       0.01915236  0.00666735   2.8726 0.0040716 **
durata_gruppo18-24 Mesi       0.01851334  0.01043678   1.7739 0.0760871 .
durata_gruppo24-30 Mesi      -0.01520654  0.00608580  -2.4987 0.0124652 *
durata_gruppo3-6 Mesi         0.32310439  0.03459764   9.3389 < 2.2e-16 ***
durata_gruppo30-36 Mesi       0.03228587  0.01211747   2.6644 0.0077124 **
durata_gruppo6-12 Mesi        0.34090931  0.01908212  17.8654 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
dF/dx is for discrete change for the following variables:

 [1] "PACCHETTOR02-CINEMA"          "PACCHETTOR03-SPORT"           "PACCHETTOR04-CINEMA + SPORT"
 [4] "PACCHETTOR05-CINEMA + CALCIO" "PACCHETTOR06-SPORT + CALCIO"  "PACCHETTOR07-TUTTO SKY"
 [7] "AREA_NIELSENNORD EST"         "AREA_NIELSENNORD OVEST"       "AREA_NIELSENSUD ISOLE"
[10] "FASCIA_REDDITOBASSO"          "FASCIA_REDDITOMEDIO"          "FASCIA_REDDITOMEDIO ALTO"
[13] "FASCIA_REDDITOMEDIO BASSO"    "FASCIA_ETAda 102 a 112"       "FASCIA_ETAda 25 a 35 anni"
[16] "FASCIA_ETAda 36 a 46 anni"    "FASCIA_ETAda 47 a 57 anni"    "FASCIA_ETAda 58 a 68 anni"
[19] "FASCIA_ETAda 69 a 79 anni"    "FASCIA_ETAda 80 a 90 anni"    "FASCIA_ETAda 91 a 101 anni"
[22] "ESG_ATTIVAZIONI"              "ESG_PROBLEMI_TECNICI"         "ESG_PROMOZIONI"
[25] "ESG_VARIAZIONI_CONTRATTUALI"  "CANALE_VENDITA_SKY_CENTER"    "CANALE_VENDITA_SKY_SERVICE"
[28] "CANALE_VENDITA_TELESELLING"   "MOP_PO"                       "MOP_CC"
[31] "OFFERTA_ONLY_DECODER"         "OFFERTA_PRONTO_SKY"           "durata_gruppo0-3 Mesi"
[34] "durata_gruppo12-18 Mesi"      "durata_gruppo18-24 Mesi"      "durata_gruppo24-30 Mesi"
[37] "durata_gruppo3-6 Mesi"        "durata_gruppo30-36 Mesi"      "durata_gruppo6-12 Mesi"
```

- durata_gruppo24-30 Mesi: valore degli effetti marginali pari a -0.01520654, si tratta di un effetto marginale negativo, quindi si verifica una riduzione della probabilità di abbandono. In corrispondenza di ogni incremento unitario di un contratto di durata compresa tra 3 e 6

mesi si registra un aumento della probabilità di abbandono pari a 0.01520654 (livello di significatività α pari a 0.01 e p-value pari a -2.4987).

- MOP_PO: valore degli effetti marginali pari a 0.11805810, si tratta di un effetto marginale positivo, quindi si verifica un aumento della probabilità di abbandono. In corrispondenza di ogni incremento unitario di pagamenti con bollettino postale si registra un amento nella probabilità di abbandono pari a 0.11805810 (livello di significatività α pari a 0 e p-value molto piccolo, pari a 2.2*10^(-16)).

Dopo aver caricato il pacchetto pscl, è possibile calcolare l'indice pseudo-R2 di MacFadden, grazie al comando generale pseudo R2:

```
> pseudoR2<-pR2(Logit) # calcolo dell'indice pseudo-R^2 di Mc-Fadden (valutazione della bontà del modello)
> pseudoR2
         llh       llhNull            G2      McFadden          r2ML          r2CU
-6493.9215912 -8178.5303621  3369.2175418     0.2059794     0.1404947     0.2699213
```

L'indice pseudo R2 di MacFadden stabilisce la valutazione di bontà del modello: esso è compreso tra 0 e 1, tuttavia in un dataset che comprende numerosissime osservazioni si considerano ottimali valori di pseudoR2 di MacFadden tra lo 0.2 e lo 0.3. Per questo motivo, poiché l'indice di McFadden del modello logit implementato è pari a 0.2059794, esso è da considerarsi ottimale.

3. IMPLEMENTARE L'ALBERO DI CLASSIFICAZIONE, COMMENTARE IL NODO RADICE; OTTENERE LA VISUALIZZAZIONE GRAFICA DEI RISULTATI MEDIANTE LA STRUTTURA AD ALBERO DELL'OUTPUT DELL'ANALISI E COMMENTARE IL NODO RADICE E I NODI TERMINALI; STABILIRE LA PROBABILITÀ DI ABBANDONO DEI CLIENTI CHE PRESENTANO COME CARATTERISTICHE durata_grupop 0-3Mesi, durata_gruppo 6-12 Mesi. MOP PO=0, ESG ATTIVAZIONI=1, durata gruppo 6-12 Mesi; STABILIRE SE NECESSARIA UN'EVENTUALE POTATURA DELL'ALBERO ATTRAVERSO UN OPPORTUNO STRUMENTO GRAFICO

Per l'implementazione dell'albero di classificazione è necessario caricare nella console i pacchetti rpart, necessario per la costruzione della struttura ad albero di classificazione, e rattle, di supporto per la visualizzazione grafica dell'albero di classificazione.
È possibile quindi implementare l'albero di classificazione, in funzione di tutte le variabili registrate nel dataset, grazie al comando `tree<-rpart(stato~.,data=SocTelePag,method="class")`.

```
> tree # richiamo l'oggetto tree per la visualizzazione delle variabili indipendenti selezionate nell'albero di
classificazione caratterizzate da maggior potere predittivo
n= 22254

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 22254 2677 0 (0.87970702 0.12029298)
   2) durata_gruppo=> 36 Mesi,12-18 Mesi,18-24 Mesi,24-30 Mesi,3-6 Mesi,30-36 Mesi 18524 1462 0 (0.92107536
 0.07892464) *
   3) durata_gruppo=0-3 Mesi,6-12 Mesi 3730 1215 0 (0.67426273 0.32573727)
     6) MOP_PO=0 2717  568 0 (0.79094590 0.20905410)
      12) ESG_ATTIVAZIONI=1 917   57 0 (0.93784079 0.06215921) *
      13) ESG_ATTIVAZIONI=0 1800  511 0 (0.71611111 0.28388889)
        26) durata_gruppo=6-12 Mesi 1724  435 0 (0.74767981 0.25232019) *
        27) durata_gruppo=0-3 Mesi 76    0 1 (0.00000000 1.00000000) *
     7) MOP_PO=1 1013  366 1 (0.36130306 0.63869694) *
```

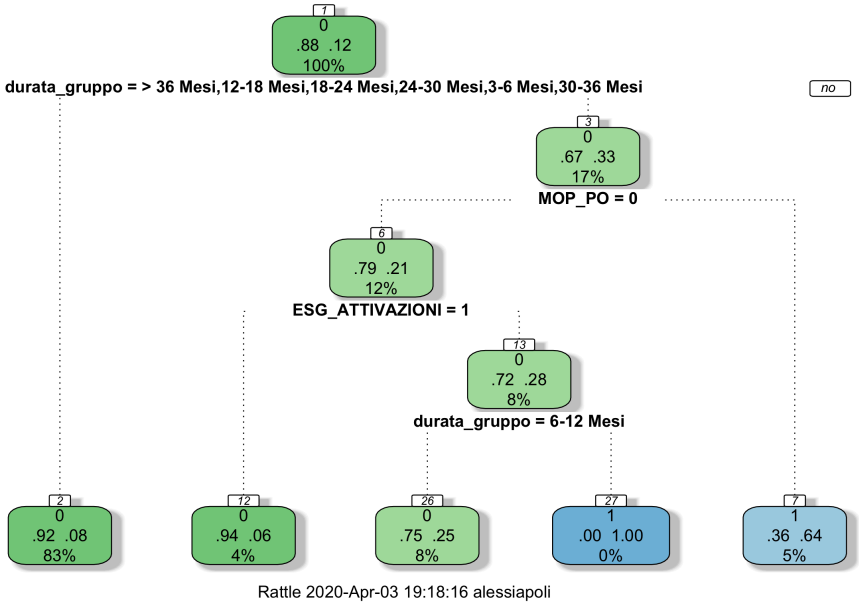
Il nodo radice raccoglie la totalità delle osservazioni contenute nel dataset, che, come si poteva evincere sin dal principio grazie al comando `dim(SocTelePag)`, sono 22254. Il nodo radice è dunque un nodo puramente descrittivo, che ripartisce le osservazioni in funzione della variabile di interesse, ossia la variabile "STATO", una variabile dicotomica che si esplicita in due modalità: 0 ossia in caso di stato attivo dell'abbonamento e quindi di non abbandono, e 1 ossia in caso di

stato inattivo dell'abbonamento e quindi di abbandono. Dunque nel nodo radice non entra in gioco nessuna variabile esplicativa e si limita a fornire l'informazione riguardo a come viene ripartito l'abbandono o il non abbandono in relazione alle osservazioni.

I nodi in analisi, in questo caso il nodo root, ovvero il nodo radice, vengono descritto riportando numero di osservazioni che sono contenute (n), la perdita (loss), la categoria modale di riferimento e, all'interno delle parentesi le frequenze relative percentuali delle osservazioni in relazione al non abbandono e all'abbandono.

Nel nodo radice, nell'output indicato dal numero 1, il numero di osservazioni raccolte, ossia il numero totale di osservazioni, corrisponde a 22254; la perdita è pari a 2677, che corrisponde al numero di osservazioni che rientrano nella categoria modale 1, nonché categoria modale di riferimento; all'interno del nodo radice circa l'87.97% delle osservazioni (19577 soggetti) non abbandonano e circa il 12.03% delle osservazioni (2677 soggetti) abbandonano.

Può essere svolta un'analisi analoga dall'osservazione del grafico dell'albero di classificazione, rappresentato grazie al comando `fancyRpartPlot(tree)`.



Come già evinto dall'analisi dell'output del comando relativo alla costruzione dell'albero in sé, grazie al comando "tree", riguardo al nodo radice. Il nodo radice è il primo nodo dall'alto all'interno del grafico: esso contiene il 100% delle osservazioni, che vengono ripartite a seconda dello "STATO" dell'abbonamento. La categoria modale di riferimento è 0, ossia quando l'abbonamento è attivo, e le osservazioni di clienti che risultano attive rappresentano il 12% (soggetti che abbandonano); invece le osservazioni di clienti che risultano inattivi rappresentano l'88% (soggetti che non abbandonano).

Il numero corrispondente a ciascun nodo terminale può essere evinto dall'output del comando "tree", contrassegnati dal simbolo asterisco *, e dalla configurazione grafica dell'albero decisionale.

- Nodo terminale numero 2: è il primo nodo da sinistra. il nodo radice viene splittato in funzione della variabile "durata_gruppo ≥ 36 Mesi, 12-18 Mesi, 18-24 Mesi, 24-30 Mesi, 3-6 Mesi, 30-36 Mesi", in funzione della quale si formano un nodo figlio, che comprende le osservazioni che non appartengono alla categoria, e un nodo terminale, che comprende le osservazioni che appartengono alla categoria (primo nodo in basso a sinistra).
  I clienti che hanno un contratto di durata compresa tra i 3 e i 6 mesi, tra i 12 e i 18 mesi, tra i 18 e i 24 mesi, tra i 24 e i 30 mesi, tra i 30 e i 36 mesi, o maggiore di 36 mesi, sono 18524: nel primo nodo terminale, quindi, sono state allocate l'83% delle osservazioni totali, con

modalità di riferimento 0, ossia il non abbandono (per questo motivo il nodo è di colore verde). All'interno del nodo si identifica che il 92% delle osservazioni appartenenti a questa categoria, ovvero 17062 clienti, hanno l'abbonamento attivo, dunque non abbandonano; invece, l'8% delle osservazioni appartenenti alla categoria, ovvero 1462 clienti, hanno abbonamento inattiva, dunque abbandonano. Il nodo illustra dunque la probabilità di abbandono qualora i clienti stipulino contratti dalla durata tra i 3 e i 6 mesi, tra i 12 e i 18 mesi, tra i 18 e i 24 mesi, tra i 24 e i 30 mesi, tra i 30 e i 36 mesi, o maggiore di 36 mesi, che è pari all'8%.

- nodo terminale 12: nel secondo nodo terminale da sinistra sono allocate il 4% del totale delle osservazioni. I 917 clienti che ne fano parte presentano le seguenti caratteristiche: durata contrattuale compresa tra 0 e 3 mesi o tra 6 e 12 mesi, MOP_PO=0 e ESG_ATTIVAZIONI=1. Il nodo terminale in questione presenta come modalità di riferimento 0, ovvero il non abbandono, quindi il 94% delle osservazioni allocate in esso non abbandona, ossia 849 clienti hanno abbonamento attivo, e il restante 6% delle osservazioni abbandona, ossia 57 clienti hanno abbonamento inattivo. In questo nodo terminale è stabilita la probabilità di non abbandono dei clienti che presentano come caratteristiche durata contrattuale compresa tra 0 e 3 mesi o tra 6 e 12 mesi, MOP_PO=0 e ESG_ATTIVAZIONI=1.
- nodo terminale 26: nel terzo nodo terminale da sinistra sono allocate l'8% del totale delle osservazioni. I 1724 clienti che ne fanno parte presentano le seguenti caratteristiche: durata contrattuale compresa tra 6 e 12 mesi, MOP_PO=0 e ESG_ATTIVAZIONI=0. Il nodo terminale ha come modalità di riferimento 0, ovvero il non abbandono, quindi il 75% delle osservazioni allocate in esso non abbandona, ossia 1289 hanno abbonamento attivo, e il restante 25% delle osservazioni abbandona, ossia 435 clienti hanno abbonamento inattivo. In questo nodo terminale è stabilita la probabilità di non abbandono dei clienti che presentano come caratteristiche durata contrattuale compresa tra 6 e 12 mesi, MOP_PO=0 e ESG_ATTIVAZIONI=0.
- nodo terminale 27: nel quarto nodo terminale da sinistra, non rientra alcuna osservazione del dataset. Infatti nessun cliente presenta le seguenti caratteristiche: durata contrattuale compresa tra 0 e 3 mesi, MOP_PO=0 e ESG_ATTIVAZIONI=0. Infatti l'8% delle osservazioni che rientrano nel nodo figlio 13, ossia 1724 clienti, ricadono interamente nel nodo terminale 12: le osservazioni in questo nodo vengono ripartite a seconda della durata del contratto, che sia tra 0 e 3 mesi o che sia tra 6 e 12 mesi (i contratti che durano più o meno di ciascuna delle due opzioni erano stati analizzati nel nodo radice). Ma dei clienti che presentano come caratteristiche MOP_PO<0.5 e ESG_ATTIVAZIONI<0.5, nessuno risulta avere contratto di durata tra 0 e 3 mesi. Dunque nel nodo terminale 27 ricade lo 0% delle osservaizoni.
- nodo terminale 7: nel quinto nodo terminale da sinistra sono allocate il 5% del totale delle osservazioni. I 1013 clienti che ne fanno parte presentano le seguenti caratteristiche: durata contrattuale compresa tra 0 e 3 mesi o tra 6 e 12 mesi, MOP_PO=1. Il nodo terminale ha come modalità di riferimento 1 (motivo per cui la cella ha sfondo blu) ovvero l'abbandono, quindi il 64% delle osservazioni allocate in esso abbandona, ossia 647 clienti hanno abbonamento inattivo, e il restante 36% delle osservazioni non abbandona, ossia 366 clienti hanno abbonamento attivo. In questo nodo terminale è stabilita la probabilità di abbandono dei clienti che presentano come caratteristiche durata contrattuale compresa tra 0 e 3 mesi o tra 6 e 12 mesi, MOP_PO=1.

Alessia Poli, 922096, EMA
a. a. 2019 2020

Il cliente che presenta come caratteristiche durata gruppo 0-3 Mesi e durata gruppo 6-12 Mesi, MOP_PO=0, ESG_ATTIVAZIONI=1 cadono nel nodo terminale 12. In particolare è opportuno osservare come l'osservazione in questioni rientri in ciascun nodo dell'albero decisionale:

- partendo dal nodo radice le osservazioni sono classificate in relazione alla durata del contratto di abbonamento; in particolare le osservazioni che presentano durata che non sia tra i 3 e i 6 mesi, o tra i 12 e i 18 mesi, o tra i 18 e i 24 mesi, o tra i 24 e i 30 mesi, o tra i 30 e i 36 mesi, o maggiore di 36 mesi, come in questo caso, ricadono nel primo nodo figlio a cui si giunge percorrendo il ramo sulla destra del nodo radice (ramo=no).
- il 17% delle osservazioni totali presentano durata che non sia tra i 3 e i 6 mesi, o tra i 12 e i 18 mesi, o tra i 18 e i 24 mesi, o tra i 24 e i 30 mesi, o tra i 30 e i 36 mesi, o maggiore di 36 mesi. Le 3730 osservazioni che ricadono nel nodo figlio 3 sono ripartite a seconda che abbiano MOP_PO=0. Le osservazioni che presentano MOP_PO=0 ricadono nel nodo figlio 6, percorrendo il ramo sulla sinistra del nodo 3 (ramo=si).
- i2% delle osservazioni totali presentano avere durata diversa da 3-6 mesi, o 12-18 mesi, o 18 -24 mesi, o 24 -30 mesi, o 30-36 mesi, o maggiore di 36 e MOP_PO=0. I 2717 clienti che rientrano nel nodo figlio 6 sono ripartite in relazione alla modalità ESG_ATTIVAZIONI=1. I clienti che presentano ESG_ATTIVAZIONI=1 ricadono nel nodo terminale 12, percorrendo il ramo sulla sinistra del nodo 6 (ramo=si).
- la variabile "durata gruppo=6-12 mesi", quindi, non risulta avere ruolo discriminatorio in questa analisi, poiché non rientra nei criteri di suddivisione della clientela che presenta le caratteristiche durata gruppo 0-3 Mesi e durata gruppo 6-12 Mesi, MOP_PO=0, ESG_ATTIVAZIONI=1.

La probabilità di abbandono dei clienti che presentano le caratteristiche durata gruppo 0-3 Mesi e durata gruppo 6-12 Mesi, MOP_PO=0 e ESG_ATTIVAZIONI=1 è analizzata nel nodo terminale 12: nel nodo ricade il 4% delle osservazioni totali, ossia 917 clienti presentano le caratteristiche elencate; di questi il 94% non abbandona (849 clienti hanno abbonamento attivo) e il restante 6% abbandona (57 clienti hanno abbonamento inattivo).

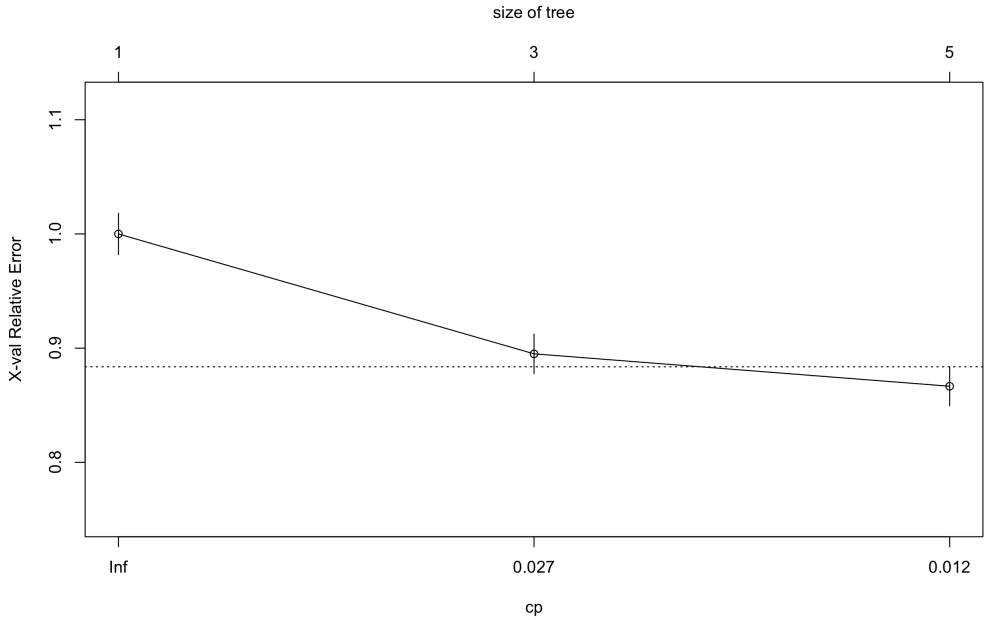
A questo punto è necessario stabilire se le dimensioni dell'albero decisionale sono da considerarsi ottimali e, in caso contrario, procedere con un'eventuale potatura al fine di ridurre il numero di nodi terminali dell'albero decisionale, riducendone la complessità: la dimensione ottimale di un albero di classificazione è infatti individuata nel numero di nodi terminali, e quindi nel parametro di complessità, a cui è associato il minimo errore di validazione.

Con il comando "princtcp(tree)" vengono determinati i valori del parametro di complessità, dell'errore di training, dell'errore di validazione e della devianza standard, in corrispondenza di un certo numero di foglie (=nsplit).

```
> # validazione dell'albero di classificazione mediante l'utilizzo del parametro di complessità (CP) e l'errore di
validazione
> printcp(tree)

Classification tree:
rpart(formula = stato ~ ., data = SocTelePag, method = "class")

Variables actually used in tree construction:
[1] durata_gruppo   ESG_ATTIVAZIONI MOP_PO

Root node error: 2677/22254 = 0.12029

n= 22254

        CP nsplit rel error  xerror     xstd
1 0.052484      0   1.00000 1.00000 0.018128
2 0.014195      2   0.89503 0.89503 0.017273
3 0.010000      4   0.86664 0.86664 0.017029
```

A questo punto è necessaria la costruzione del grafico che permette di individuare il valore di parametro di complessità (CP) a cui è associato il minor errore di validazione: in corrispondenza di questa associazione è individuata la dimensione ottimale dell'albero decisionale, ossia il numero ottimale di foglie (nodi terminali).

```
tree$cptable[which.min(tree$cptable[,"xerror"]),"CP"] # identificazione del valore del parametro di complessità a cui è associato il minimo xerror (errore di validazione)
```



Il grafico riporta sull'asse delle ascisse inferiore i parametri di complessità, sull'asse delle ordinate l'errore di validazione e sull'asse delle ascisse superiore il numero di nodi associato a ciascun parametro di complessità.
Il minor errore di validazione è identificato in corrispondenza del parametro di complessità corrispondente a 5 nodi terminali, ossia 0.01: poiché 5 è il numero dei nodi terminali dell'albero implementato, l'albero non va potato ed è da considerarsi soddisfacente. Dunque il modello costruito non ha necessità di essere migliorato.

4. RAPPRESENTARE GRAFICAMENTE LE CURVE ROC DI ENTRAMBI I MODELLI IMPLEMENTATI E STABILIRE QUALE DEI DUE MODELLI RISULTI CARATTERIZZATO DA MAGGIOR CAPACITÀ PRODUTTIVA.

Al termine dell'implementazione dei due metodi predittivi di data mining, è necessario confrontarli al fine di determinare il modello con la miglior capacità previsiva. Lo strumento di confronto è la curva ROC (Receiver Operating Characteristic) e il corrispondente indice AUC (sul grafico indicato con la sigla AUROC e corrisponde al valore dell'area sottostante alla curva ROC).
Innanzitutto bisogna caricare il pacchetto InformationValue, necessario per la visualizzazione grafica della curva e la determinazione dell'indice, e calcolare le probabilità di abbandono rispetto a ciascun cliente.
La curva ROC è calcolata sulla base delle frequenze assolute calcolate sulla matrice di confusione, avente per righe evento e non-evento osservati e sulle colonne evento e non-evento previsti: il grafico riporterà infatti sulle ascisse la proporzione di falsi positivi, ossia la proporzione di non

eventi previsti come eventi, e sulle ordinate la sensibilità, ossia la proporzione di eventi previsti come tali.

```
print("Matrice di confusione per il modello logit");table(SocTelePag$stato,prob_logit>0.65) # matrice di confusione con threshold>cutoff=0.65
```

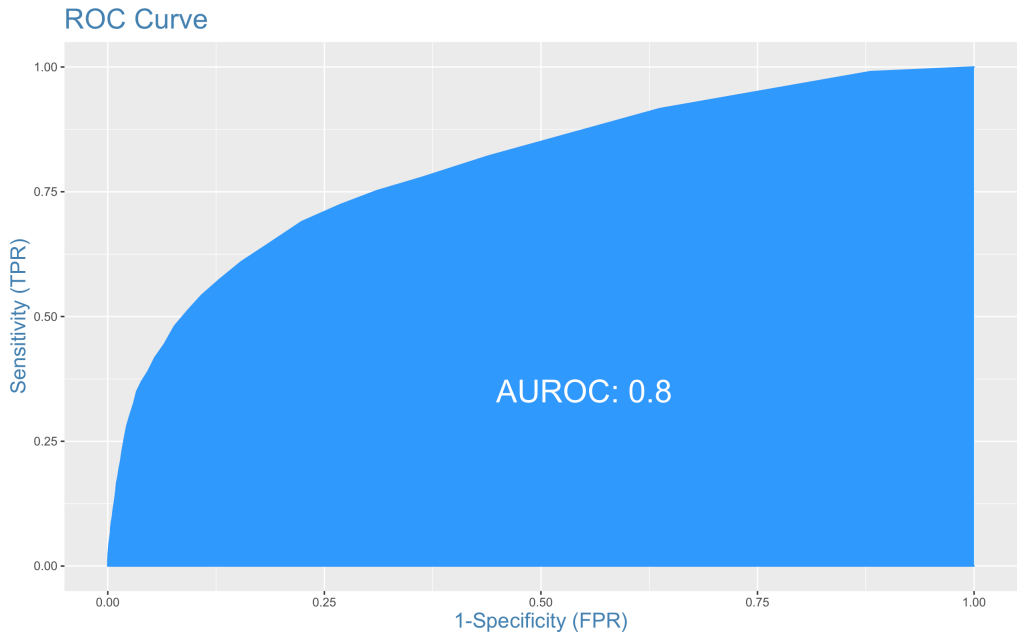
```
[1] "Matrice di confusione per il modello logit"

   FALSE  TRUE
0 19483    94
1  2425   252
```

```
print("Matrice di confusione per gli alberi di classificazione");table(SocTelePag$stato,prob_tree_1>0.65) # matrice di confusione con threshold>cutoff=0.65
```

```
[1] "Matrice di confusione per gli alberi di classificazione"

   FALSE  TRUE
0 19577     0
1  2601    76
```

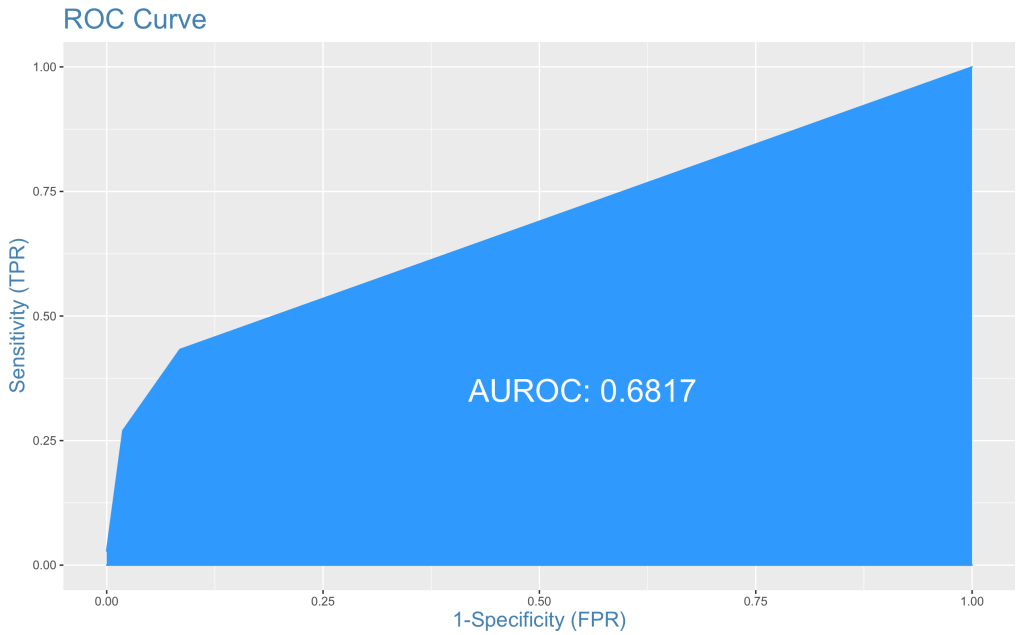
Dopo aver costruito la matrice di confusione per entrambi i modelli, è possibile implementare la curva ROC per ciascuno dei due, che associerà a ciascun punto della curva un livello di cut-off.

```
ROC_logit<-plotROC(SocTelePag$stato,prob_logit) # grafico della curva ROC con valore dell'indice AUROC del modello di regressione logistica
```




```
ROC_tree<-plotROC(SocTelePag$stato,prob_tree_1) # grafico della curva ROC e relativo indice AUROC del modello degli alberi di classificazione
```

Alessia Poli, 922096, EMA
a. a. 2019 2020



L'indice AUROC assume valori compresi tra 0 e 1: più è vicino ad 1, migliore è la capacità predittiva del modello a cui si riferisce.

Quindi in questo caso: dalla comparazione dell'indice AUC (AUROC nel grafico) (0.8 nel caso del modello logit vs 0.6817 nel caso degli alberi di classificazione) e sulla base del criterio basato sulla matrice di confusione, il modello di regressione logistica è preferibile all'albero di classificazione in quanto è caratterizzato da una maggior accuratezza previsiva.