Customer churn prediction

Michele Puglia

2024-02-24

Contents

Introduzione

Questo report si propone di illustrare la costruzione e applicazione di modelli lineari generalizzati (GLM) attraverso un esempio pratico. I GLM rappresentano un'estensione della classica regressione lineare, adattandosi a situazioni in cui la distribuzione della variabile risposta non segue la normale (ad esempio alla famiglia ED) e consentendo l'analisi di relazioni più complesse tra le variabili rispetto al caso lineare. Il progetto analizza la problematica relativa alla perdita di clienti da parte di un istituto bancario. Per farlo, si utilizza la regressione logistica al fine di modellare la probabilità che un individuo abbandoni o rimanga cliente della banca (binaria: 0,1)

Logistic regression

Si tratta di un modello di regressione utile a modellare la probabilità di successo di una variabile di risposta binaria in funzione di variabili predittive. L'obiettivo principale è predire con precisione la probabilità con cui un'osservazione appartiene a una delle due classi.. Tale modello prevede che si possa scegliere la link function tra 3 alternative: logit, probit e clog-log. Tale scelta è cruciale per delineare la forma più adatta di relazione tra le variabili indipendenti e la probabilità di appartenenza alle classi, in base ai dati a disposizione.

Dataset bank

I dati sono stati acquisti acquisiti da Kaggle, una piattaforma online per il data science (https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset). In particolare, il dataset fa riferimento a dati di clienti dell'istituto ABC Multistate Bank, analizzando i casi di abbandono ovvero l'interruzione dell'utilizzo dei servizi offerti dalla banca e la chiusura dei propri conti. Un cliente che abbandona è una perdita per l'azienda sia in termini finanziari che reputazionali,per questo motivo è importante affrontare in modo proattivo i potenziali rischi di abbandono, sviluppando strategie di fidelizzazione mirate e migliorando la soddisfazione dei clienti. L'obiettivo principaleè comprendere e prevedere i fattori che influenzano l'abbandono dei clienti attraverso la costruzione di un modello adatto al contesto.

I dati a disposizione (10.000 osservazioni) considerano il comportamento dei clienti in relazione ad alcuni fattori che potrebbero influenzare la decisione di abbandono (churn). In particolare, si fa riferimento a 12 variabili specifiche, tra cui:

- **Credit Score**: Punteggio assegnato ad ogni cliente rispetto al proprio livello di affidabilità creditizia.
- **Country**: Paese di residenza del cliente (Germania, Francia, Spagna). Può implicare differenti normative e politiche bancarie.
- **Gender**: Genere del cliente (Femmina=1, Maschio=0).
- **Age**: Età del cliente.
- **Tenure**: Durata del rapporto tra il cliente e la banca, misurata in anni.
- **Balance**: Saldo del conto bancario del cliente, espresso in Euro.
- **Products Number**: Numero di prodotti bancari detenuti dal cliente presso la banca.
- **Credit Card**: Indicazione rispetto al possesso di una carta di credito. Un cliente può possedere o meno una carta di credito.(1=possiede, 0=non possiede)
- **Active Member**: Stato di attività del cliente. Un cliente può essere attivo (1)o inattivo (0).
- **Estimated Salary**: Stipendio stimato del cliente, espresso in Euro.
- **Churn**: Variabile binaria di risposta che indica se il cliente ha abbandonato o meno la banca (1=churn, 0=no churn).

Analisi descrittiva

Innanzitutto, è possibile visualizzare una panoramica del dataset e le principali statistiche descrittive per ciascuna variabile

```
## Rows: 10,000
## Columns: 12
## $ customer_id      <int> 15634602, 15647311, 15619304, 15701354, 15737888, 155~
## $ credit_score     <int> 619, 608, 502, 699, 850, 645, 822, 376, 501, 684, 528~
## $ country          <chr> "France", "Spain", "France", "France", "Spain", "Spai~
## $ gender           <chr> "Female", "Female", "Female", "Female", "Female", "Ma~
## $ age              <int> 42, 41, 42, 39, 43, 44, 50, 29, 44, 27, 31, 24, 34, 2~
## $ tenure           <int> 2, 1, 8, 1, 2, 8, 7, 4, 4, 2, 6, 3, 10, 5, 7, 3, 1, 9~
## $ balance          <dbl> 0.00, 83807.86, 159660.80, 0.00, 125510.82, 113755.78~
## $ products_number  <int> 1, 1, 3, 2, 1, 2, 2, 4, 2, 1, 2, 2, 2, 2, 2, 2, 1, 2,~
## $ credit_card      <int> 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1,~
## $ active_member    <int> 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1,~
## $ estimated_salary <dbl> 101348.88, 112542.58, 113931.57, 93826.63, 79084.10, ~
## $ churn            <int> 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,~
```

```
##   customer_id        credit_score     country             gender
##  Min.   :15565701   Min.   :350.0   Length:10000       Length:10000
##  1st Qu.:15628528   1st Qu.:584.0   Class :character   Class :character
##  Median :15690738   Median :652.0   Mode  :character   Mode  :character
##  Mean   :15690941   Mean   :650.5
##  3rd Qu.:15753234   3rd Qu.:718.0
##  Max.   :15815690   Max.   :850.0
##       age            tenure          balance       products_number
##  Min.   :18.00   Min.   : 0.000   Min.   :     0   Min.   :1.00
##  1st Qu.:32.00   1st Qu.: 3.000   1st Qu.:     0   1st Qu.:1.00
##  Median :37.00   Median : 5.000   Median : 97199   Median :1.00
##  Mean   :38.92   Mean   : 5.013   Mean   : 76486   Mean   :1.53
##  3rd Qu.:44.00   3rd Qu.: 7.000   3rd Qu.:127644   3rd Qu.:2.00
##  Max.   :92.00   Max.   :10.000   Max.   :250898   Max.   :4.00
##   credit_card     active_member    estimated_salary        churn
##  Min.   :0.0000   Min.   :0.0000   Min.   :    11.58   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 51002.11   1st Qu.:0.0000
##  Median :1.0000   Median :1.0000   Median :100193.91   Median :0.0000
##  Mean   :0.7055   Mean   :0.5151   Mean   :100090.24   Mean   :0.2037
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:149388.25   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :199992.48   Max.   :1.0000
```

Pre-processing

Questa fase è fondamentale e richiedel'esecuzione di diverse operazioni al fine di preparare e pulire i dati per l'analisi successiva. In particolare, è processo che mira a migliorare la qualità e l'adeguatezza dei dati.

Conversione di tipo

Le variabili churn, country, gender, active_member e credit_card vengono convertite in factor per essere considerate come delle variabili categoriali. Queste trasformazioni possono rendere più chiara la rappresentazione delle variabili nel modello, migliorando l'interpretazione e ottimizzando le prestazioni

Controllo di eventuali valori mancanti

Avere dei dati completi è fondamentale per riuscire ad eseguire analisi e modellazioni accurate. I valori mancanti potrebbero introdurre delle distorsioni nei dati e quindi influire negativamente sulle prestazioni del modello.

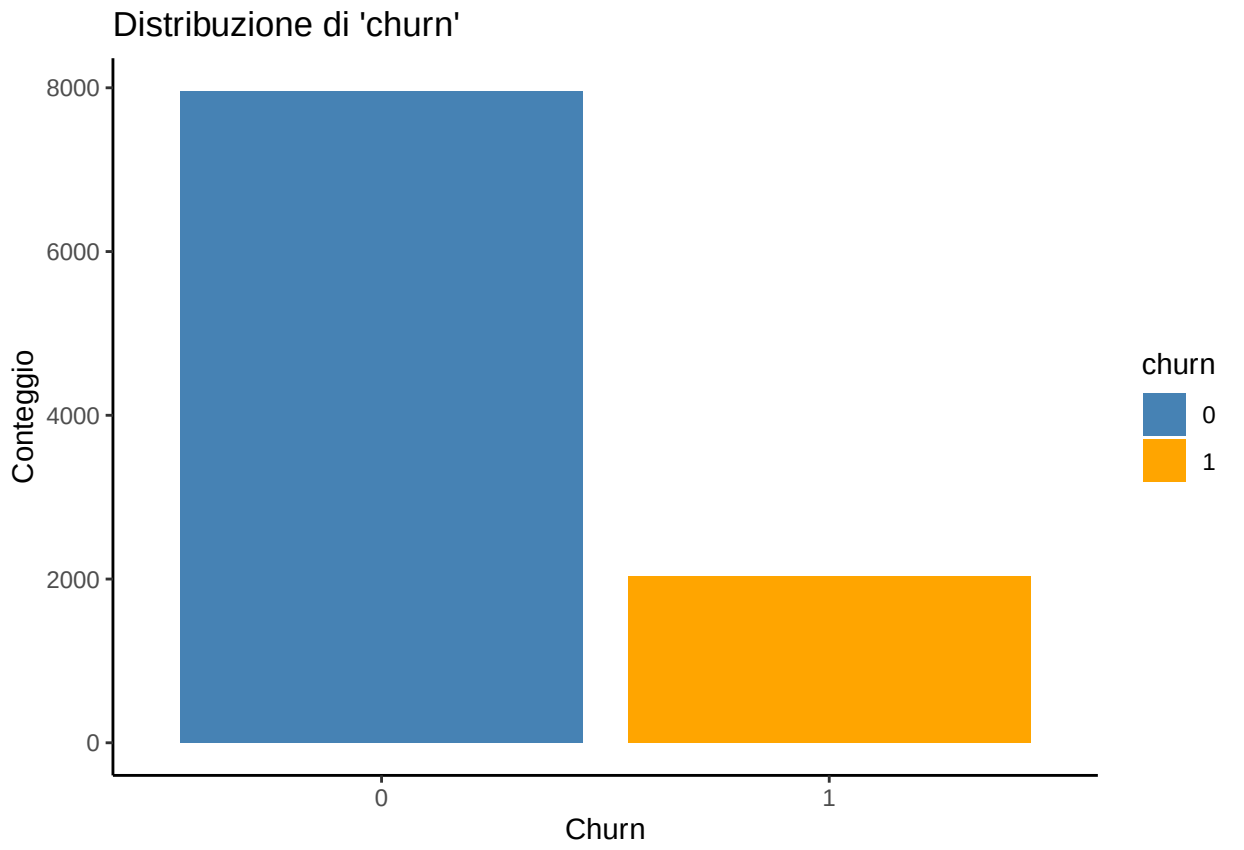
```
## [1] FALSE
```

Data visualization

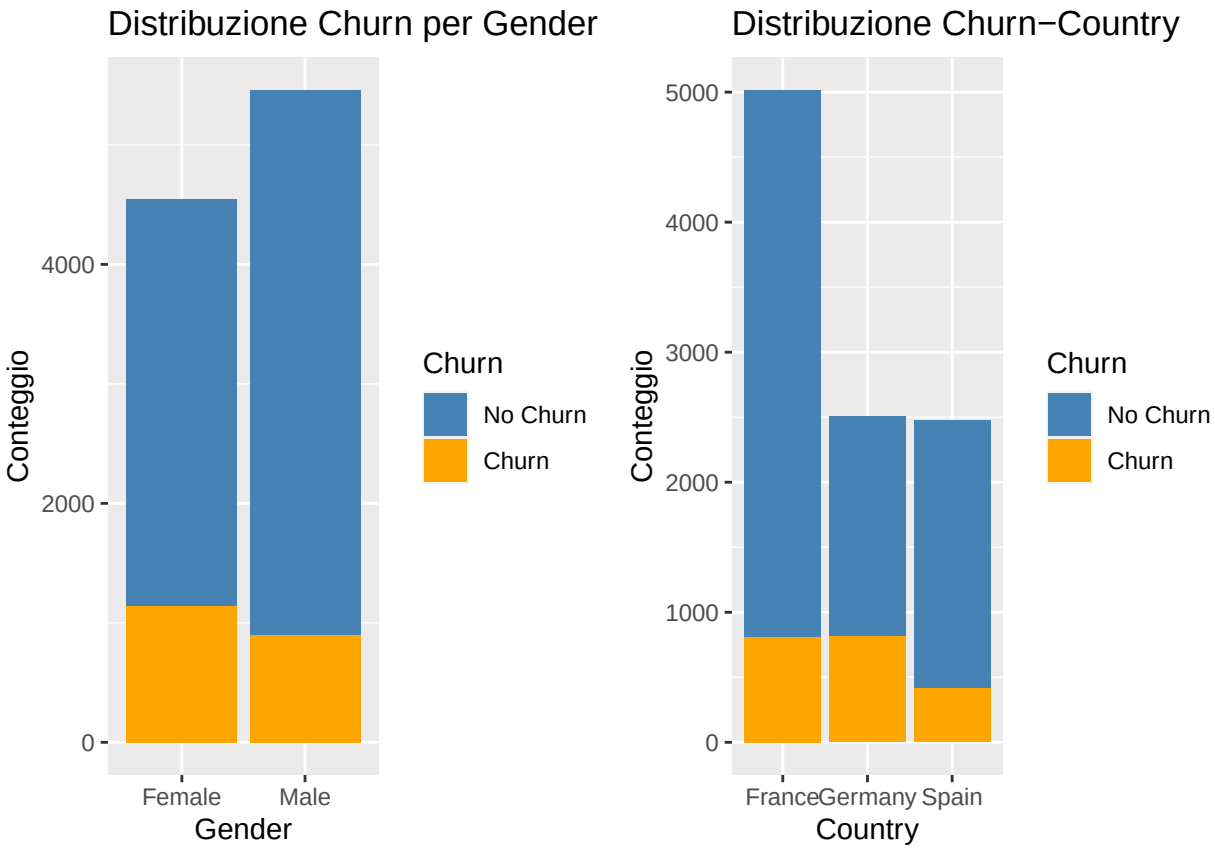
A questo punto si considera la rappresentazione grafica di alcune variabili di interesse al fine di esaminare le distribuzione di tali variabili, individuare pattern e identificare eventuali anomalie.

- **churn**: Innanzitutto, considero la variabile di risposta churn. Il barplot rappresenta la frequenza delle due classi ed indica che la maggior parte dei dati riguardano casi in cui il cliente non abbandona (circa l'80%).

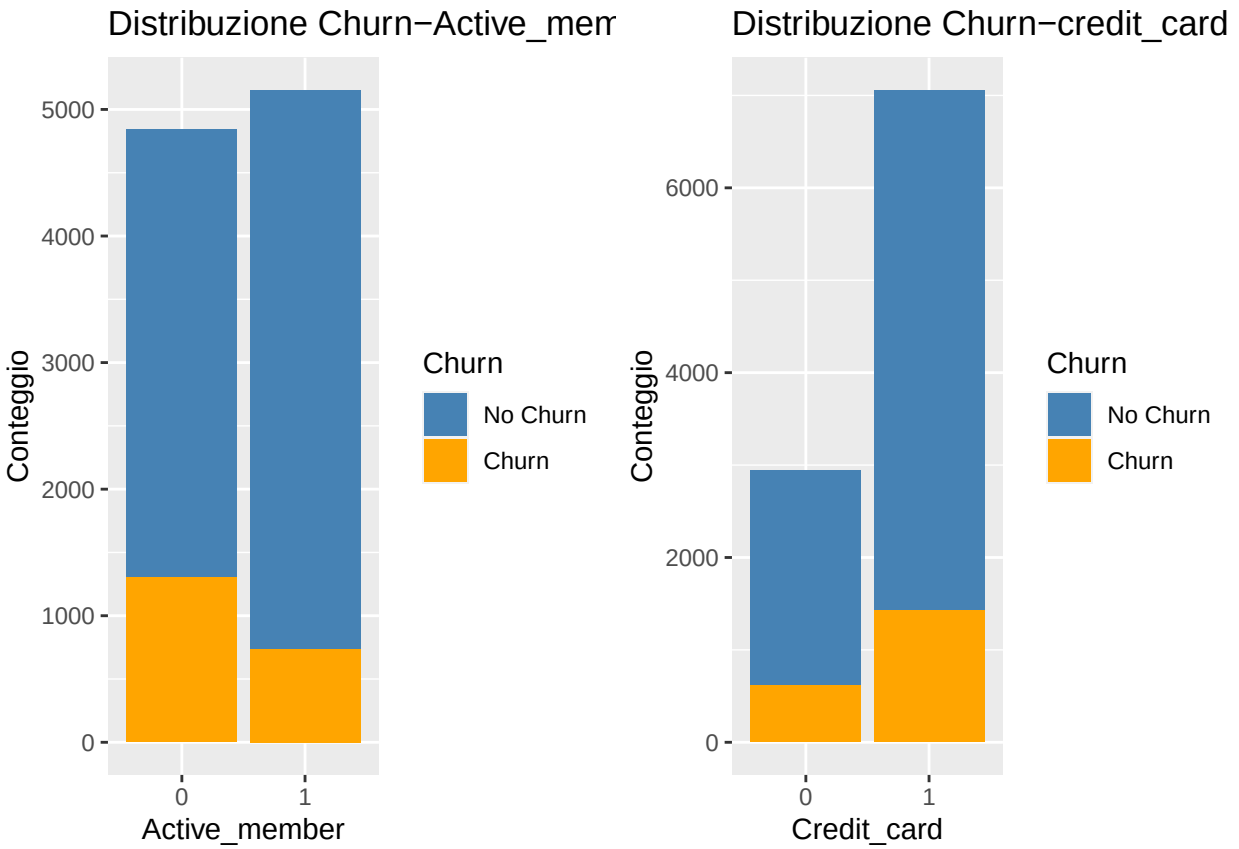
```
##   churn    n proportion
## 1     0 7963     0.7963
## 2     1 2037     0.2037
```



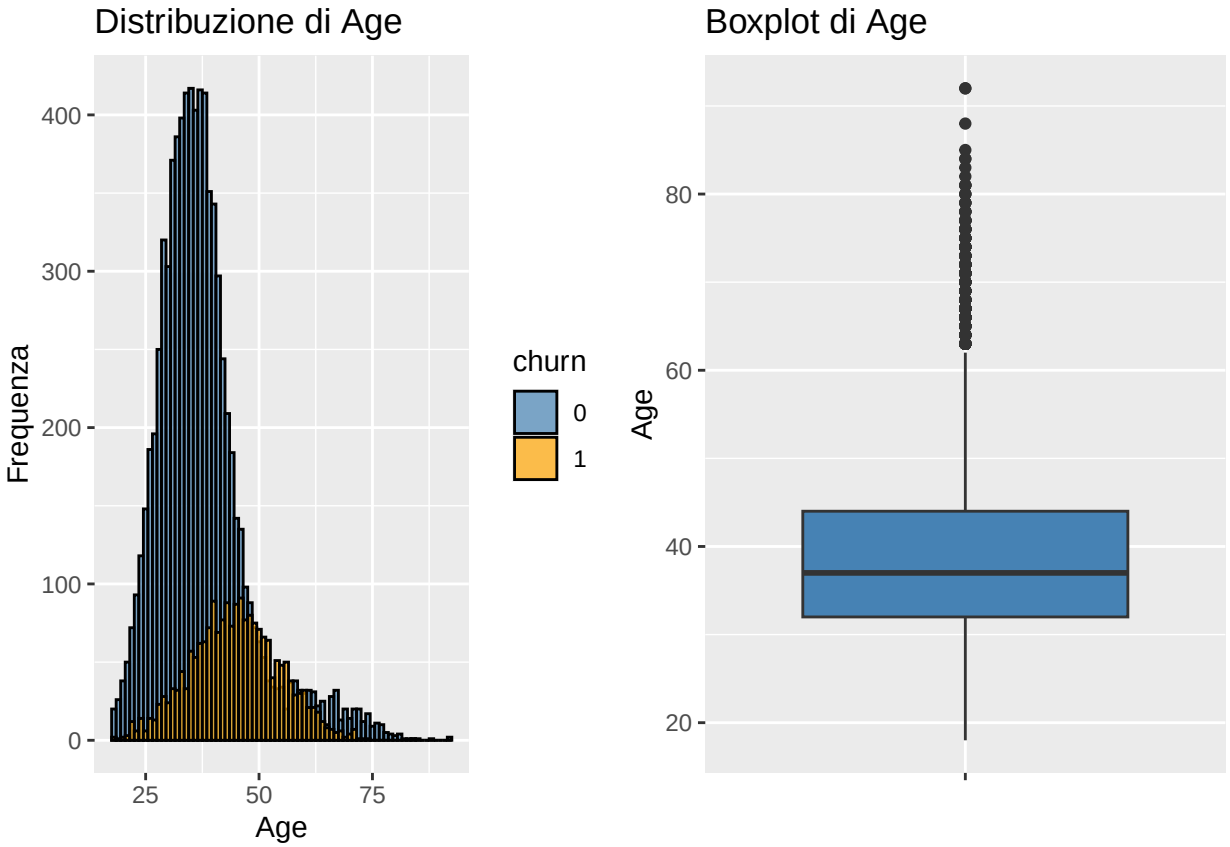
- **gender**: Le osservazioni sono costituite da un 55% di rappresentanza maschile e un 45% femminile. Si può notare una propensione di churn leggermente maggiore per le donne pari al 25% contro il 16% degli uomini.
- **country**: la maggior parte delle informazioni riguarda clienti francesi (circa il 50%), mentre spagnoli e tedeschi rappresentano entrambi una porzione del 25%. Inoltre, i clienti che vivono in Germania sembrano essere i più propensi ad abbandonare l'istituto bancario con una probabilità pari al 32%



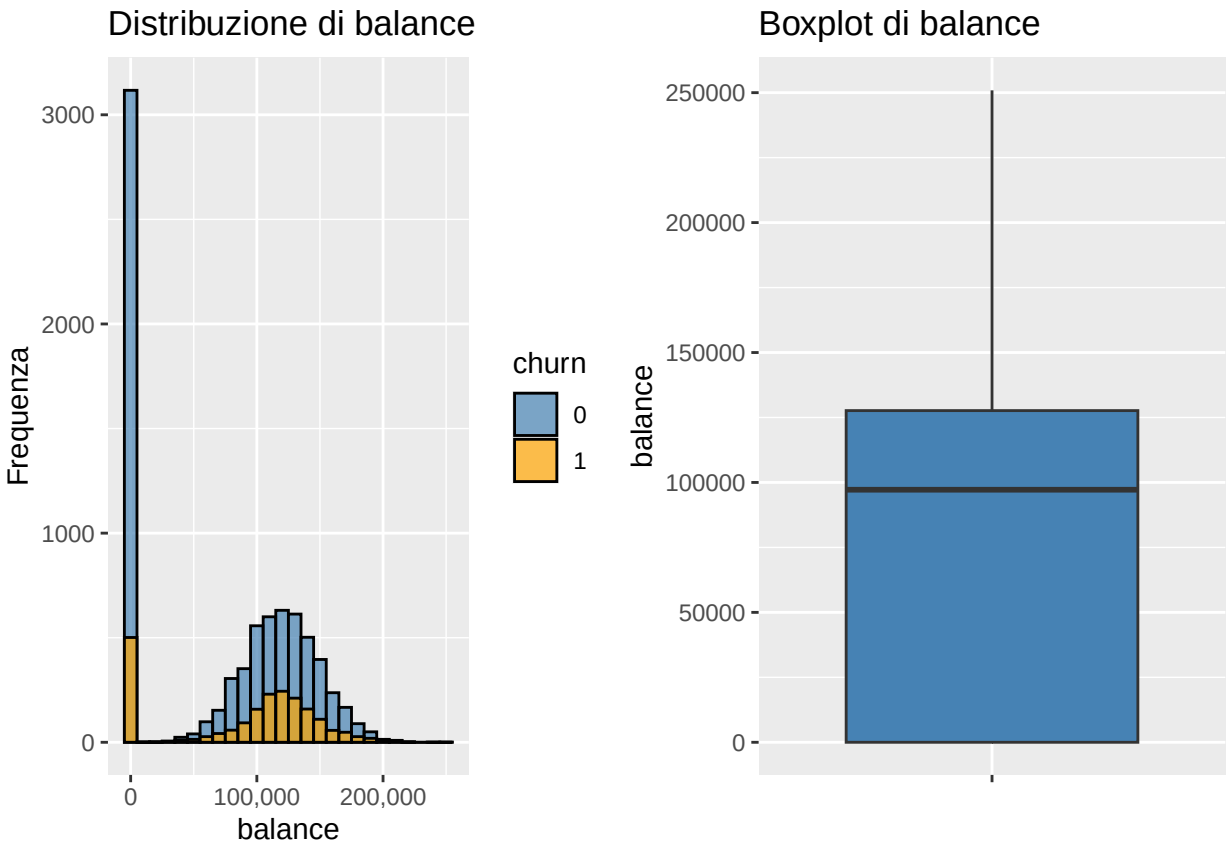
- **active_member**: La percentuale di clienti attivi è del 52% contro il 48% degli inattivi.In termini di churn si può notare come il 27% dei clienti inattivi tendano ad abbandonare il proprio istituto bancario contro il 16% di quelli attivi.
- **credit_card**: Il 70% di clienti possiede una carta di credito mentre il 30% ne è sprovvisto. In termini di churn, le probabilità di abbandono sono pari al 20% in entrambi i casi.



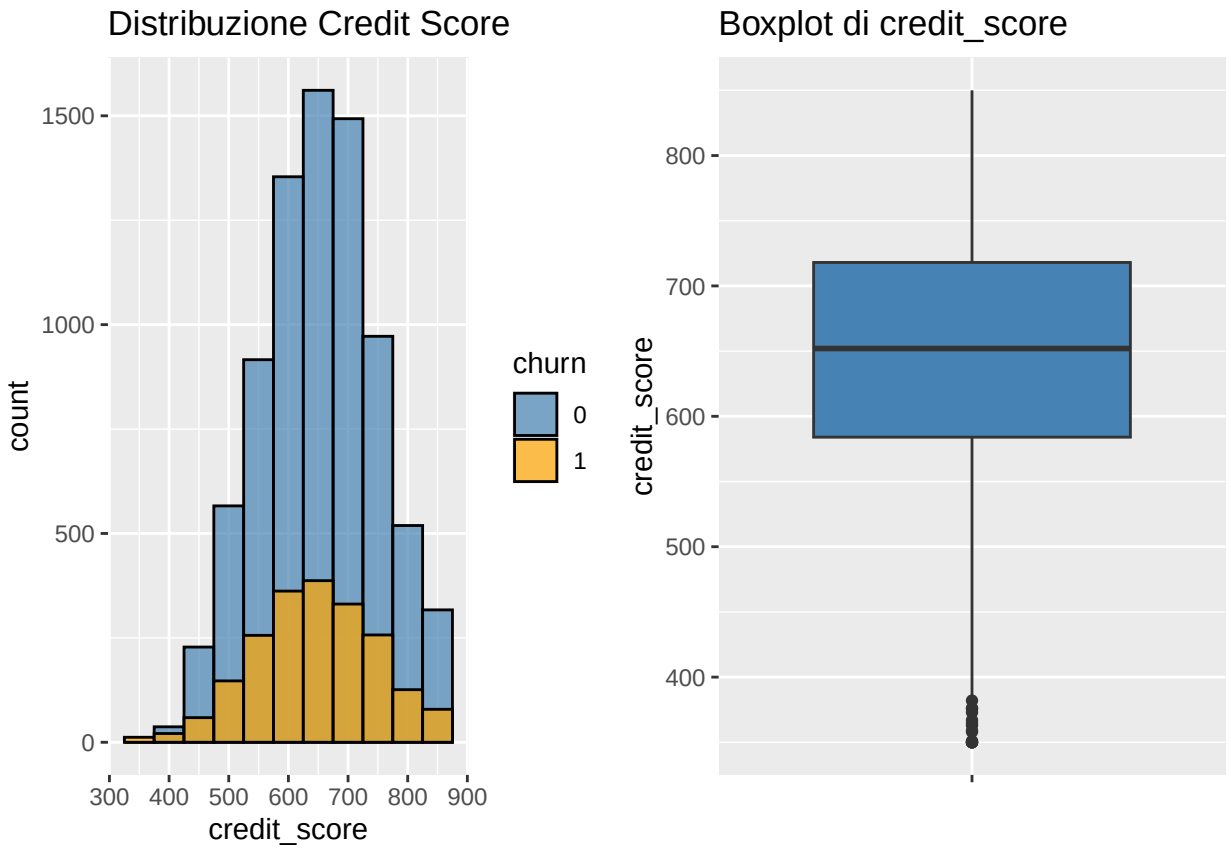
- **age**: L'istogramma della variabile age indica una distribuzione asimmetrica a destra con molti valori in corrispondenza di età giovani(tra i 30 e i 45 anni).Infatti,nel relativo boxplot si può notare come la mediana si posizioni in corrispondenza dei 37 anni e quindi più vicina al primo quartile rispetto al terzo. Inoltre, sono presenti dei potenziali outliers in corrispondenza di valori di età superiori ai 60 anni. Tali valori non corrispondono ad errori o anomalie ma riflettono la presenza di valori che si discostano significativamente dalla massa centrale dei dati. Infatti, i clienti con un'età avanzata rappresentano un sottogruppo di modeste dimensioni (359 ossservazioni) rispetto alla maggioranza dei clienti. Considerando la scelta di churn, si può notare che i clienti più anziani (tra i 40 e i 70 anni) riflettono una maggiore propensione all'abbandono rispetto ai clienti più giovani.



- **balance**: Per la variabile balance si può notare la presenza di molti dati in corrispondenza di valore 0 (conti bancari vuoti), questo può essere spiegato dal fatto che tali saldi siano relativi a conti chiusi oppure aperti di recente Inoltre, si evidenzia una propensione all'abbandono per i clienti con saldi maggiori a 85.000 e questo potrebbe essere dovuto a condizioni più favorevoli in altre banche, ad esempio la ricezione di offerte di interessi di risparmio più alti in altri istituti bancari.

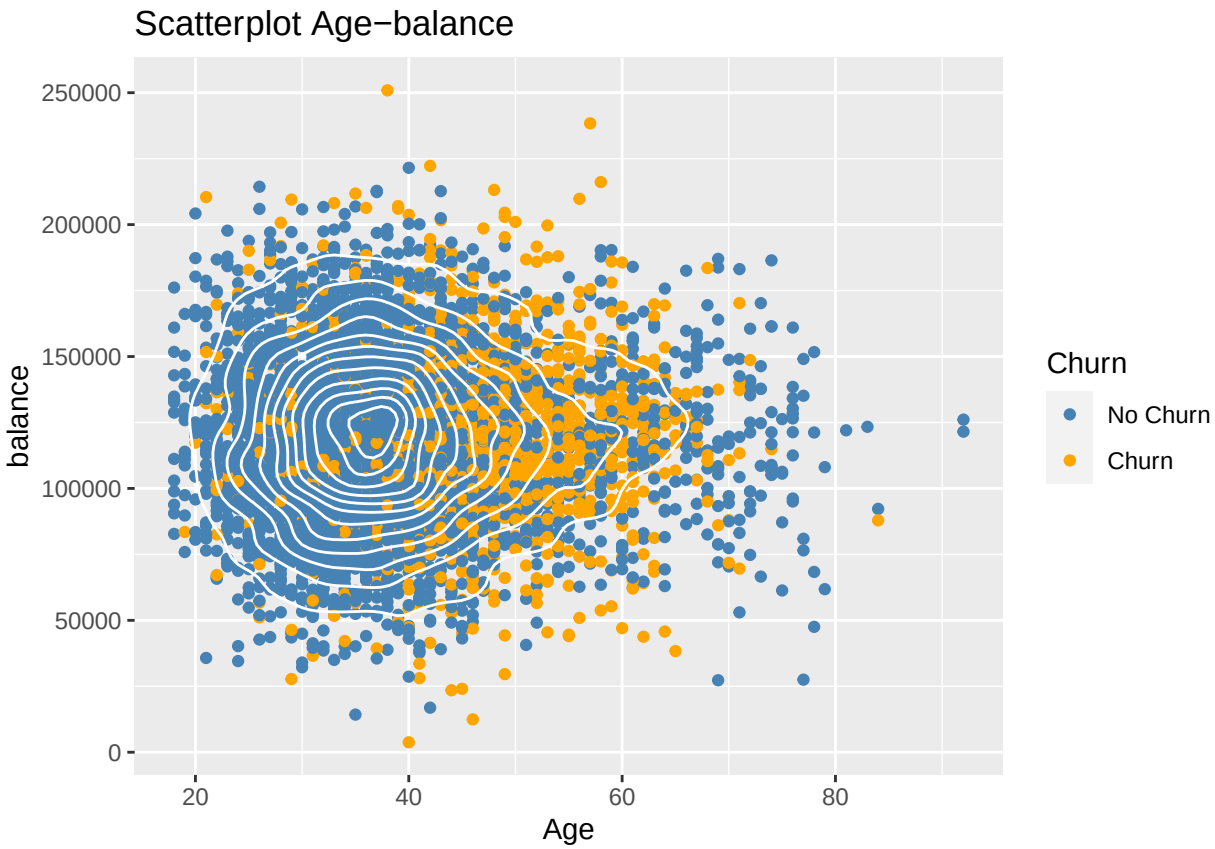


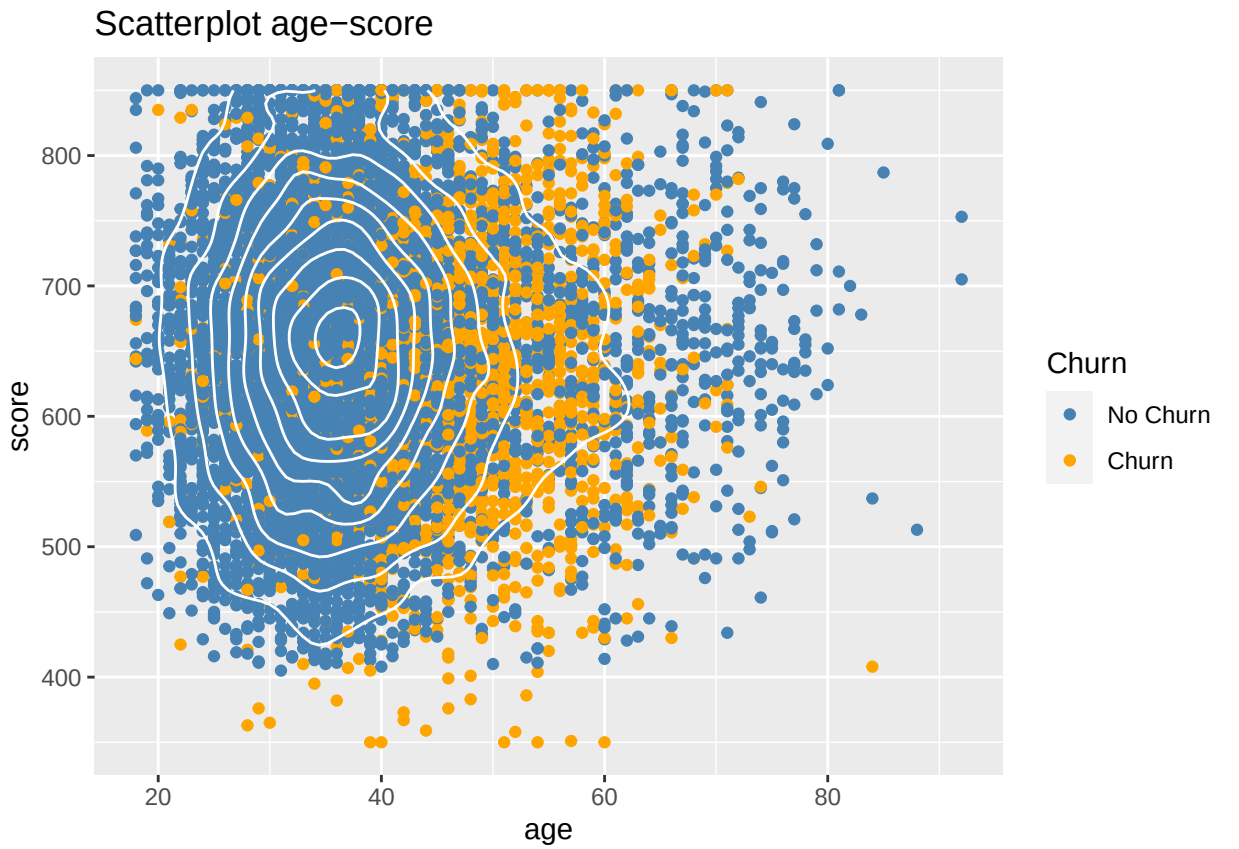
- **credit_score**: la distribuzione dello score mostra una leggera asimmetria a sinistra, la mediana si posizione in corrispondenza del valore 652 mentre la maggior parte delle osservazioni è compresa tra 584 e 718. Sono presenti dei potenziali outliers in corrispondenza di valori molto bassi di score, in particolare al di sotto del valore 400, dovuti alla presenza di pochi clienti con un basso livello di affidabilità creditizia (solo 15 osservazioni). Considerando le sclete di churn, il credit_score non sembrerebbe incidire notevolmente sulla decisione di clienti ma si può notare una propensione all'abbandono molto elevata per i casi di punteggio al di sotto di 400.



Gli scatterplot (o grafici a dispersione) sono una tipologia di grafici utilizzati per visualizzare la relazione tra due variabili quantitativo in modo da visualizzare il grado di correlazione tra di esse. In questo report sono stati considerati due casi:

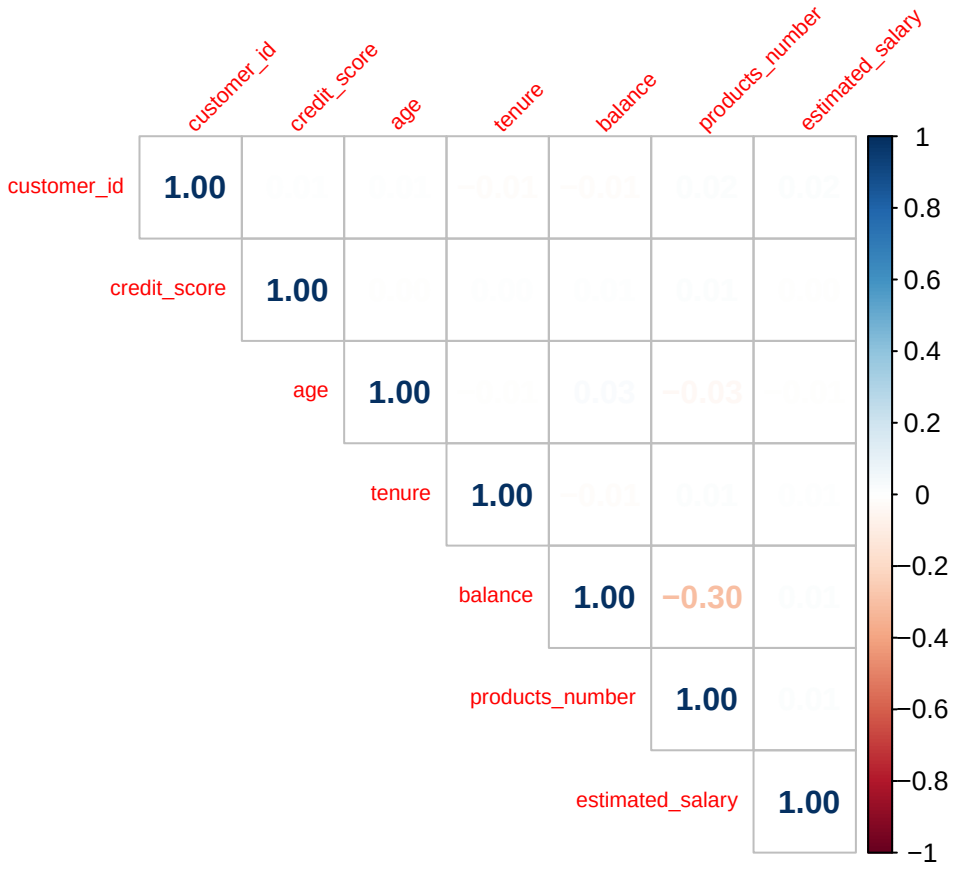
- **Age-balance**: ha lo scopo di trovare eventuali relazioni tra l'età dei clienti e i rispettivi saldi. Le linee di densità indicano una concentrazione significativa intorno ai 30-40 anni e ai saldi tra 50.000 e 200.000 ma in generale non è presente una particolare relazione tra le variabili. In base al colore dei punti si distinguono i casi di churn (arancione) dai casi di No churn (blue) e si può notare come la maggiore concentrazione di churn sia in corrispondenza di età superiori ai 40 anni e saldi superiori agli 80.000. Nel garfico, sono stati esclusi i valori di balance pari a 0 poichè riferendosi a conti appena aperti o chiusi non sono rilevanti ai fini dell'analisi.
- **age-credit_score**: in questo caso si vuole analizzare l'età rispetto ai credit_score assegnati dalla banca.Anche in queso caso, il grafico non mostra una particolare relazione tra le due variabili. Si può notare una concentrazione significativa di punti per età comprese tra i 30 e 40 anni e score tra 600 e 700. In termini di churn, invence, mentre la variabile di credit_score non sembra indicare una particolare influenza, sono presenti molti casi di abbandono per età superiori ai 40 anni.


Scatterplot Age−balance
250000
200000
150000
100000
50000
0
balance
20
40
60
80
Age
Churn
No Churn
Churn



Analisi della correlazione

Misura del grado di relazione lineare tra le variabili presenti nel dataset. Dall'analisi risulta che non sono presenti valori particolarmente elevati di correlazione tra le variabili. Il valore più alto, pari a -0.3, è associato alla relazione tra la variabile balance e la variabile products_number. Tale risultato indica una correlazione negativa tra le 2 variabili e suggerisce che i clienti con saldi più alti tendono a detenere meno prodotti bancari.



Splitting data

I dati vengono separati in train e test set. Il train set verrà utilizzato per addestrare il modello e apprendere le relazioni utili ad effettuare le successive previsioni. Il test set invece è utile per valutare le prestazioni del modello una volta addestrato.

Il train set conterrà il 70% dei dati

```
## [1] 7001   12
```

Il test set conterrà il 30% dei dati

```
## [1] 2999   12
```

Logit models

Costruzione del modello GLM. In questo caso la variabile dipendente è binaria e quindi occorre una regressione logistica binaria con cui modellare la probabilità di successo della variabile di risposta. Inizialmente, la funzione di collegamento (link function) scelta è il logit indicato come il logaritmo degli odds e quindi una funzione che permette di trasformare la probabilità di successo in un'equazione lineare (mappa la probabilita 0,1 su una scala continua tra meno infinito e più infinito). Il modello risultante presenta diverse informazioni:

- **stima dei coefficienti**: Permette di comprendere come i cambiamenti nelle variabili predittive sono associati a cambiamenti nei log-odds della variabile di risposta.Ad esempio, il coefficiente per age è

7.254e-02, indicando che un aumento di un'unità in age è associato a un incremento di 0.07254 nei log-odds di "churn". Al contrario, essere un "Active member" è associato a una diminuzione di 1.089 nei log-odds della variabile di risposta "churn"

- **significatività**: Per ciascuna variabile è specificato un p-value che permette di indicare la significatività statistica ovvero quanto bene ciascuna variabile predrittice è in grado di prevedere il valore della variabile risposta nel modello. Nel caso di age, ad esempio, si ha un livello di significatività molto elevato mentre per la variabile tenure il livello di significatività è moderato.
- **Null deviance e Residual deviance**: La Null Deviance indica quanto bene la variabile risposta possa essere prevista dal modello nullo (solo intercetta). La residual deviance invece mostra quanto bene la variabile risposta possa essere prevista dal modello con p variabili predittive. Valori di devianza bassi indicano una capacità di previsione migliore. In questo caso, la residual risulta minore della null deviance il che suggerisce che il modello, con le variabili predittive incluse, spiega una parte significativa della variabilità di churn.

```
##
## Call:
## glm(formula = churn ~ ., family = binomial(link = "logit"), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3573  -0.6545  -0.4521  -0.2625   2.9978
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -7.630e+00  7.098e+00  -1.075   0.2824
## customer_id       2.665e-07  4.521e-07   0.590   0.5555
## credit_score     -4.911e-04  3.386e-04  -1.450   0.1470
## countryGermany    8.328e-01  8.118e-02  10.259  < 2e-16 ***
## countrySpain      4.696e-02  8.490e-02   0.553   0.5802
## genderMale       -5.637e-01  6.538e-02  -8.622  < 2e-16 ***
## age               7.254e-02  3.096e-03  23.429  < 2e-16 ***
## tenure           -2.197e-02  1.120e-02  -1.961   0.0499 *
## balance           2.665e-06  6.178e-07   4.314 1.61e-05 ***
## products_number  -1.051e-01  5.633e-02  -1.867   0.0620 .
## credit_card1     -1.408e-02  7.157e-02  -0.197   0.8440
## active_member1   -1.089e+00  6.896e-02 -15.798  < 2e-16 ***
## estimated_salary  8.517e-08  5.685e-07   0.150   0.8809
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7077.6  on 7000  degrees of freedom
## Residual deviance: 5955.0  on 6988  degrees of freedom
## AIC: 5981
##
## Number of Fisher Scoring iterations: 5
```

In base ai risultati ottenuti con il primo modello è possibile costruire un modello alternativo che tenga conto esclusivamente delle variabili significative ovvero quelle per cui il p-value risulta minore del livello 0.05. In particolare, sono state selezionate le variabili: country, gender, age, tenure, balance e active_member. Dal confronto tra residual e null deviance possiamo affermare che anche questo modello riesce a spiegare una parte significativa della variabilità della variabile churn.

```
##
## Call:
## glm(formula = churn ~ country + gender + age + tenure + balance +
##     active_member, family = binomial(link = "logit"), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3648  -0.6549  -0.4509  -0.2630   2.9892
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.957e+00  1.560e-01 -25.362  < 2e-16 ***
## countryGermany  8.157e-01  8.069e-02  10.110  < 2e-16 ***
## countrySpain    4.664e-02  8.488e-02   0.550   0.5827
## genderMale     -5.612e-01  6.532e-02  -8.591  < 2e-16 ***
## age             7.280e-02  3.094e-03  23.530  < 2e-16 ***
## tenure         -2.204e-02  1.120e-02  -1.968   0.0491 *
## balance         2.972e-06  5.948e-07   4.996 5.85e-07 ***
## active_member1 -1.094e+00  6.889e-02 -15.874  < 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: 7077.6  on 7000  degrees of freedom
## Residual deviance: 5961.1  on 6993  degrees of freedom
## AIC: 5977.1
##
## Number of Fisher Scoring iterations: 5
```

A questo punto, avendo a disposizione un modello più complesso (bank.fit) ed uno più semplice (bank.fit2), è possibile confrontare la bontà di adattamento di tali modelli utilizzando il test LRT (likelihood ratio test).Questo test confronta la massima verosimiglianza dei modelli e cerca di determinare se l'aggiunta di parametri (modello più complesso) porta ad un miglioramento significativo nella verosimiglianza rispetto al modello più semplice. La statistica di test segue approssimativamente una distribuzione chi-quadro con un numero di gradi di libertà pari alla differenza nel numero di parametri tra i due modelli. Il test è costituito da un'ipotesi nulla H0: i modelli sono equivalenti e dall'ipotesi alternativa H1: i modelli sono diversi. La decisione scaturisce dall'analisi del p-value associato e in corrispondenza di valori inferiori a 0.05 si dedice di rifiutare H0 e viceversa.

In questo caso si ottiene un p-value pari a 0.2998 perciò non si rifiuta H0 e quindi si può affermare che i due modelli sono equivalenti. Per il principio di parsimonia il modello da preferire è il più semplice che in questo caso corrisponde a bank.fit2.

```
## Analysis of Deviance Table
##
## Model 1: churn ~ customer_id + credit_score + country + gender + age +
##     tenure + balance + products_number + credit_card + active_member +
##     estimated_salary
## Model 2: churn ~ country + gender + age + tenure + balance + active_member
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      6988     5955.0
## 2      6993     5961.1 -5  -6.0661   0.2998
```

Si può effettuare un ulteriore confronto tra i due modelli considerando i criteri AIC e BIC criteri:

- **AIC**: L'Akaike information criterion riguarda il grado di bilanciamento tra complessità e bontà di adattamento. La formula tiene conto della log-verosimiglianza e di un termine di penalizzazione sul numero di parametri (utile ad evitare sovradattamento). Nel confronto, valori di AIC più bassi indicano un modello preferibile.
- **BIC**: Il Bayesian Information Criterion è simile all'AIC ma contiene un termine di penalizzazione maggiore perchè utilizza il logaritmo per riuscire a penalizzare maggiormente i modelli con più parametri. Anche per questo criterio, un valore più basso di BIC indica un modello preferibile.

In questo caso, in accordo con il test LRT, sia AIC che BIC indicano che il modello bank.fit2 è preferibile

```
##           df      BIC
## bank.fit  13 6070.085
## bank.fit2  8 6031.882
```

```
##           df      AIC
## bank.fit  13 5980.985
## bank.fit2  8 5977.051
```

Probit e Clog-log models

Come detto inizialmente, la regressione logistica prevede la possibilità di utilizzare delle link function alternative. Dopo aver costruito il modello logit è possibile creare dei modelli con i link probit e clog-log, per poi confrontarli tra loro sulla base alle perfomance ottenute.

- **clog-log**: utilizza la funzione complementare al logaritmo negativo del complemento della variabile casuale. Risulta utile in caso di squilibri nei dati ovvero se una classe è molto più frequente dell'altra. Come per il modello logit, dapprima si considerano tutte le variabili per poi creare un secondo modello che includa solo quelle significative e tramite l'anova test si sceglie il migliore.

```
##
## Call:
## glm(formula = churn ~ ., family = binomial(link = "cloglog"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2999  -0.6511  -0.4657  -0.2937   2.8458
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -4.048e+00  5.863e+00  -0.690   0.4900
## customer_id       6.651e-08  3.735e-07   0.178   0.8587
## credit_score     -3.901e-04  2.781e-04  -1.403   0.1606
## countryGermany    6.521e-01  6.614e-02   9.858  < 2e-16 ***
## countrySpain      4.386e-02  7.309e-02   0.600   0.5484
## genderMale       -4.555e-01  5.440e-02  -8.374  < 2e-16 ***
## age               5.547e-02  2.379e-03  23.321  < 2e-16 ***
## tenure           -2.011e-02  9.246e-03  -2.175   0.0296 *
## balance           2.242e-06  5.241e-07   4.278 1.89e-05 ***
## products_number  -8.947e-02  4.609e-02  -1.941   0.0523 .
## credit_card1     -1.879e-02  5.924e-02  -0.317   0.7511
```

```
## active_member1   -9.835e-01  5.832e-02 -16.864  < 2e-16 ***
## estimated_salary  2.168e-08  4.693e-07   0.046   0.9631
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7077.6  on 7000  degrees of freedom
## Residual deviance: 5978.9  on 6988  degrees of freedom
## AIC: 6004.9
##
## Number of Fisher Scoring iterations: 8
```

Il p-value è pari a 0.2565 quindi i modelli sono equivalenti e si sceglie il modello più semplice, bank.fit.clog2

```
## Analysis of Deviance Table
##
## Model 1: churn ~ customer_id + credit_score + country + gender + age +
##     tenure + balance + products_number + credit_card + active_member +
##     estimated_salary
## Model 2: churn ~ country + gender + age + tenure + balance + active_member
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      6988     5978.9
## 2      6993     5984.3 -5  -5.3464   0.3751
```

- **probit**: utilizza la funzione di distribuzione cumulativa Normale standard che come il logit mappa la probabilità su una scala continua (tra meno infinito e più infinito) ma presenta delle code più leggere e non consente un'interpretazione diretta in termini di odds.

```
##
## Call:
## glm(formula = churn ~ ., family = binomial(link = "probit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3351  -0.6661  -0.4545  -0.2360   3.2480
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -4.535e+00  4.018e+00  -1.129   0.2590
## customer_id       1.618e-07  2.559e-07   0.632   0.5271
## credit_score     -3.216e-04  1.916e-04  -1.678   0.0933 .
## countryGermany    4.782e-01  4.658e-02  10.267  < 2e-16 ***
## countrySpain      2.864e-02  4.715e-02   0.608   0.5435
## genderMale       -3.219e-01  3.690e-02  -8.723  < 2e-16 ***
## age               4.160e-02  1.750e-03  23.769  < 2e-16 ***
## tenure           -1.221e-02  6.348e-03  -1.923   0.0544 .
## balance           1.518e-06  3.464e-07   4.382 1.18e-05 ***
## products_number  -5.632e-02  3.229e-02  -1.744   0.0811 .
## credit_card1     -1.180e-02  4.046e-02  -0.292   0.7705
## active_member1   -5.872e-01  3.804e-02 -15.436  < 2e-16 ***
## estimated_salary  6.685e-08  3.216e-07   0.208   0.8353
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7077.6  on 7000  degrees of freedom
## Residual deviance: 5958.6  on 6988  degrees of freedom
## AIC: 5984.6
##
## Number of Fisher Scoring iterations: 5
```

Il p-value è pari a 0.09653 quindi si conclude che i modelli sono equivalenti e sis scegli ancora una volta il più semplice (bank.fit.probit2) in base al principio di parsimonia

```
## Analysis of Deviance Table
##
## Model 1: churn ~ customer_id + credit_score + country + gender + age +
##     tenure + balance + products_number + credit_card + active_member +
##     estimated_salary
## Model 2: churn ~ country + gender + age + balance + active_member
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      6988     5958.6
## 2      6994     5969.3 -6  -10.747  0.09653 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Infine, è possibile effettuare un confronto tra i tre modelli migliori con le tre diverse funzioni di collegamento.Riprendendo i criteri AIC e BIC, i risultati a confronto indicano che il modello con le migliori performance è il modello che utlizza la link function logit e che include le sole variabili significative, ovvero bank.fit2. Infatti, AIC e BIC per quel modello risultano i più bassi.

```
##        logit   probit   loglog
## AIC 5977.051 5983.324 6000.293
## BIC 6031.882 6031.301 6055.123
```

Modello migliore

Una volta individuato il modello migliore si può procedere con ulteriori analisi su di esso. In particolare, verranno esaminati i coefficienti, gli effetti marginali e verranno realizzate delle previsioni su dati nuovi rispetto a quelli di train.

Coefficienti

Calcolo dei coefficienti e dei rispettivi intervalli di confidenza. I coefficienti positivi indicano un aumento delle probabilità di churn, mentre quelli negativi suggeriscono una diminuzione. Inoltre, la dimensione dei coefficienti indicano quanto forte sia tale effetto. In questo caso, i coefficienti positivi per i clienti tedeschi e spagnoli indicano una maggiore probabilità di churn rispetto ai francesi (l'effetto è maggiore per i tedeschi 8.157291e-01), il coefficiente negativo di genderMale indica che gli uomini hanno probabilità minori di effettuare churn rispetto alle donne, i clienti più anziani potrebbero avere una probabilità più elevata di churn rispetto a quelli più giovani, la maggiore durata del rapporto cliente-banca tende a diminuire la probabilità di churn, saldi più elevati indurrebbero ad una maggiore propensione all'abbandono (anche se l'effetto è lieve 2.971927e-06), infine i membri attivi hanno una probabilità inferiore di churn rispetto a quelli inattivi.

```
## In attesa che venga eseguita la profilazione...

##                       betaHat         2.5 %        97.5 %
## (Intercept)    -3.957454e+00 -4.265602e+00 -3.653834e+00
## countryGermany  8.157291e-01  6.578837e-01  9.742265e-01
## countrySpain    4.664044e-02 -1.205583e-01  2.122494e-01
## genderMale     -5.611525e-01 -6.894226e-01 -4.333456e-01
## age             7.279864e-02  6.677121e-02  7.890153e-02
## tenure         -2.203728e-02 -4.400194e-02 -1.019023e-04
## balance         2.971927e-06  1.806379e-06  4.138620e-06
## active_member1 -1.093588e+00 -1.229413e+00 -9.593028e-01
```

Se si considerano gli odd ratio si tiene conto della variazione percentuale nei rapporti di odds per un aumento unitario nelle variabili predittorie. In questo caso, considerando un aumento unitario di una variabile per volta, ceteris paribus le altre, possiamo concludere che: countryGermany fa aumentare l'odds di churn di un fattore di 2,26 mentre countrySpain di un fattore di 1.04,GenderMale lo fa aumentare di 0.57, age di 1.07, tenure del 0.97, balance di 1 e active_member di 0.33.

```
##                                2.5 %     97.5 %
## (Intercept)    0.01911172 0.01404341 0.02589167
## countryGermany 2.26082344 1.93070209 2.64911745
## countrySpain   1.04774522 0.88642543 1.23645616
## genderMale     0.57055113 0.50186577 0.64833641
## age            1.07551395 1.06905086 1.08209777
## tenure         0.97820377 0.95695210 0.99989810
## balance        1.00000297 1.00000181 1.00000414
## active_member1 0.33501243 0.29246432 0.38315995
```

Prediction

Sulla base del modello addestrato e considerando dei nuovi dati, è possibile eseguire delle previsioni rispetto alla decisione dei clienti di abbandonare l'istituto bancario.

Il primo caso considera un uomo e una donna con le stesse identiche caratteristiche e quindi vuole predire la decisione di churn in base alla differenza di genere. Il risulato indica che la probabilità di churn per una donna (21,7%) è maggiore rispetto a quella di un uomo (13,7%)

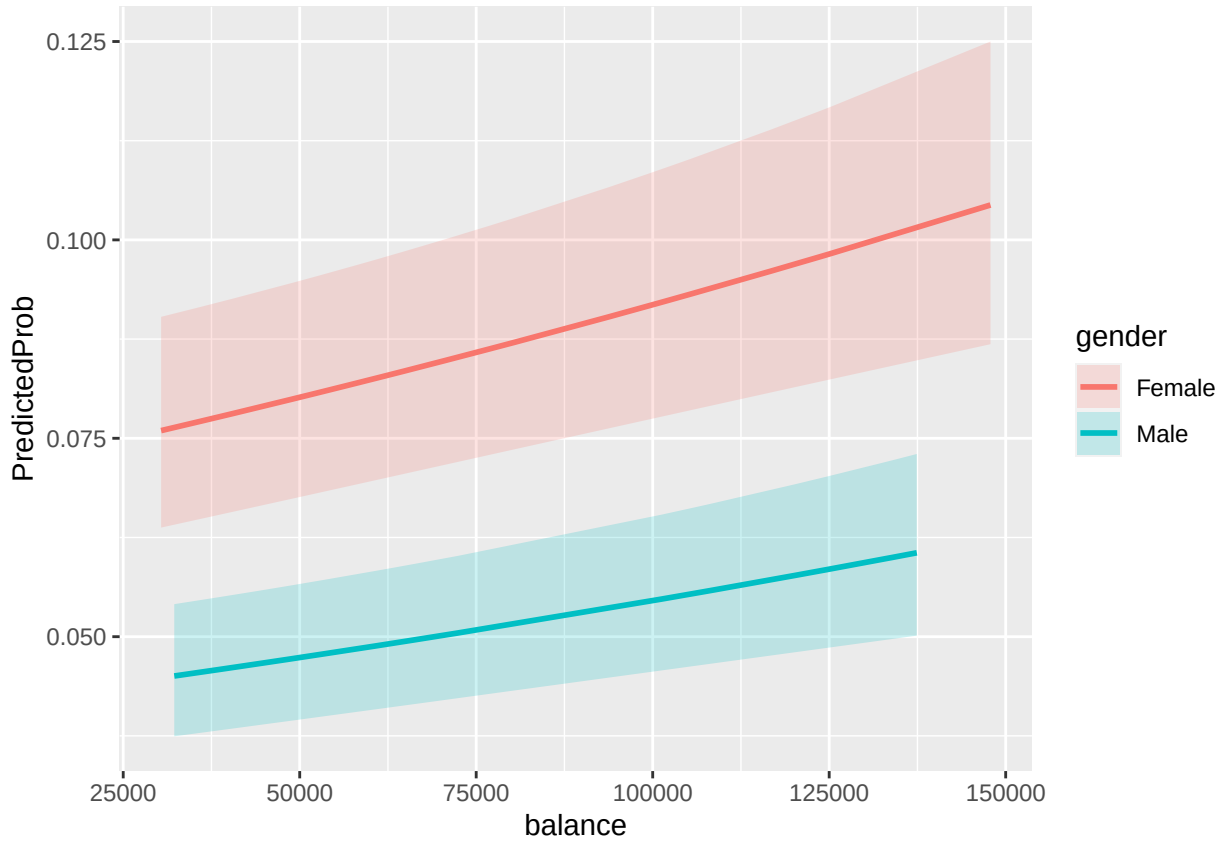
```
##         1         2
## 0.1370284 0.2177135
```

Il secondo caso analizza l'impatto dell'attività/inattività di un cliente rispetto alla scelta di churn. Si considerano livelli di acitve_member opposti per due uomini di mezz'età che vivono in Spagna, clienti da 5 anni e che hanno un saldo medio . I risultati indicano che la probabilità di churn per un cliente inattivo è pari al 19% mentre per un cliente attivo è pari al 7,3%. Si può concludere che i clienti attivi hanno una propensione minore all'abbandono rispetto agli inattivi.

```
##         1         2
## 0.1910973 0.0733398
```

Il terzo caso concerne la scelta di churn in base a differenti valori del saldo. Vengono considerati 50 nuovi clienti che vivono in Spagna, con età di 35 anni e da 6 anni clienti attivi. I risultati indicano che le previsioni in corrispondenza di valori di balance più elevati anche la percentuale di prediction aumenta. Ad esempio, in corrispondenza di un uomo che detiene un saldo pari a 49337.01 ho una prediction di churn del 4,7% mentre per un saldo di 120264.97 ottengo una prediction di churn del 5,7%

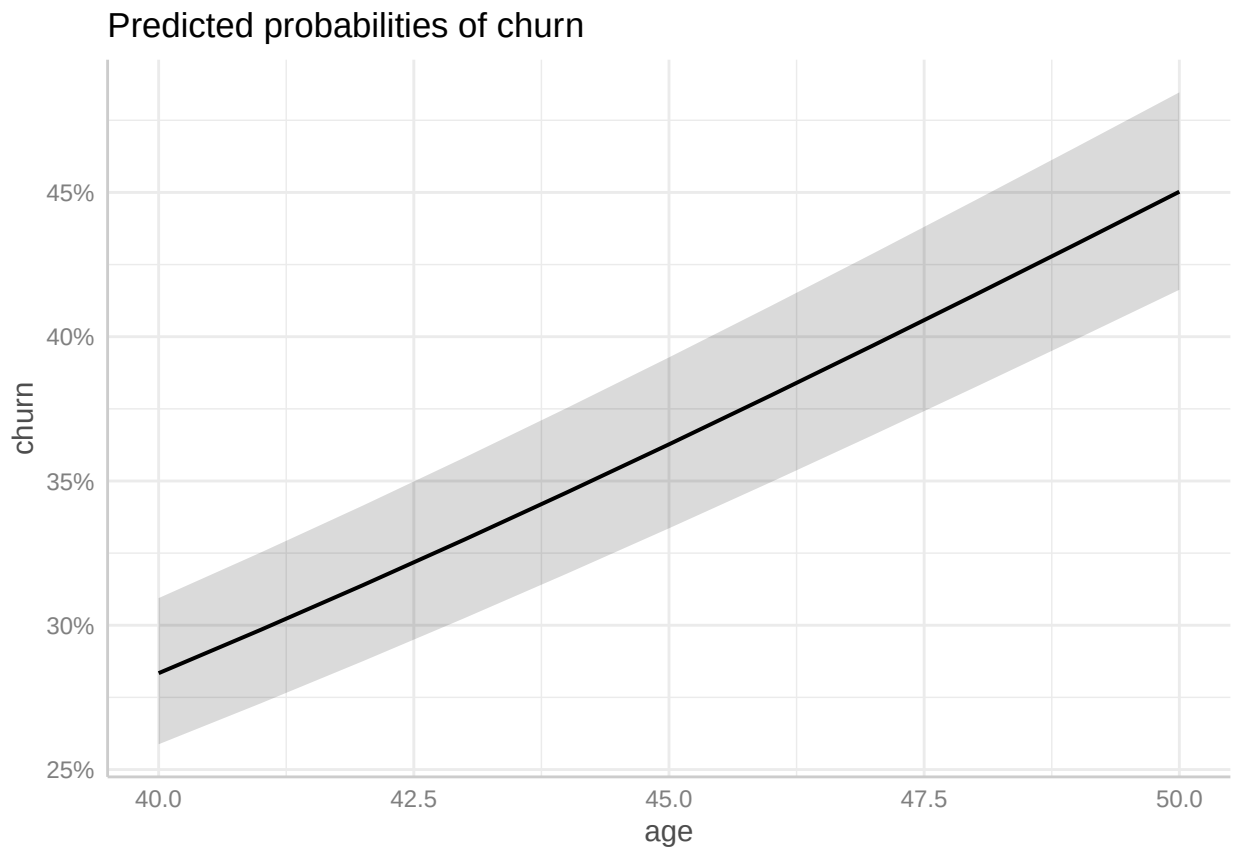
```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



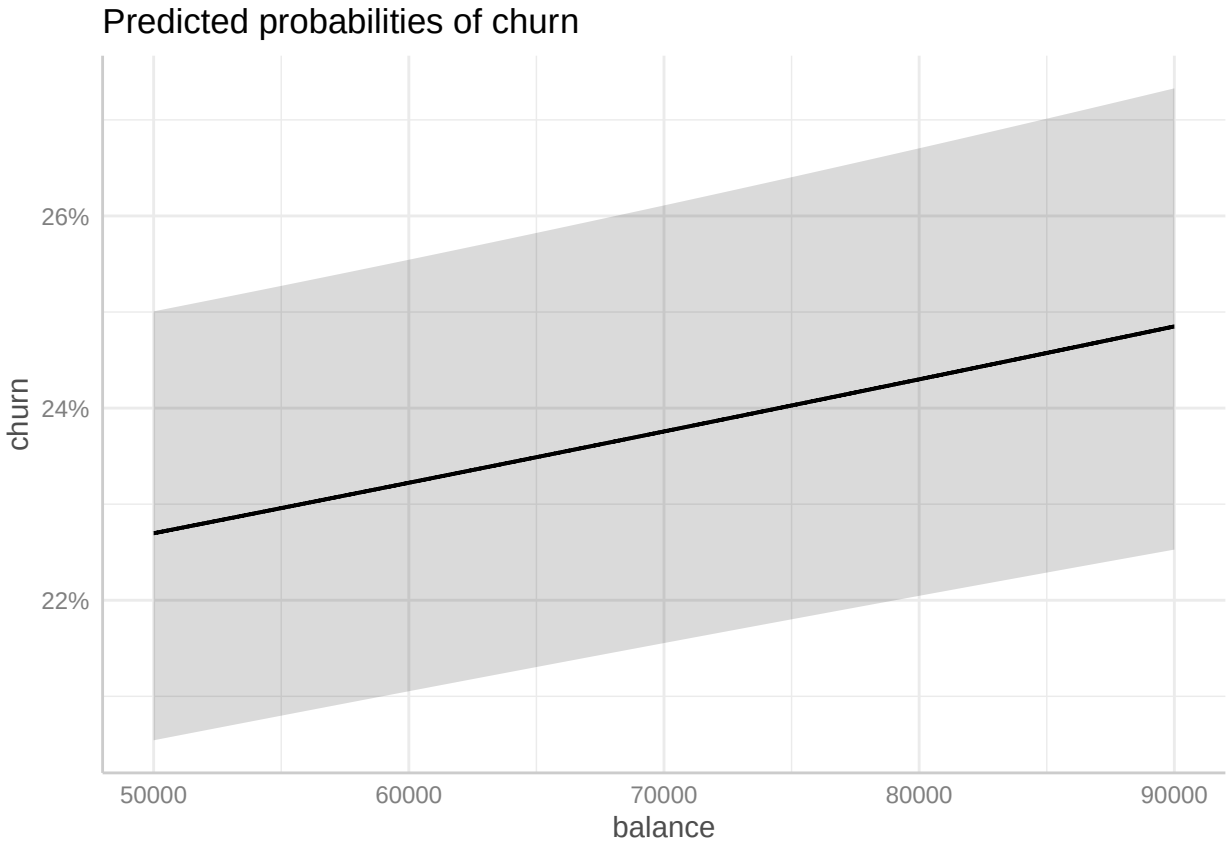
Marginal effects

Gli effetti marginali rappresentano il cambiamento stimato nella variabile risposta associato ad un aumento unitario nelle variabili indipendenti, ceteris paribus tutte le altre. In altre parole permettono di calcolare l'effetto di una singola variabile sul valore atteso della variabile dipendente.

In questo primo caso, è possibile verificare l'effetto delle variabili age e balance sulla variabile risposta chrurn. Considerando un aumento unitario di age tra i 40 e i 50 anni possiamo notare che all'aumentare dell'età aumenta anche la previsione che il cliente effettui un churn. Ad esempio, per età pari a 45 anni il valore di predicted è 0.36 (con un intervallo di confidenza tra 0.33 e 0.39) mentre per un aumento unitario a 46 anni il predicted aumenta a 0.38 (con un intervallo di confidenza tra 0.35 e 0.41). Graficamente, è possibile notare la relazione positiva tra età e probabilità prevista di churn e i relativi intervalli di confidenza.



Per analizzare gli effetti marginali rispetto alla variabile balance sono stati considerati degli aumenti diversi da age perchè l'aumento unitario di 1 per il saldo avrebbe una significatività limitata. Perciò, rispetto al saldo consideriamo degli aumenti dell'ordine di 5.000. Anche in questo caso, è possibile notare che l'aumento del saldo comporta un aumento della probabilità di churn prevista. Si passa da una probabilità del 23% per saldi compresi tra 50.000 e 60.000 ad una probabilità del 25% per saldi pari a 90.000.



Model validation

La validazione di un modello (in questo caso GLM) è fondamentale per garantire che il modello sia in grado di generalizzare bene su nuovi dati e che le sue previsioni siano affidabili nella pratica. Utilizzando tecniche di divisione del dataset e valutazioni appropriate, è possibile valutare l'efficacia del GLM e apportare eventuali miglioramenti necessari.

Accuracy

L'accuracy rappresenta la percentuale di previsioni corrette rispetto al totale delle previsioni effettuate dal modello. In questo caso, sul train-set è dell'81.46% mentre per il test-set corrisponde all'80%. Il valore della metrica di test indica che il modello ha una buona capacità di generalizzazione su nuovi dati. Inoltre, non essendo presente un'elevata discrepanza tra i due valori di accuracy è possibile dedurre che non sia presente overfitting.

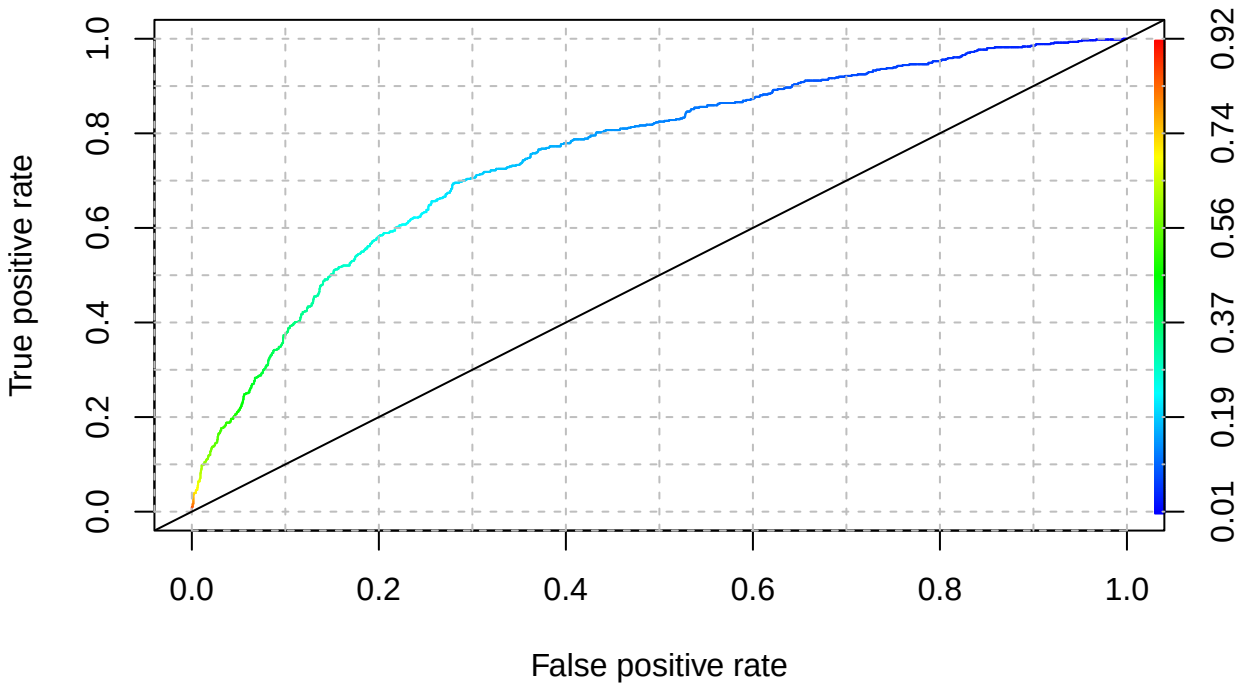
```
## [1] 0.8145979
```

```
## [1] 0.8026009
```

AUC-ROC

La curva AUC-ROC (Area Under the Receiver Operating Characteristic Curve) è una metrica di valutazione delle prestazioni di un modello di classificazione binaria. Essa visualizza la capacità del modello di discrim-

inare tra le classi positive e negative variando la soglia di decisione. In generale, un AUC di 0.7553 è positivo e indica che il modello ha una buona capacità di distinguere tra casi positivi e negativi.




```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
## Area under the curve: 0.7553
```

Precision, Recall, F1-score

Poichè l'accuracy non è la metrica migliore in caso di classi sbilanciate, è doveroso considerare ulteriori metriche. La precision misura la precisione del modello nelle previsioni positive. Un valore più alto di precision indica che il modello fa pochi falsi positivi. La recall misura la capacità del modello di identificare tutte le istanze positive. Un valore più alto di recall indica che il modello fa pochi falsi negativi.F1-score è una media armonica tra precision e recall. È utile quando si desidera una misura bilanciata tra queste due metriche.Nel set di addestramento, il modello mostra una precision discreta (62.12%), ma una bassa recall (23.00%), indicando che il modello ha difficoltà nell'identificare correttamente le istanze positive. L'F1-score di 0.34 suggerisce un compromesso tra precision e recall. Sul set di test, la precision è leggermente migliorata (54.50%), ma la recall rimane bassa (18.82%). L'F1-score di 0.28 conferma il compromesso tra precision e recall anche in questo caso.

```
## Precision sul set di addestramento: 0.6212121
```

```
## Recall sul set di addestramento: 0.230014
```

```
## F1-score sul set di addestramento: 0.3357216

## Precision sul set di test: 0.5450237

## Recall sul set di test: 0.188216

## F1-score sul set di test: 0.2798054
```

Confusion matrix

La matrice di confusione mostra il numero di predizioni corrette e errate fatte dal modello per entrambe le classi.In generale, queste statistiche indicano che il modello ha una precisione e una specificità relativamente alte, ma una sensibilità bassa. Questo suggerisce che il modello ha una buona capacità di identificare le istanze negative, ma può avere difficoltà nell'identificare correttamente le istanze positive

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2292   96
##          1  496  115
##
##                Accuracy : 0.8026
##                  95% CI : (0.7879, 0.8167)
##     No Information Rate : 0.9296
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1957
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.54502
##             Specificity : 0.82209
##          Pos Pred Value : 0.18822
##          Neg Pred Value : 0.95980
##              Prevalence : 0.07036
##          Detection Rate : 0.03835
##    Detection Prevalence : 0.20373
##       Balanced Accuracy : 0.68356
##
##        'Positive' Class : 1
##


##
##        0    1
##   0 96.0  4.0
##   1 81.2 18.8
```

Conclusion

L'obiettivo principale di questa analisi è sfruttare le conoscenze acquisite riguardo ai Modelli Lineari Generalizzati (GLM) per sviluppare un modello affidabile in grado di offrire prestazioni elevate nella previsione

dei comportamenti dei clienti bancari, specialmente in relazione alla decisione di churn. La scelta del dataset è stata motivata dalla rilevanza pratica e dall'interesse concreto inerente a questo contesto. Inoltre, la selezione della regressione logistica come modello è stata dettata dalla natura binaria della variabile di risposta, il churn. Il processo analitico è iniziato con un'approfondita analisi descrittiva e grafica dei dati, finalizzata a ottenere informazioni utili per i modelli. Successivamente, durante la fase di modellazione, sono stati esplorati diverse alternative, considerando la significatività delle variabili e valutando diverse opzioni di funzioni di collegamento (link function). Il modello ottimale, identificato come il più performante, è stato selezionato per analisi più approfondite (effetti marginali e previsioni). Le conclusioni derivanti dall'analisi del modello sono state approfondite per comprendere come si adatta ai dati e per estrarre informazioni rilevanti per comprendere i comportamenti dei clienti. In particolare, il modello indica diversi aspetti interessanti, tra cui: il modo in cui la nazione di riferimento possa influenzare i clienti in base alle differenti normative regole vigenti, infatti per i diversi paesi considerati sono stati ottenuti dei risultati diversi. Inoltre, sembrerebbe che da un punto di vista prettamente statistico le donne siano più inclini ad effettuare dei churn. Un ulteriore risultato significativo è legato al saldo dei clienti perchè i clienti con maggiori capitali investiti nell'istituto sembrano essere i pù propensi ad abbandonare forse perchè attratti maggiormente da opportunità di risparmio altrove. Infine, sulla base dei risultati ottenuti nella fase di validazione, è importante sottolineare che a causa della natura sensibile dei dati bancari l'accesso a un numero maggiore di informazioni potrebbe migliorare la generalizzabilità delle previsioni, consentendo una comprensione più approfondita e una maggiore precisione nel modellare il churn dei clienti.