21 VARIABILI

7043 OSSERVAZIONI

Telco Customer Churn

Il dataset contiene informazioni riguardo il tasso di abbandono dei clienti nel settore delle telecomunicazioni

Manuel Manzo
(matricola: s1106217)

Variabili del dataset

Le variabili del dataset contengono informazioni riguardanti le caratteristiche del cliente, i servizi acquistati e gli importi pagati

customerID	**MultipleLines**	**PaperlessBilling**
gender	**InternetService**	**PaymentMethod**
Partner	**OnlineBackup**	**MonthlyCharges**
Dependents	**DeviceProtection**	**TotalCharges**
tenure	**TechSupport**	**SeniorCitizen**
StreamingMovies	**Contract**	**OnlineSecurity**
PhoneService	**StreamingTV**	**Churn**

Churn

Obiettivo del modello

Proporre alle imprese operanti nel settore delle telecomunicazioni un modello, basato su tecniche di data mining che permetta di **PREDIRE SE UN CLIENTE ABBANDONERÀ LA NOSTRA IMPRESA**

CHURN è la variabile dipendente

Utilità del modello per le imprese

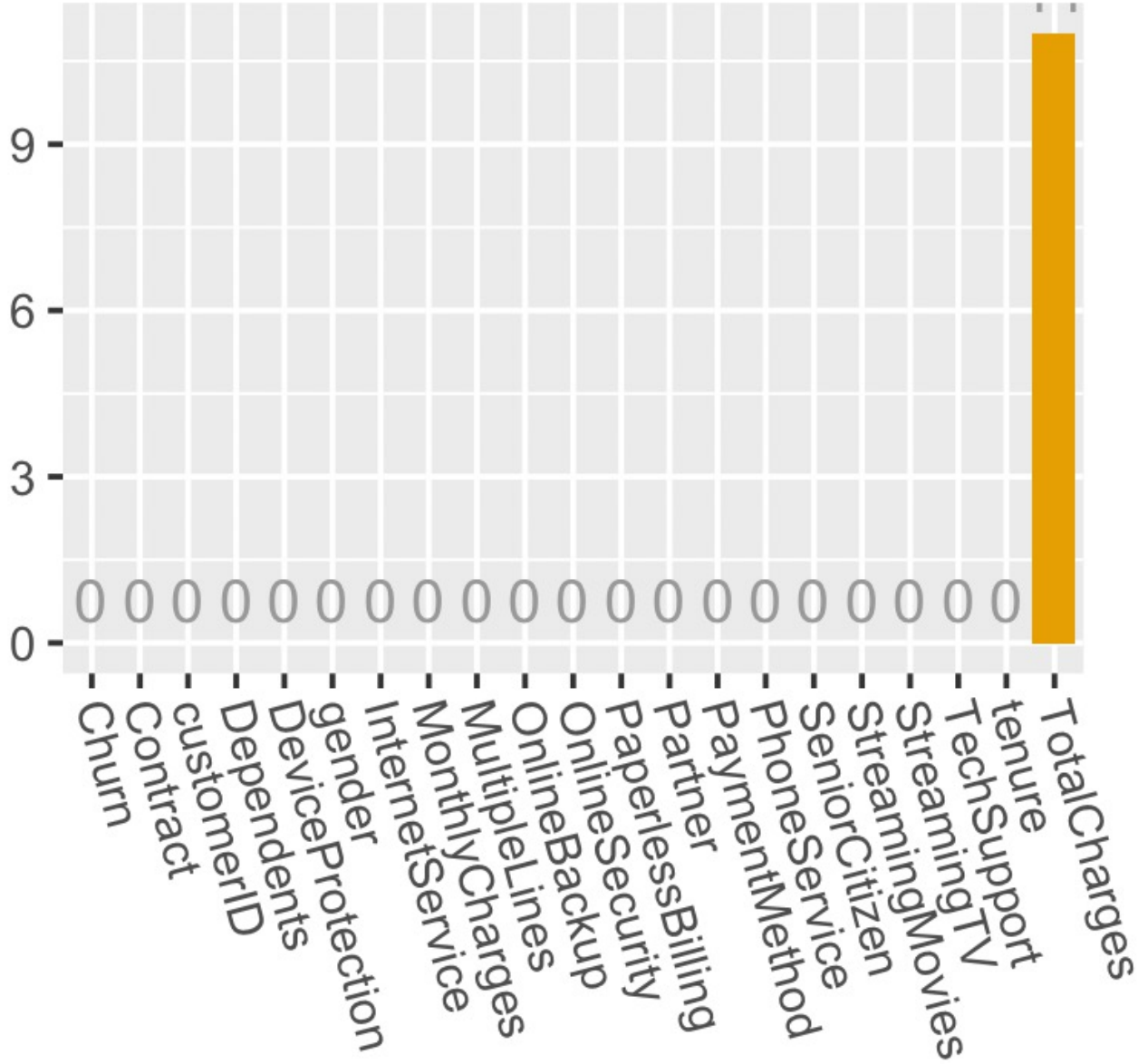
Le imprese operanti nel TELCO potranno adottare **campagne comunicative e pubblicitarie** indirizzare ai clienti che probabilmente abbandoneranno l'impresa

Data Cleaning

Duplicati
Non sono presenti duplicati

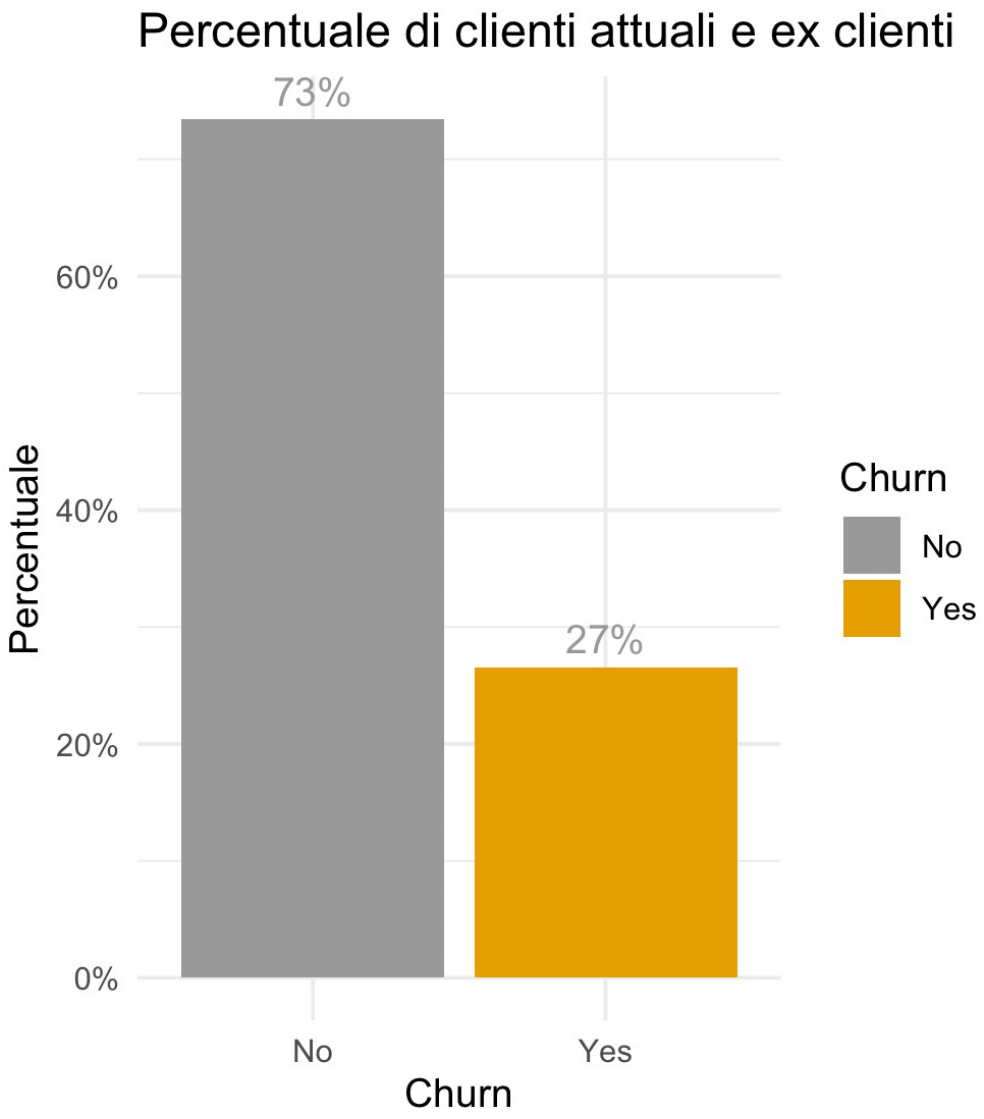
Valori mancanti
11 NA nella variabile TotalCharges

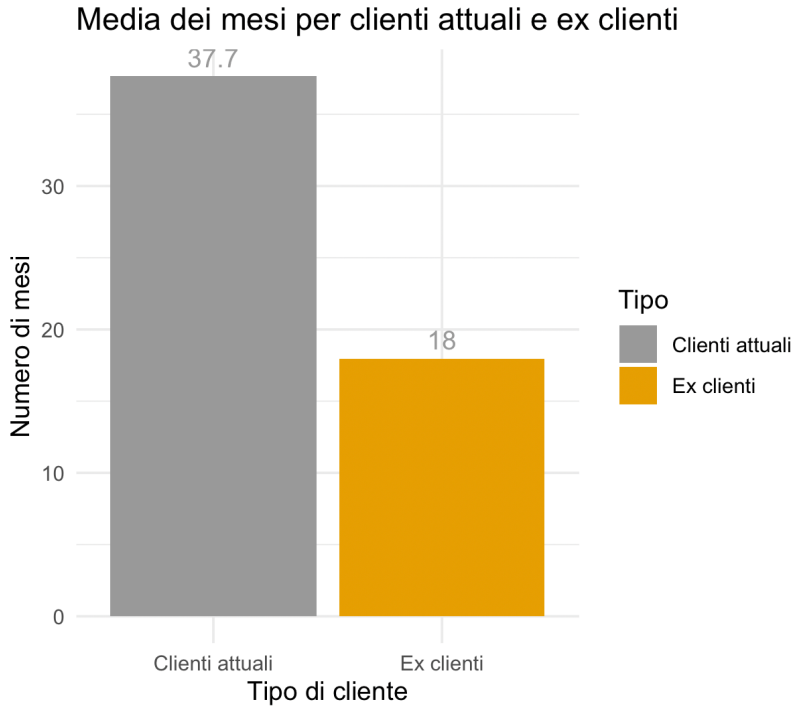
Outlier
Non sono presenti outlier

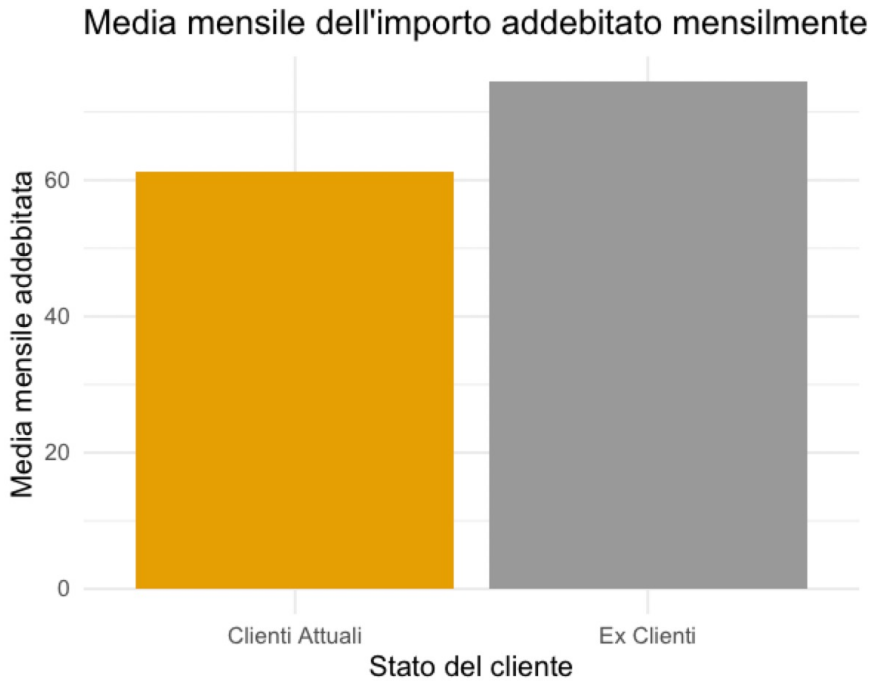


Statistiche descrittive

VARIABILI QUALITATIVE

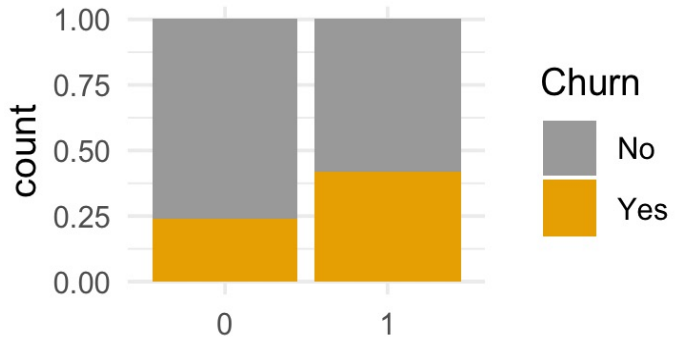




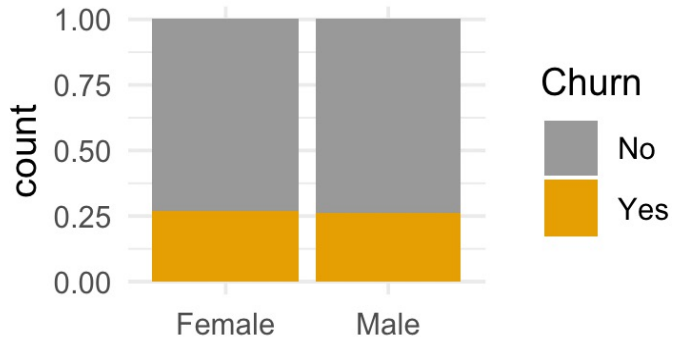


Statistiche descrittive

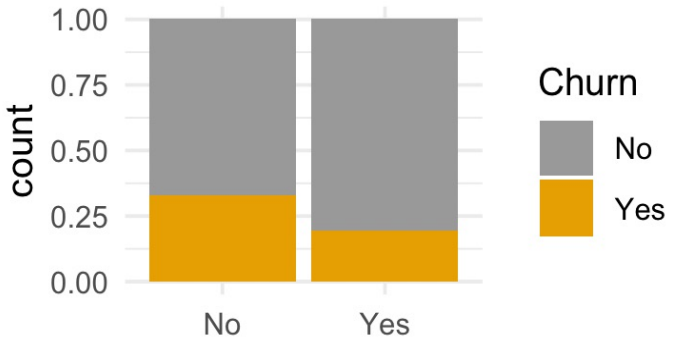
ALTRE VARIABILI QUALITATIVE



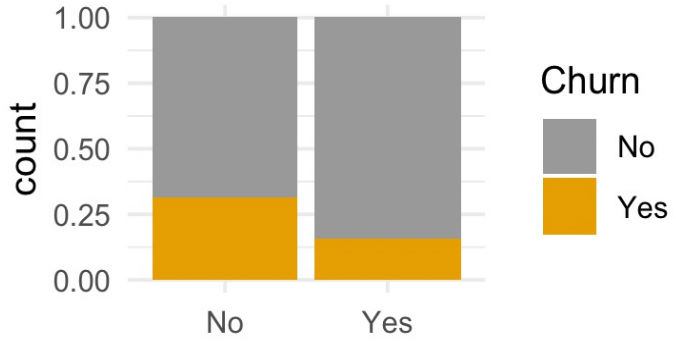
SeniorCitizen



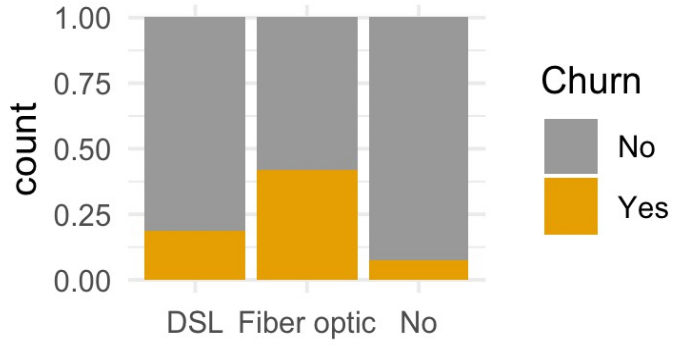
gender



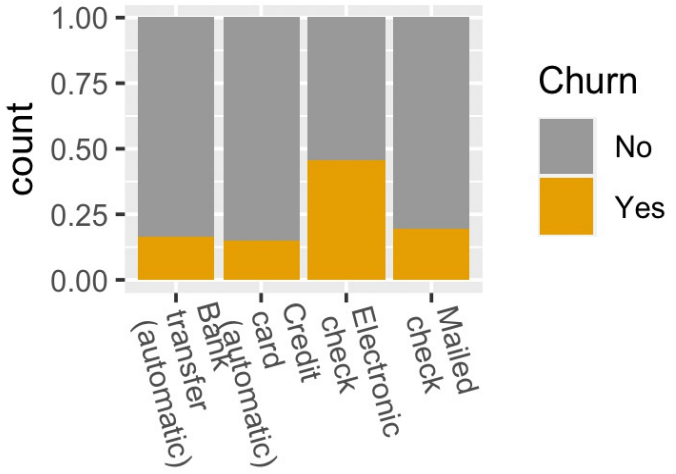
Partner



Dependents



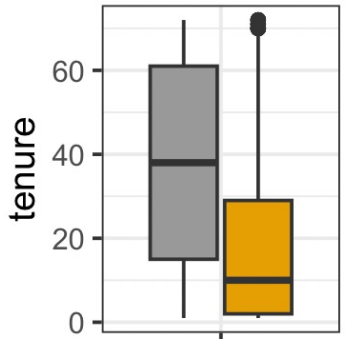
InternetService

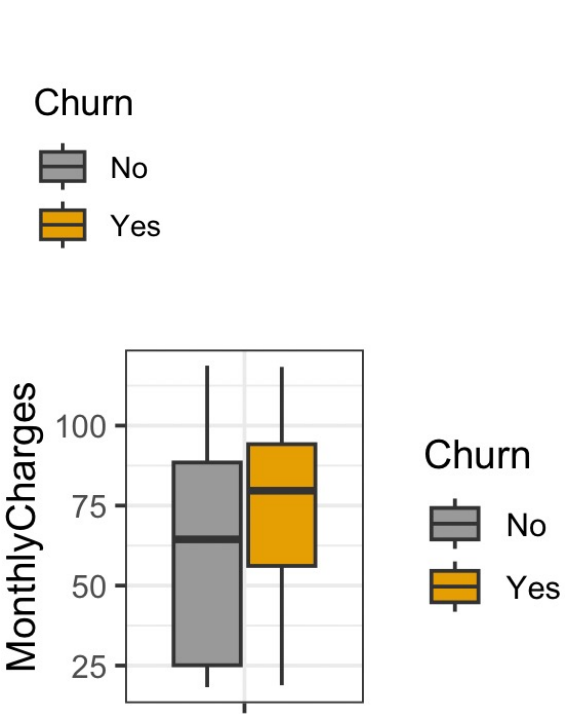


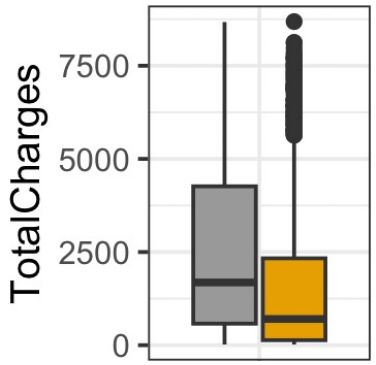
PaymentMethod

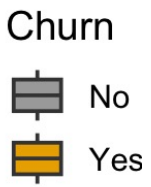
Statistiche descrittive e correlazione

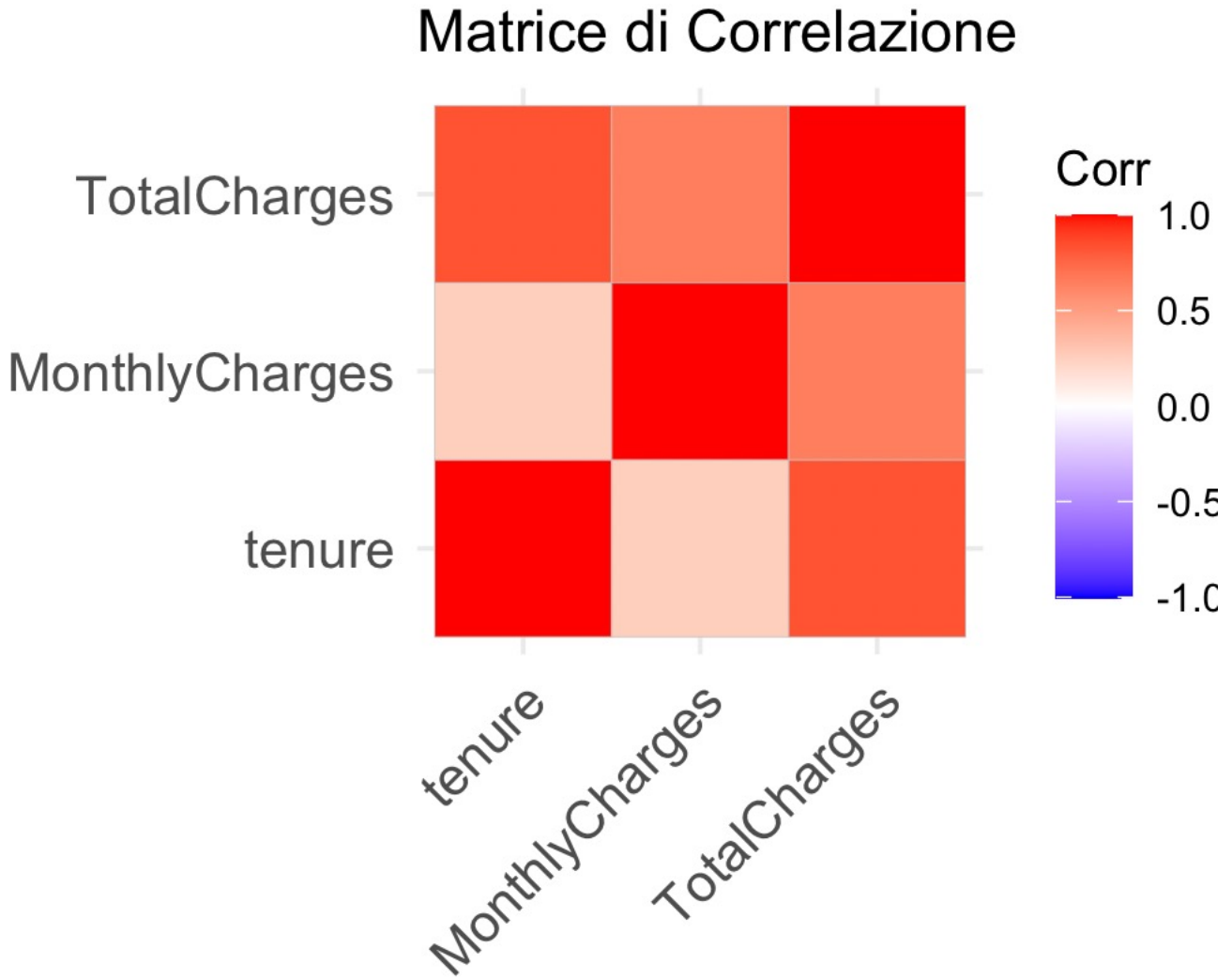
VARIABILI QUANTITATIVE











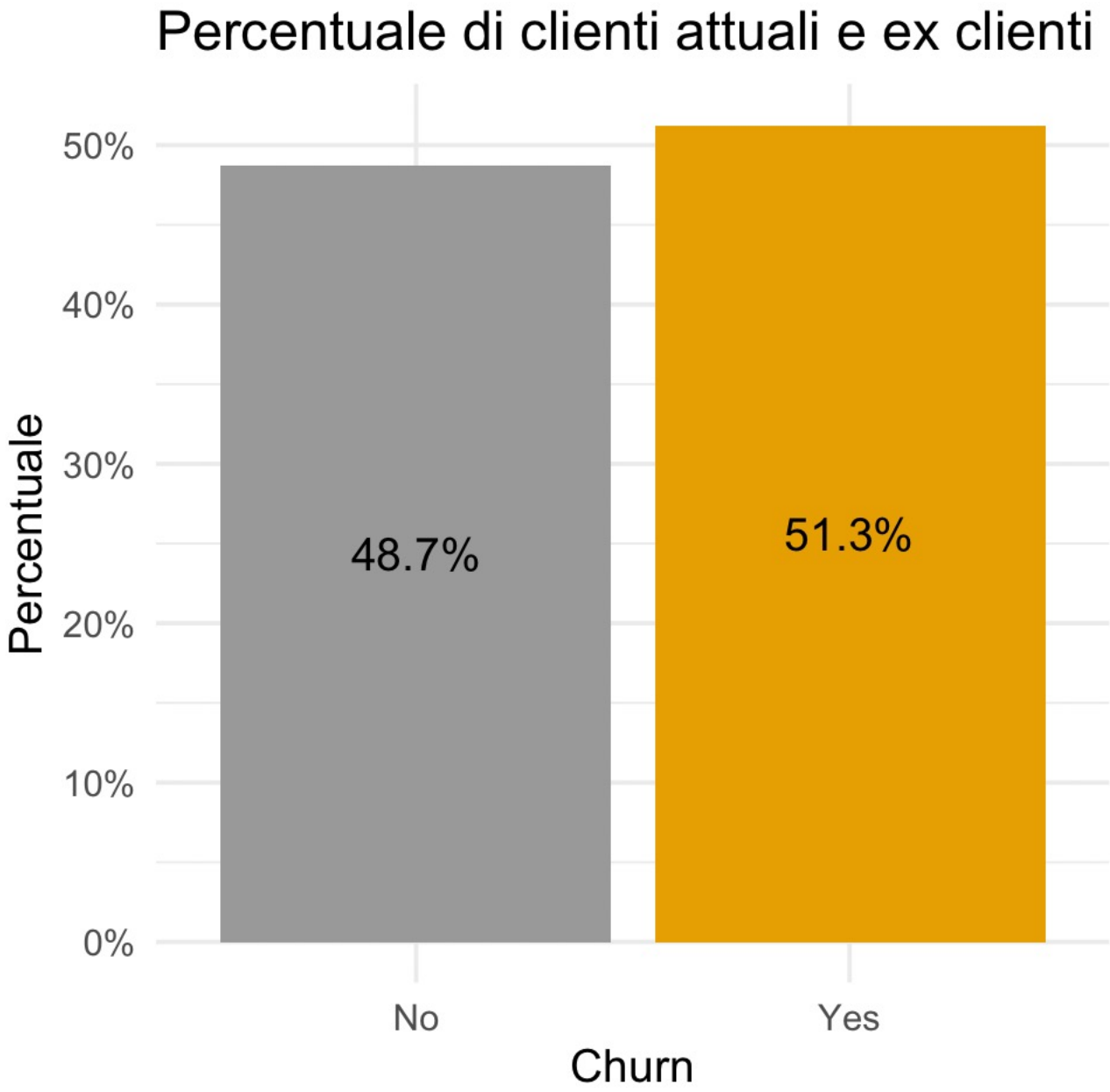
Bilanciamento del training set

Variabile dipendente sbilanciata
solamente il 27% sono ex-clienti

Divisione in training e test-set
70% training test | 30% test set

Bilanciamento variabile dipendente
Utilizzo la funzione SMOTE
per bilanciare il training set

```
> train_data = SMOTE(Churn ~ ., data = nobalanced_train_data,
perc.over = 150, perc.under = 190)
> summary(train_data$Churn=="Yes") #ora Churn è quasi perfetta
mente bilanciata (circa 49%-51%)
   Mode    FALSE     TRUE
logical     2487     2618
```



Regressione Logistica

Modello FULL

```
Call:
glm(formula = Churn ~ ., family = binomial(link = logit), data = train_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3725  -0.8694   0.4519   0.8658   2.6155

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                             0.224452   0.183957   1.220 0.222415
genderMale                             -0.058329   0.066309  -0.880 0.379046
SeniorCitizen1                          0.511887   0.085536   5.984 2.17e-09 ***
PartnerYes                              0.040571   0.072620   0.559 0.576380
DependentsYes                           0.171738   0.078593   2.185 0.028878 *
tenure                                 -1.120414   0.125321  -8.940  < 2e-16 ***
PhoneServiceYes                        -0.772875   0.125245  -6.171 6.79e-10 ***
MultipleLinesYes                        0.270782   0.077874   3.477 0.000507 ***
InternetServiceFiber optic              0.216715   0.097973   2.212 0.026967 *
InternetServiceNo                      -0.791499   0.168405  -4.700 2.60e-06 ***
OnlineSecurityYes                      -0.190516   0.075712  -2.516 0.011858 *
OnlineBackupYes                         0.102399   0.075136   1.363 0.172929
DeviceProtectionYes                     0.072415   0.076746   0.944 0.345390
TechSupportYes                          0.101295   0.077312   1.310 0.190126
StreamingTVYes                          0.206092   0.080173   2.571 0.010152 *
StreamingMoviesYes                      0.213170   0.080288   2.655 0.007929 **
ContractOne year                       -0.433317   0.089177  -4.859 1.18e-06 ***
ContractTwo year                       -0.910854   0.121311  -7.508 5.98e-14 ***
PaperlessBillingYes                     0.005172   0.071466   0.072 0.942303
PaymentMethodCredit card (automatic)    0.182838   0.105432   1.734 0.082884 .
PaymentMethodElectronic check           0.244660   0.094127   2.599 0.009343 **
PaymentMethodMailed check              -0.219966   0.107840  -2.040 0.041375 *
MonthlyCharges                          0.330635   0.112888   2.929 0.003402 **
TotalCharges                            0.154892   0.140656   1.101 0.270802
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7073.7  on 5104  degrees of freedom
Residual deviance: 5471.4  on 5081  degrees of freedom
AIC: 5519.4
```

Modello Stepwise

```
Call:
glm(formula = Churn ~ SeniorCitizen + Dependents + tenure + PhoneService +
    MultipleLines + InternetService + OnlineSecurity + OnlineBackup +
    StreamingTV + StreamingMovies + Contract + PaymentMethod +
    MonthlyCharges, family = binomial(link = logit), data = train_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3813  -0.8538   0.4448   0.8620   2.5418

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                              0.28900    0.16298   1.773 0.076192 .
SeniorCitizen1                           0.51205    0.08527   6.005 1.91e-09 ***
DependentsYes                            0.17981    0.07479   2.404 0.016202 *
tenure                                  -0.98750    0.05324 -18.548  < 2e-16 ***
PhoneServiceYes                         -0.78582    0.12395  -6.340 2.30e-10 ***
MultipleLinesYes                         0.26795    0.07748   3.458 0.000543 ***
InternetServiceFiber optic               0.18655    0.09561   1.951 0.051030 .
InternetServiceNo                       -0.75834    0.16117  -4.705 2.54e-06 ***
OnlineSecurityYes                       -0.18363    0.07540  -2.435 0.014877 *
OnlineBackupYes                          0.11268    0.07488   1.505 0.132393
StreamingTVYes                           0.20973    0.08000   2.622 0.008753 **
StreamingMoviesYes                       0.22170    0.08009   2.768 0.005636 **
ContractOne year                        -0.42347    0.08842  -4.789 1.67e-06 ***
ContractTwo year                        -0.86320    0.11791  -7.321 2.46e-13 ***
PaymentMethodCredit card (automatic)     0.17682    0.10510   1.682 0.092498 .
PaymentMethodElectronic check            0.23178    0.09377   2.472 0.013446 *
PaymentMethodMailed check               -0.21404    0.10705  -1.999 0.045565 *
MonthlyCharges                           0.43084    0.09493   4.538 5.67e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7073.7  on 5104  degrees of freedom
Residual deviance: 5476.5  on 5087  degrees of freedom
AIC: 5512.5

Number of Fisher Scoring iterations: 4
```

Modello ottimizzato per evitare multicollinearità

```
Call:
glm(formula = Churn ~ SeniorCitizen + Dependents + tenure + PhoneService +
    MultipleLines + OnlineSecurity + StreamingTV + StreamingMovies +
    Contract + PaymentMethod + MonthlyCharges, family = binomial(link = logit),
    data = train_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4474  -0.8906   0.4151   0.8771   2.4680

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                              0.58606    0.14491   4.044 5.25e-05 ***
SeniorCitizen1                           0.53104    0.08514   6.237 4.45e-10 ***
DependentsYes                            0.18668    0.07449   2.506  0.01221 *
tenure                                  -1.01256    0.05233 -19.350  < 2e-16 ***
PhoneServiceYes                         -1.06498    0.11195  -9.513  < 2e-16 ***
MultipleLinesYes                         0.22647    0.07688   2.946  0.00322 **
OnlineSecurityYes                       -0.16149    0.07420  -2.177  0.02951 *
StreamingTVYes                           0.15648    0.07913   1.978  0.04798 *
StreamingMoviesYes                       0.16385    0.07900   2.074  0.03807 *
ContractOne year                        -0.45188    0.08744  -5.168 2.37e-07 ***
ContractTwo year                        -0.92208    0.11657  -7.910 2.57e-15 ***
PaymentMethodCredit card (automatic)     0.17876    0.10491   1.704  0.08839 .
PaymentMethodElectronic check            0.25208    0.09354   2.695  0.00704 **
PaymentMethodMailed check               -0.23599    0.10631  -2.220  0.02643 *
MonthlyCharges                           0.78040    0.05872  13.291  < 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: 7073.7  on 5104  degrees of freedom
Residual deviance: 5503.2  on 5090  degrees of freedom
AIC: 5533.2

Number of Fisher Scoring iterations: 4
```

Regressione Logistica

SCELTA CUT-OFF

Utilizzo di un ciclo for per la scelta del cut-off

cut-off scelto è 0,5: si è optato per questo cut-off, nonostante non massimizzi l'accuracy, poiché riesce a garantire un buon livello di sensitività (essenziale per l'obiettivo del lavoro)

	cutoff	accuracy	specificity	sensitivity
1	0.1	0.3780835	0.1543928	0.99642857
2	0.2	0.4943074	0.3178295	0.98214286
3	0.3	0.5896584	0.4573643	0.95535714
4	0.4	0.6646110	0.5826873	0.89107143
5	0.5	0.7191651	0.6821705	0.82142857
6	0.6	0.7594877	0.7713178	0.72678571
7	0.7	0.7964896	0.8624031	0.61428571
8	0.8	0.7851044	0.9399225	0.35714286
9	0.9	0.7490512	0.9948320	0.06964286

VIF

Controllo della multicollinearità

	GVIF	Df	GVIF^(1/(2*Df))
SeniorCitizen	1.063660	1	1.031339
Dependents	1.054146	1	1.026716
tenure	1.954203	1	1.397928
PhoneService	1.230210	1	1.109148
MultipleLines	1.333432	1	1.154743
OnlineSecurity	1.134663	1	1.065205
StreamingTV	1.426967	1	1.194557
StreamingMovies	1.420298	1	1.191763
Contract	1.443686	2	1.096145
PaymentMethod	1.282938	3	1.042400
MonthlyCharges	2.395097	1	1.547610

MATRICE DI CONFUSIONE

```
          Reference
Prediction  Yes   No
       Yes  460  492
       No   100 1056

               Accuracy : 0.7192
                 95% CI : (0.6994, 0.7383)
    No Information Rate : 0.7343
    P-Value [Acc > NIR] : 0.9448

                  Kappa : 0.4116

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8214
            Specificity : 0.6822
         Pos Pred Value : 0.4832
         Neg Pred Value : 0.9135
             Prevalence : 0.2657
         Detection Rate : 0.2182
   Detection Prevalence : 0.4516
      Balanced Accuracy : 0.7518

       'Positive' Class : Yes
```

Regressione Logistica

R2

```
 McFadden
0.2220148
```

LIKELIHOOD RATIO TEST

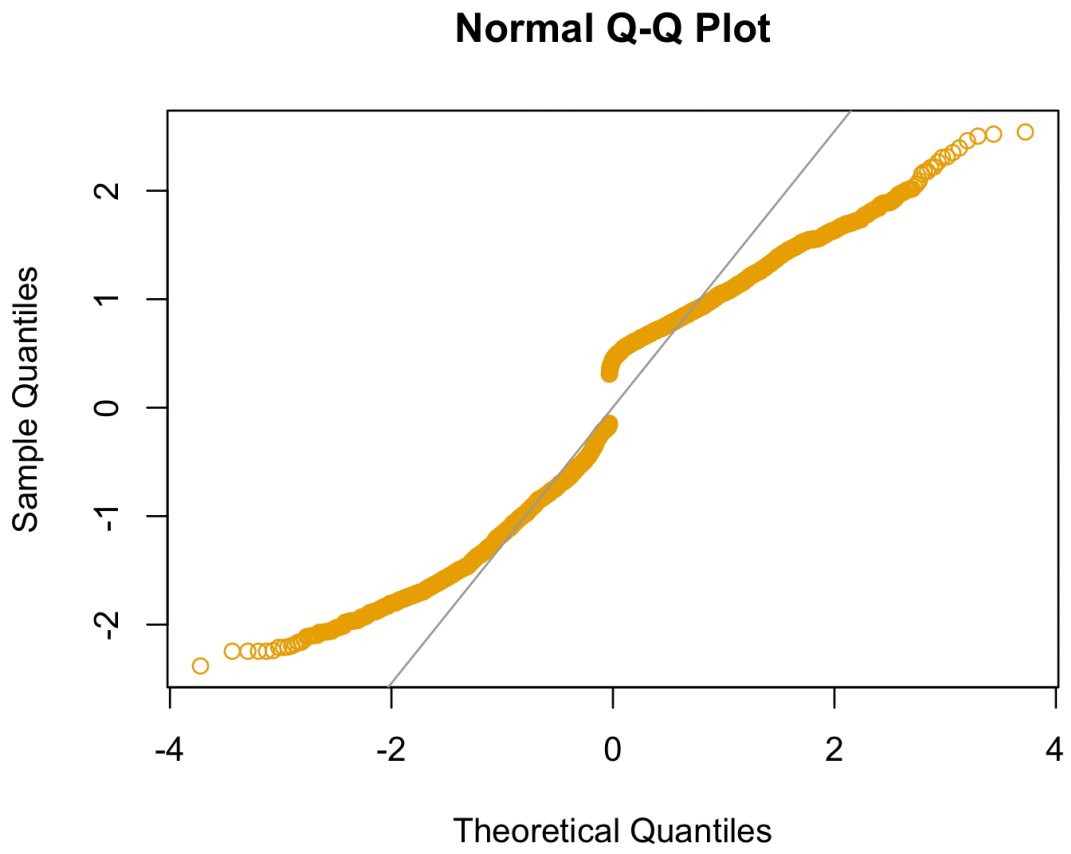
```
Model 1: Churn ~ SeniorCitizen + Dependents + tenure + PhoneService +
    MultipleLines + OnlineSecurity + StreamingTV + StreamingMovies +
    Contract + PaymentMethod + MonthlyCharges
Model 2: Churn ~ 1
  Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
1      5090     5503.2
2      5104     7073.7 -14  -1570.5 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

WALD TEST

```
Model 1: Churn ~ SeniorCitizen + Dependents + tenure + PhoneService +
    MultipleLines + OnlineSecurity + StreamingTV + StreamingMovies +
    Contract + PaymentMethod + MonthlyCharges
Model 2: Churn ~ 1
  Res.Df  Df      F    Pr(>F)
1   5090
2   5104 -14 72.905 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

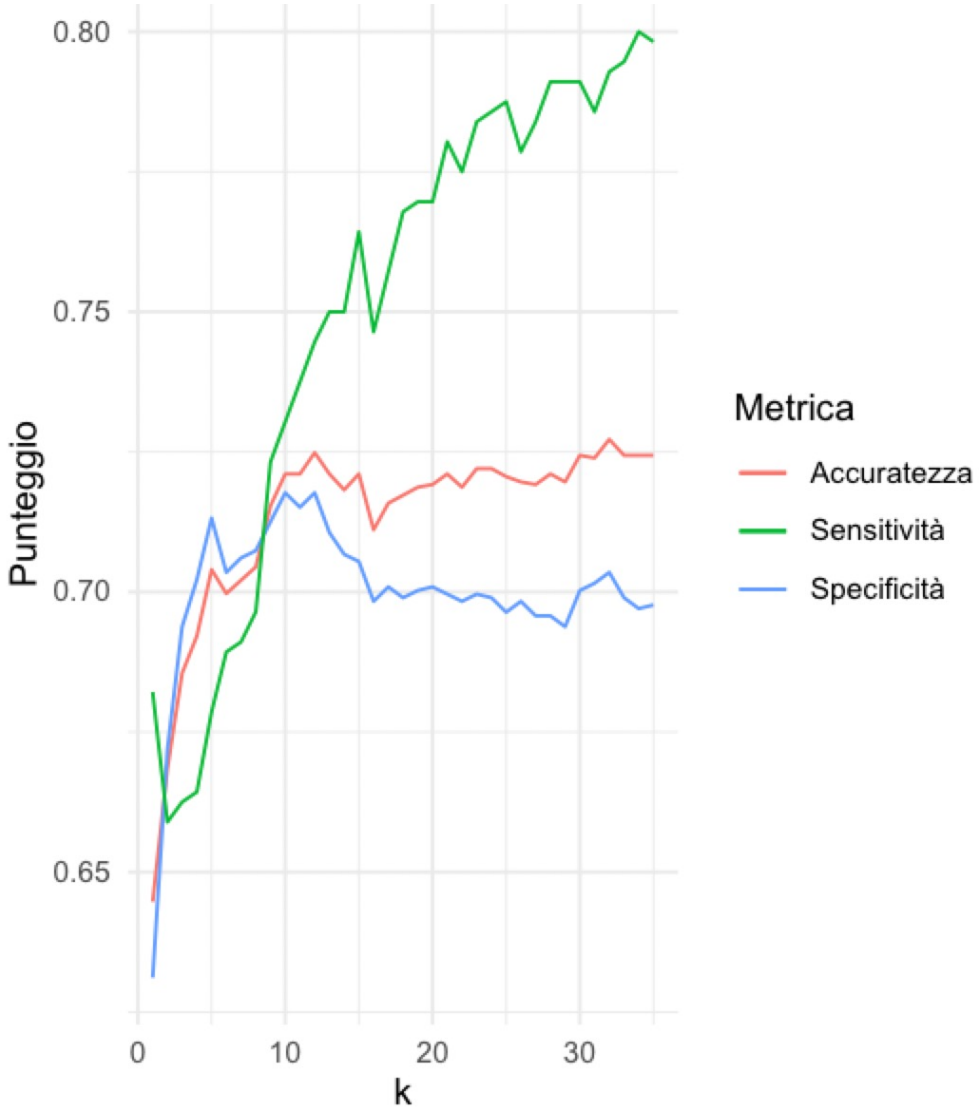
MEDIA DEI RESIDUI

```
> round(mean(residuals))
[1] 0
```

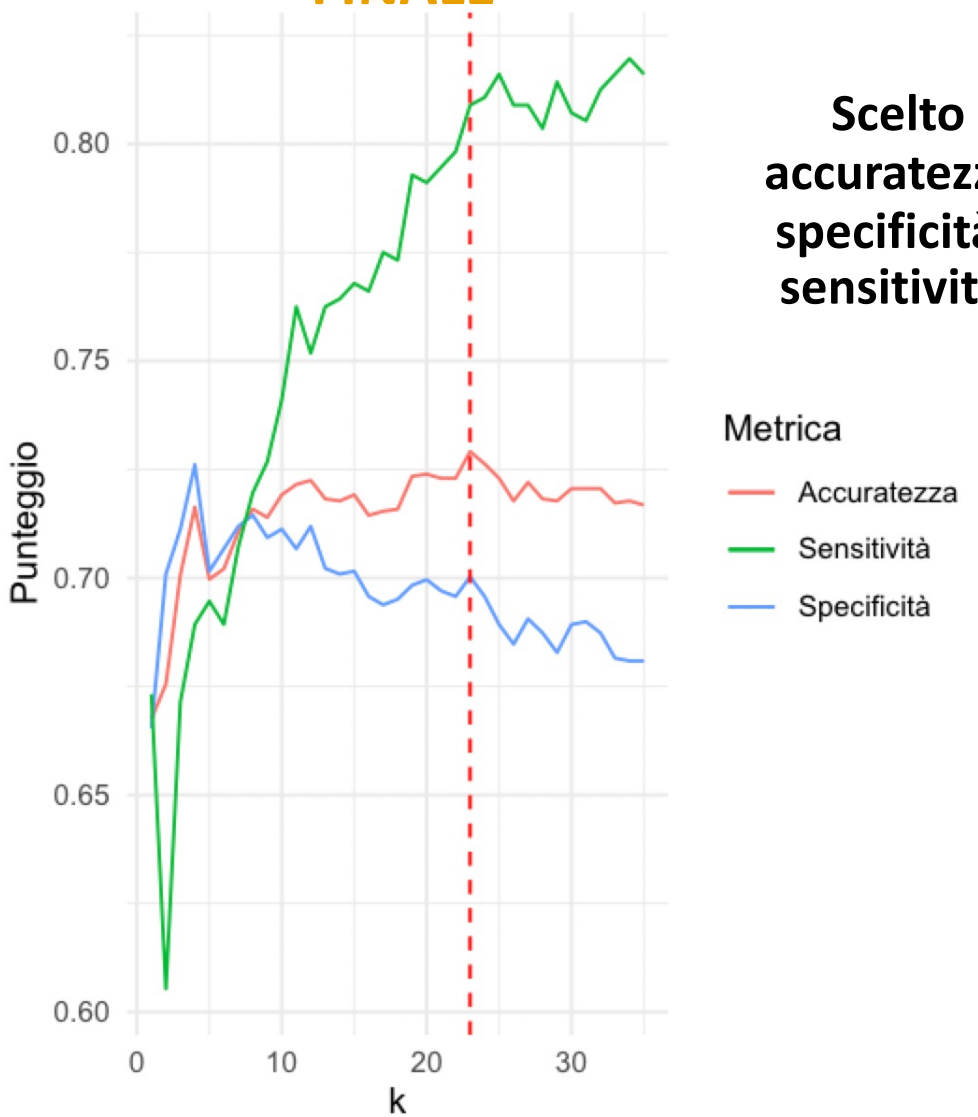


K-nn

Scelta di k in un modello FULL



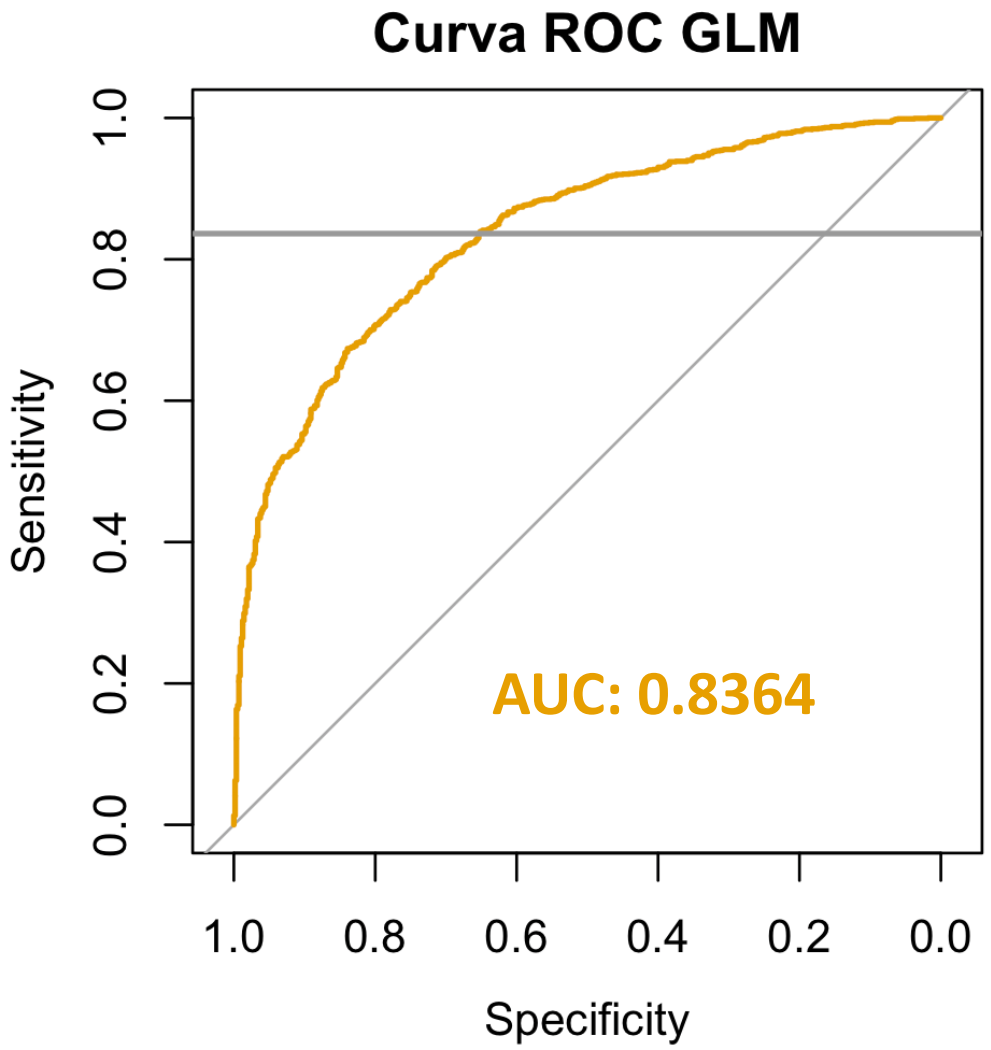
Scelta di k nel modello FINALE

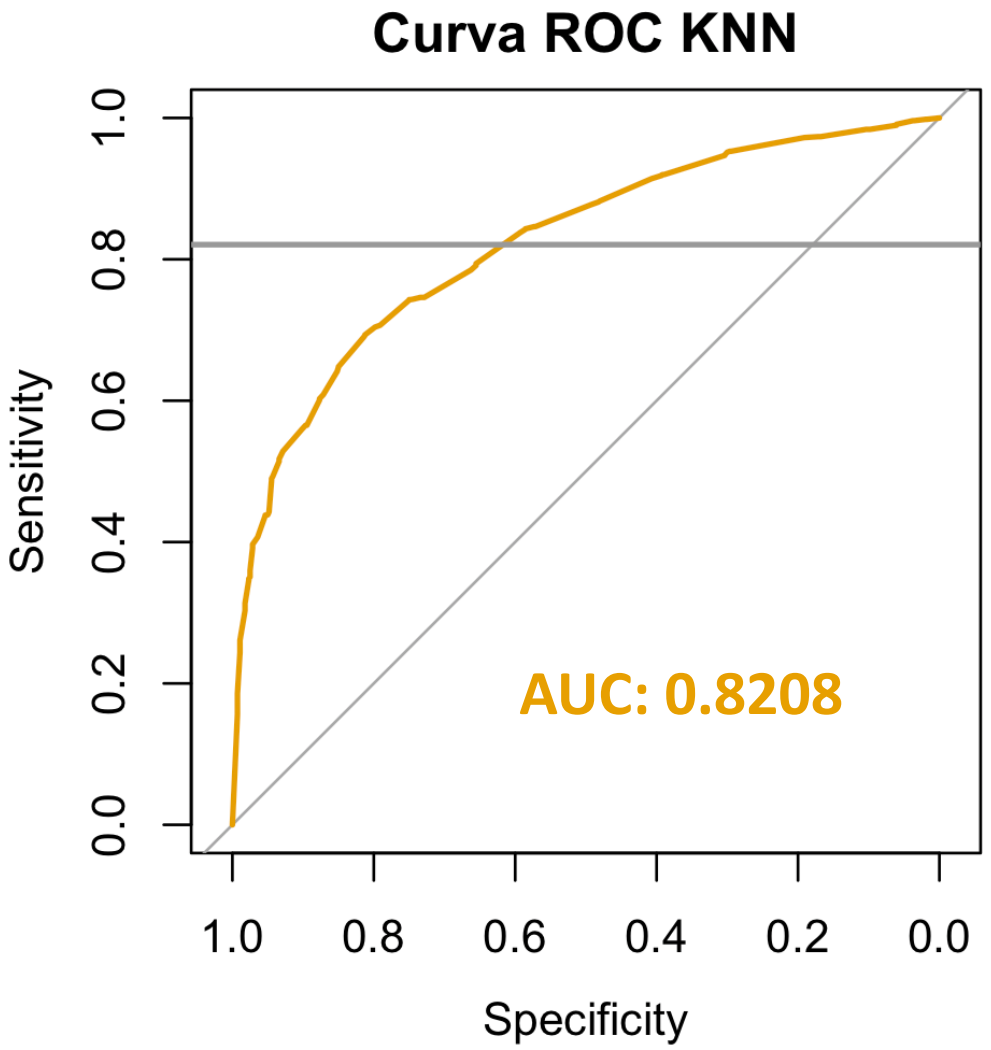


Scelto K=23
accuratezza 0.729
specificità 0.700
sensitività 0.809

Confronto K-nn e regressione logistica

Curve ROC di entrambi i modelli





Grazie

per l’attenzione!!

Manuel Manzo
(matricola: s1106217)