CUSTOMER CHURN ANALYSIS

PREDICTION AND PREVENTION

17 Gennaio 2020

Telecom Italia (TIM)

Autore: Pietro Ruffo

INDICE

Introduzione

Customer Churn Analysis

Prediction and Prevention

Il nuovo anno si prospetta ricco di sfide per il settore delle telecomunicazioni. La diffusione capillare di internet su tutto il Paese ha rivoluzionato non solo usi e costumi dei consumatori ma anche le logiche del mercato di riferimento in termini di servizi richiesti e interazioni cliente-impresa.

La Custumer Churn Analysis è uno strumento tipico dell'e-business e in generale di tutte le imprese che operano nell'economia digitale. Questo genere di studio, infatti, consente di valutare la propensione al rischio "churn" del cliente ovvero la tendenza dello user alla cancellazione dell'abbonamento e alla sua conseguente fuoriscita dalla baseclienti.

Si dimostrerà come ormai anche un'azienda leader nel settore telecom non può prescindere da questo tipo di approccio.
E' necessario un cambio di paradigma:

Acquisire un nuovo cliente costa molto più che trattenerne uno.

Nella digital economy un cliente soddisfatto è un asset che, se gestito adeguatemente, garantisce all'impresa un flusso di guadagni costante e continuo nel tempo.
Capire quali sono le nuove tendenze e i prodotti che più influiscono sul grado di soddisfazione degli user risulta essere di primaria importanza per adottare strategie di marketing efficaci che garantiscano una customer retention duratura.
Le moderne tecniche di analisi dati consentono di individuare le criticità attualmente presenti nella baseclienti e offrono al management la possibilità di individuare le soluzioni ottime rivolte a creare una customer experience telecom di qualità con un conseguente incremento delle performance aziendali.

Struttura Dataset

Per l'analisi dei dati ci si è avvalsi del software statistico *R-studio*.

Il dataset impiegato nello studio contiente 21 attributi relativi ad un campione di 7043 clienti Telecom.

```
> glimpse(telecomdata) #capiamo la struttura del dataset e la natura delle variabili al suo interno
Observations: 7,043
Variables: 21
$ customerID       <chr> "7590-VHVEG", "5575-GNVDE", "3668-QPYBK", "7795-CFOCW", "9237-HQITU", "9305-CDSKC", "1452-KIOVK", "6713-OKOMC", "7892-P...
$ gender           <chr> "Female", "Male", "Male", "Male", "Female", "Female", "Male", "Female", "Female", "Male", "Male", "Male", "Male", "Male...
$ SeniorCitizen    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0,...
$ Partner          <chr> "Yes", "No", "No", "No", "No", "No", "No", "No", "Yes", "No", "Yes", "No", "Yes", "No", "No", "Yes", "No", "No", "Yes",...
$ Dependents       <chr> "No", "No", "No", "No", "No", "No", "Yes", "No", "No", "Yes", "Yes", "No", "No", "No", "No", "Yes", "No", "Yes", "Yes",...
$ tenure           <dbl> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 58, 49, 25, 69, 52, 71, 10, 21, 1, 12, 1, 58, 49, 30, 47, 1, 72, 17, 71, 2,...
$ PhoneService     <chr> "No", "Yes", "Yes", "No", "Yes", "Yes", "Yes", "No", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Ye...
$ MultipleLines    <chr> "No phone service", "No", "No", "No phone service", "No", "Yes", "Yes", "No phone service", "Yes", "No", "No", "No", "Y...
$ InternetService  <chr> "DSL", "DSL", "DSL", "DSL", "Fiber optic", "Fiber optic", "Fiber optic", "DSL", "Fiber optic", "DSL", "DSL", "No", "Fib...
$ OnlineSecurity   <chr> "No", "Yes", "Yes", "Yes", "No", "No", "No", "Yes", "No", "Yes", "Yes", "No internet service", "No", "No", "Yes", "Yes"...
$ OnlineBackup     <chr> "Yes", "No", "Yes", "No", "No", "No", "Yes", "No", "No", "Yes", "No", "No internet service", "No", "Yes", "No", "Yes", ...
$ DeviceProtection <chr> "No", "Yes", "No", "Yes", "No", "Yes", "No", "No", "Yes", "No", "No", "No internet service", "Yes", "Yes", "Yes", "Yes"...
$ TechSupport      <chr> "No", "No", "No", "Yes", "No", "No", "No", "No", "Yes", "No", "No", "No internet service", "No", "No", "Yes", "Yes", "N...
$ StreamingTV      <chr> "No", "No", "No", "No", "No", "Yes", "Yes", "No", "Yes", "No", "No", "No internet service", "Yes", "Yes", "Yes", "Yes",...
$ StreamingMovies  <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "Yes", "No", "No", "No internet service", "Yes", "Yes", "Yes", "Yes", ...
$ Contract         <chr> "Month-to-month", "One year", "Month-to-month", "One year", "Month-to-month", "Month-to-month", "Month-to-month", "Mont...
$ PaperlessBilling <chr> "Yes", "No", "Yes", "No", "Yes", "Yes", "Yes", "No", "Yes", "No", "Yes", "No", "No", "Yes", "Yes", "No", "No", "No", "N...
$ PaymentMethod    <chr> "Electronic check", "Mailed check", "Mailed check", "Bank transfer (automatic)", "Electronic check", "Electronic check"...
$ MonthlyCharges   <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89.10, 29.75, 104.80, 56.15, 49.95, 18.95, 100.35, 103.70, 105.50, 113.25, 20...
$ TotalCharges     <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820.50, 1949.40, 301.90, 3046.05, 3487.95, 587.45, 326.80, 5681.10, 5036.30, 2...
$ Churn            <chr> "No", "No", "Yes", "No", "Yes", "Yes", "No", "No", "Yes", "No", "No", "No", "No", "Yes", "No", "No", "No", "No", "Yes",...
```

Le variabili possono essere classificate in 4 macro gruppi: informazioni anagrafiche, servizi basic, servizi premium e condizioni contrattuali.

Informazioni anagrafiche (tra parentesi la relativa traduzione italiana impiegata nel modello):

- *customerID(idcliente),* codice identificativo del cliente;
- *gender(sesso),* sesso del cliente *(male, female)*;
- *SeniorCitizen(pensionato),* se il cliente è pensionato o meno *(Yes, No);*
- *Partner(partner),* se il cliente ha un partner o meno *(Yes, No);*
- *Dependents(persone a carico)* se il cliente ha persone a carico o meno *(Yes, No).*

Servizi basic:

- *PhoneService(linea telefonica)* se il cliente ha attivato la linea telefonica o meno *(Yes, No);*
- *InternetService(internet)* Il tipo connessione internet utilizzata dal cliente (*DSL, Fiber optic, No);*
- *TechSupport(assistenza tecnica)* se il cliente ha usufruito dell'assistenza tecnica *(Yes, No, No internet service).*

Servizi premium:

- *MultipleLines*(più linee telefoniche) se il cliente ha attivato il servizio di linee telefoniche multiple *(Yes, No, No phone service);*
- *OnlineSecurity(sicurezza online)* se il cliente ha attivato il servizio di sicurezza online *(Yes, No, No internet service);*

- *OnlineBackup*(backup) se il cliente ha attivato il servizio di backup online *(Yes, No, No internet service);*
- *DeviceProtection*(protezione dispositivo) se il cliente ha attivato il servizio di protezione dispositivo *(Yes, No, No internet service);*
- *StreamingTV*,(streamingtv) se il cliente ha attivato il servizio di tv in streaming *(Yes, No, No internet service);*
- *StreamingMovies(streamingfilm),* se il cliente gode del servizio di film in streaming *(Yes, No, No internet service).*

Condizioni contrattuali:

- *Tenure(permanenza)* da quanti mesi il soggetto è cliente;
- *Contract(contratto)* tipo di contratto *(Month-to-month, One year, Two year)*
- *PaperlessBilling(pagamento elettronico)*se il cliente paga in forma elettronica (*Yes, No)*
- *PaymentMethod(metodo di pagamento)modalità di pagamento usata dal cliente (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic));*
- *MonthlyCharges*(spesa mensile) del cliente,
- *TotalCharges(*spesa totale) del cliente*;*
- *Churn(disdetta)*se il cliente ha disdetto o meno.

Si tratta nel complesso di un dataset costituito da variabili categoriali ad eccezione di spesa mensile, spesa totale, disdetta e tenure che sono invece delle variabili quantitative.

Il set di dati è decisamente ricco e offre diversi spunti di analisi, motivo per cui, **rispetto l'obiettivo di questo report verranno presi in considerazione solo alcuni caratteri specifici.** Questo consentirà la costruzione di un modello coerente per la formulazione di *insights* utili agli stakeholders.

Dataset pre-processing

In primo luogo, occorre definire l'obiettivo ovvero il carattere rispetto al quale si procederà alla modellazione dei dati. Il carattere Churn, sarà la variabile risposta cioè quella grandezza rispetto alla quale si valuterà la propensione del cliente a disattivare piuttosto che mantenere l'abbonamento ai servizi telecom.
Per quanto riguarda i possibili fattori determinanti tale propensione ci si focalizzerà sui seguenti caratteri: sesso, partner, persone a carico, linea telefonica, assistenza tecnica, streaming tv, streaming film, pagamento elettronico, permanenza e spesa mensile.
Si avrà così uno spaccato completo del dataset poiché le variabili sopra elencate appartengono a ciascuno dei macrogruppi presenti: Informazioni anagrafiche, servizi basic, servizi premium e condizioni contrattuali.
Una volta individuate le informazioni utili alla costruzione del modello si procederà alla fase di cleaning del set di dati al fine di garantire un'analisi dati chiara ed informativa.

Rimozione dei dati in eccesso:

```
#escludiamo preventivamente la variabili "customerID","Multiplelines","Onlinesecurity","Online backup","Device protection" "Seniocitizen",
#"Contract","Paymentmethod" e "TotalCharges" poiché irrilevanti ai fini della nostra analisi.
telecomdata$customerID<-NULL
telecomdata$MultipleLines<-NULL
telecomdata$OnlineSecurity<-NULL
telecomdata$OnlineBackup<-NULL
telecomdata$DeviceProtection<-NULL
telecomdata$SeniorCitizen<-NULL
telecomdata$Contract<-NULL
telecomdata$PaymentMethod<-NULL
telecomdata$TotalCharges<-NULL
```

Si attesta l'assenza di valori mancanti nel dati residui che potrebbero compromettere l'analisi:

```
> #verifichiamo la presenza di valori mancanti nel dataset
> sum(is.na(telecomdata))
[1] 0
```

Si procede alla conversione in "factors" delle variabili categoriali così da renderle "leggibili" nella stima del modello. Seguirà l'accorpamento delle risposte "No" e "No internet service" per le variabili Internet Service, Tech Support, Streaming Movies e Streaming TV così da renderle binarie per un ulteriore semplificazione:

```
#procediamo nella conversione delle variabili tramite la funzione as.factor
telecomdata$gender <-as.factor(telecomdata$gender)
telecomdata$Churn <- as.factor(telecomdata$Churn)
telecomdata$Partner <- as.factor(telecomdata$Partner)
telecomdata$Dependents<- as.factor(telecomdata$Dependents)
telecomdata$PhoneService<- as.factor(telecomdata$PhoneService)
telecomdata$PaperlessBilling <- as.factor(telecomdata$PaperlessBilling)
#accorpiamo le risposte "No" e "No phone service" per le seguenti variabili
telecomdata$InternetService <- as.factor(telecomdata$InternetService)
telecomdata$TechSupport <- factor( with(telecomdata, replace(TechSupport, TechSupport %in% c( "No", "No internet service"),"No") ) )
telecomdata$StreamingMovies<- factor( with(telecomdata, replace( StreamingMovies, StreamingMovies %in% c( "No", "No internet service"),"No") ) )
telecomdata$StreamingTV <- factor( with(telecomdata, replace( StreamingTV, StreamingTV %in% c( "No", "No internet service"),"No") ) )
```

Rinomina delle variabili, delle risposte e definizione delle categorie di riferimento:

```
#rinominiamo le variabili quantitative
permanenza <- telecomdata$tenure
spesamensile<-telecomdata$MonthlyCharges
#rinominiamo le variabili categoriali, le risposte e riordiniamo le categorie (la prima sarà quella di riferimento)
disdetta <-factor(telecomdata$Churn, levels=c("Yes","No"), labels=c("si", "no"))
sesso <- factor(telecomdata$gender, levels=c("Female","Male"), labels=c("donne", "uomini"))
partner <- factor(telecomdata$Partner, levels=c("Yes","No"), labels=c("con", "single"))
personeacarico<-factor(telecomdata$Dependents, levels=c("Yes","No"), labels=c("con", "senza"))
lineatelefonica<-factor(telecomdata$PhoneService, levels=c("Yes","No"), labels=c("con l.t.", "senza l.t."))
internet<-factor(telecomdata$InternetService, levels=c("Fiber optic","DSL", "No"), labels=c("fibra", "adsl", "no"))
assistenzatecnica<-factor(telecomdata$TechSupport, levels=c("Yes","No"), labels=c("si", "no"))
streamingTV<-factor(telecomdata$StreamingTV, levels=c("Yes","No"), labels=c("si", "no"))
streamingfilm<-factor(telecomdata$StreamingMovies, levels=c("Yes","No"), labels=c("si", "no"))
pagamentoelettronico <- factor(telecomdata$PaperlessBilling, levels=c("Yes","No"), labels=c("si", "no"))
```

Il dataset risultante avrà questo aspetto:

```
> str(new.telecomdata)
'data.frame':   7043 obs. of  12 variables:
 $ sesso               : Factor w/ 2 levels "donne","uomini": 1 2 2 2 1 1 2 1 1 2 ...
 $ partner             : Factor w/ 2 levels "con","single": 1 2 2 2 2 2 2 2 1 2 ...
 $ personeacarico      : Factor w/ 2 levels "con","senza": 2 2 2 2 2 2 1 2 2 1 ...
 $ permanenza          : num  1 34 2 45 2 8 22 10 28 62 ...
 $ lineatelefonica     : Factor w/ 2 levels "con l.t.","senza l.t.": 2 1 1 2 1 1 1 2 1 1 ...
 $ internet            : Factor w/ 3 levels "fibra","adsl",..: 2 2 2 2 1 1 1 2 1 2 ...
 $ assistenzatecnica   : Factor w/ 2 levels "si","no": 2 2 2 1 2 2 2 2 1 2 ...
 $ streamingTV         : Factor w/ 2 levels "si","no": 2 2 2 2 2 1 1 2 1 2 ...
 $ streamingfilm       : Factor w/ 2 levels "si","no": 2 2 2 2 2 1 2 2 1 2 ...
 $ pagamentoelettronico: Factor w/ 2 levels "si","no": 1 2 1 2 1 1 1 2 1 2 ...
 $ spesamensile        : num  29.9 57 53.9 42.3 70.7 ...
 $ disdetta            : Factor w/ 2 levels "si","no": 2 2 1 2 1 1 2 2 1 2 ...
```

Terminata la fase di pre-processing il set di dati è pronto per essere analizzato.

Analisi esplorativa

Prima di iniziare la costruzione del modello è preferibile effettuare un'analisi esplorativa dei dati allo scopo di ricavare preventivamente informazioni significative rispetto l'obiettivo dello studio.
Per apprezzare la distribuzione delle frequenze di ciascun carattere rispetto alla variabile risposta (disdetta) si costruiscono tabelle a doppia entrata.

```
> xtabs(~ personeacarico + disdetta, data=new.telecomdata)
              disdetta
personeacarico   si   no
         con    326 1784
         senza 1543 3390
> xtabs(~ internet + disdetta, data=new.telecomdata)
        disdetta
internet   si   no
   fibra 1297 1799
   adsl   459 1962
   no     113 1413
> xtabs(~ assistenzatecnica+disdetta, data=new.telecomdata)
                 disdetta
assistenzatecnica   si   no
               si  310 1734
               no 1559 3440
> xtabs(~ streamingfilm+disdetta, data=new.telecomdata)
            disdetta
streamingfilm   si   no
           si  818 1914
           no 1051 3260
> xtabs(~ streamingTV+disdetta, data=new.telecomdata) #streamingfilm e streamingtv hanno frequenze molto simili rispetto la y
          disdetta
streamingTV   si   no
         si  814 1893
         no 1055 3281
> xtabs(~ pagamentoelettronico+disdetta, data=new.telecomdata)
                   disdetta
pagamentoelettronico   si   no
                  si 1400 2771
                  no  469 2403

> xtabs(~ partner+disdetta, data=new.telecomdata)
        disdetta
partner    si   no
  con     669 2733
  single 1200 2441
> xtabs(~ lineatelefonica+disdetta, data=telecomdata)
               disdetta
lineatelefonica   si   no
     con l.t.   1699 4662
     senza l.t.  170  512
> tabsex<-xtabs(~ sesso + disdetta, data=new.telecomdata)
> tabsex#le frequenze sono distribuite in modo pressocché identico tra i due sessi
        disdetta
sesso      si   no
  donne   939 2549
  uomini  930 2625
```

Questa prima analisi descrittiva offre due importanti spunti relativi ai caratteri sesso, streaming tv e streaming film .

In primo luogo, si dimostra in forma preventiva come **la disdetta del cliente sia del tutto indipendente rispetto al sesso.**

```
> oddsratio(tabsex, log=F) #or prossimo ad 1, sinonimo di indipendenzatra x e y
 odds ratios for sesso and disdetta

[1] 1.039782
```

Successivamente**, si attesta una forte dipendenza tra le variabili streamingfilm e streamingtv** (anche in virtù della somiglianza tra i due servizi)

```
> #dipendenza tra streamingtv e streamignfilm
> tabstreaming<-xtabs(~ streamingTV+streamingfilm, data=new.telecomdata)
> chisq.test(tabstreaming, correct = F)  #pvalue molto basso

        Pearson's Chi-squared test

data:  tabstreaming
X-squared = 2001.5, df = 1, p-value < 2.2e-16
```

Queste informazioni saranno molto utili nella fase di costruzione del modello.

Stima modello

Applicando il metodo della **regressione logistica** si andrà alla ricerca del modello statistico ottimo capace di stimare con precisione la propensione dello user ad uscire dalla baseclienti e soprattutto definire quelli che sono i fattori che più incidono su questa tendenza.
Si è partiti da un modello comprensivo di tutte le variabili precedentemente selezionate.

```
> #stimiamo il modello completo di tutte le variabili
> logistic1<- glm(disdetta~ sesso+partner+streamingTV+personeacarico+permanenza+lineatelefonica+internet+assistenzatecnica+streamingfilm+pagamentoelettr
onico+spesamensile, family=binomial(link=logit))
> summary(logistic1)#le variabili sesso, partner e linea telefonica risultano essere poco significative

Call:
glm(formula = disdetta ~ sesso + partner + streamingTV + personeacarico +
    permanenza + lineatelefonica + internet + assistenzatecnica +
    streamingfilm + pagamentoelettronico + spesamensile, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9880  -0.7056   0.3680   0.6584   1.9143

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -2.277940   0.765234  -2.977 0.002913 **
sessouomini               0.005382   0.064021   0.084 0.933003
partnersingle             0.044552   0.075883   0.587 0.557130
streamingTVno             0.439613   0.111670   3.937 8.26e-05 ***
personeacaricosenza      -0.334639   0.086118  -3.886 0.000102 ***
permanenza                0.044346   0.002043  21.706  < 2e-16 ***
lineatelefonicasenza l.t. -0.141171   0.201513  -0.701 0.483581
internetadsl              1.612685   0.207076   7.788 6.81e-15 ***
internetno                3.050722   0.425162   7.175 7.21e-13 ***
assistenzatecnicano      -0.505615   0.092553  -5.463 4.68e-08 ***
streamingfilmno           0.448903   0.111367   4.031 5.56e-05 ***
pagamentoelettronicono    0.460057   0.072642   6.333 2.40e-10 ***
spesamensile              0.015307   0.007444   2.056 0.039759 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8150.1  on 7042  degrees of freedom
Residual deviance: 6038.8  on 7030  degrees of freedom
AIC: 6064.8

Number of Fisher Scoring iterations: 5
```

Questo primo modello mostra come i β dei regressori sesso, partner e lineatelefonica siano significativamente diversi da zero, sembrerebbero non “spiegare” bene il rischio churn del cliente.

Si proveranno dunque delle interazioni per attestare un evenutale effetto simultaneo delle variabili sulla decisione finale del customer. Il modello logistic3 include contemporaneamente due interazioni quella tra i regressori partner e linea telefonica e quella tra sesso e permanenza. In altre parole, con la prima interazione si vuole verificare se un cliente con un compagno/a e provvisto di linea telefonica sia maggiormente propenso alla disattivazione, mentre con la seconda interazione si verifica se il numero di mesi di utilizzo dei servizi telecom e il sesso del cliente possano influire congiuntamente sulla scelta di rescidere il contratto.

La prima interazione si rivela priva di significatività motivo per cui i regressori partner e linea telefonica verranno eliminati dal modello. La seconda interazione sembrerebbe avere un impatto sulla risposta, tuttavia, in virtù della natura della problematica e della verifica effettuata nella fase di analisi

esplorativa il regressore sesso verrà rimosso. Possiamo dunque affermare con un buon grado di certezza **che il sesso del cliente non influenza la scelta nel rescindere il contratto con la società**. Il dato più significativo però è un altro: **la linea telefonica non incide sulla customer retention**. In altre parole**, il core business della società non contribuisce significativamente alla soddisfazione dello user**. Questo è un primo segnale di come il mercato delle telecomunicazioni sia cambiato e conseguentemente anche le pretese dei consumatori.

Occore a questo punto dell'analisi interrogarsi su quale possa essere adesso il key product offerto dalla compagnia. A tal proposito, l'analisi dei dati offre uno spunto importante, infatti, la costruzione dei modelli presenta una caratteristica costante: l'impatto del servzio internet.

I servizi di connessione rete si candidano come sostituiti dei servizi di telecomunicazione nella definizione del core business aziendale dal momento che la nuova generazione di clienti si dimostra sempre più sensibile ed interessata ai servizi e alle piattaforme disponibili in rete.

Allo scopo di comprovare questo sentiment, si è messo a confronto i modelli logisti5 e logistic7. Il primo comprendentente le varibili internet e streaming film e il secondo privo dei servizi di connessione e dei servizi streaming. Il modello logistic5 ha una maggiore capacità di adattamento ai dati e quindi in grado di stimare con più precisione la probabilità di rescissione del contratto da parte dello user.

```
> anova(logistic7,logistic5, test= "Chisq") #il test anova dimostra che il modello contenente internet e streamingfilm si adatta meglio
Analysis of Deviance Table

Model 1: disdetta ~ personeacarico + assistenzatecnica + pagamentoelettronico
Model 2: disdetta ~ personeacarico + internet + assistenzatecnica + streamingfilm +
    pagamentoelettronico
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      7039     7505.9
2      7036     6919.0  3    586.9 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

E' importante evidenziare come il modello logistic5 contenga solo una delle due variabili relative ai servizi in streaming ,ovvero streaming film, a causa della dipendenza tra i due regressori evidenziata nelle analisi esplorative dei dati. Un'informazione ridondante potrebbe influenzare negativamente la validità dello studio.
Successivamente, la costruzione del modello è stata incentrata sul ruolo giocato dalle variabili quantitative "permanenza" e "spesa mensile". Quest'ultima variabile a seguito di un'ulteriore prova con interazione non sembrerebbe influenzare la scelta finale del cliente, sintomo dell'essenzialità dei servizi di connessione internet, **il cliente non è attento al risparmio quando si tratta di avere una connessione rapida e di qualità.**

Il modello finale risulta essere il seguente:

```
> logistic11<-glm(disdetta~ personeacarico+internet+assistenzatecnica+streamingfilm+pagamentoelettronico+permanenza, family=binomial(link=logit))
> summary(logistic11)

Call:
glm(formula = disdetta ~ personeacarico + internet + assistenzatecnica +
    streamingfilm + pagamentoelettronico + permanenza, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9453  -0.7250   0.3731   0.6606   1.8315

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           -0.606887   0.124849  -4.861 1.17e-06 ***
personeacaricosenza   -0.318710   0.078212  -4.075 4.60e-05 ***
internetadsl           1.104528   0.074549  14.816  < 2e-16 ***
internetno             2.281753   0.119466  19.100  < 2e-16 ***
assistenzatecnicano   -0.589507   0.081109  -7.268 3.65e-13 ***
streamingfilmno        0.389962   0.073236   5.325 1.01e-07 ***
pagamentoelettronicono 0.479159   0.072224   6.634 3.26e-11 ***
permanenza             0.044713   0.001718  26.024  < 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: 8150.1  on 7042  degrees of freedom
Residual deviance: 6071.7  on 7035  degrees of freedom
AIC: 6087.7

Number of Fisher Scoring iterations: 5
```

Validazione modello

Una volta stimato il modello occorre validarlo per poter trarre delle informazioni rilevanti dallo stesso.

Il modello logistic11 si adatta a circa il 79% dei dati di partenza ciò significa che potrà essere utilizzato in futuro per stimare la propensione al rischio churn del cliente in un gran numero di casi.

```
> #tabella corretta classificazione
> tabcorrclass <- table(disdetta,logistic11$fitted >0.5)
> tabcorrclass

disdetta FALSE TRUE
      si   898  971
      no   482 4692
> sum(diag(tabcorrclass)/sum(tabcorrclass))
[1] 0.7936959
```

Per comprovare la validità del modello si è proceduti ad un ulteriore test sulla bontà di adattamento (hosmer-lemeshow test):

```
> hl<-hoslem.test(logistic11$y, fitted(logistic11), g=10 )
> hl

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  logistic11$y, fitted(logistic11)
X-squared = 3.4278, df = 8, p-value = 0.9047

> cbind(hl$expected,hl$observed)
                   yhat0    yhat1  y0  y1
[0.187,0.337] 517.801153 188.1988 517 189
(0.337,0.506] 441.844966 323.1550 437 328
(0.506,0.634] 280.530586 382.4694 274 389
(0.634,0.743] 211.370517 471.6295 215 468
(0.743,0.824] 150.878237 554.1218 157 548
(0.824,0.873] 106.367296 599.6327 112 594
(0.873,0.91]   74.977233 628.0228  78 625
(0.91,0.948]   49.235655 653.7643  51 652
(0.948,0.976]  26.051747 685.9483  19 693
(0.976,0.994]   9.942611 687.0574   9 688
```

La differenza tra le frequenze osservate nel set di dati e quelle stimate dal modello logistic11 è molto piccola sintomo di una elevata bontà di adattamento del modello che si presta adesso alla fase di interpretazione.

Utilizzo del modello: Churn Prediction

Per utilizzare il modello a scopo predittivo è necessario il calcolo dei coefficenti beta appartenenti a ciascun regressore del modello finale nonché dei relativi intervalli di confidenza , seguirà il calcolo delle probabilità.

Il modello finale è il seguente (con arrotondamento alla seconda cifra decimale):

$$Y = -0.60 - 0.31X_1 + 1.1X_2 + 2.28X_3 - 0.59X_4 + 0.39X_5 + 0.48X_6 + 0.04X_7$$

- Y = logit [prob (disdetta = “si”)]
- (-0.60) = intercetta
- X_1 = persone a carico (senza)
- X_2 = internet (ADSL)
- X_3= internet (no)
- X_4 = assistenza tecnica (no)
- X_5 = streaming film (no)
- X_6 = pagamento elettronico (no)
- X_7 = permanenza

La seguente tabella contiene l’esponenziale dei cofficenti sopra riportati ed i relativi intervalli di confidenza al 95%:

	β	Exp(β)	Estremo inferiore	Estremo superiore
Intercetta	−0.60	0.54	0.43	0.7
X_1 = persone a carico (senza)	− 0.31	0.73	0.62	0.85
X_2 = internet (ADSL)	+1.1	3.02	2.61	3.49
X_3= internet (no)	+2.28	9.79	7.75	12.38
X_4 = assistenza tecnica (no)	−0.59	0.55	0.47	0.65
X_5 = streaming film (no)	+0.39	1.48	1.28	1.70
X_6 = pagamento elettronico (no)	+0.48	1.61	1.40	1.86
X_7 = permanenza	+0.04	1.04	1.04	1.05

Le cifre sopra riportate sono state arrotondate al secondo decimale.

Segue l'interpretazione dei coefficienti, passaggio chiave per comprendere l'impatto di ciascun regressore sulla decisione finale del cliente.

Si rileva un maggiore grado di soddisfazione tra i giovani ed in generale coloro i quali non hanno persone a carico. La propensione dei clienti a disdire se non hanno persone a carico è circa 0.73 volte l'analoga propensione di coloro i quali hanno persone a carico. L'intervallo di confidenza va da 0.62 a 0.85. In altre parole, la tendenza al churn tra coloro che non hanno persone a proprio carico è almeno del 38% al massimo del 15% inferiore rispetto a coloro i quali hanno persone a carico.

E' evidente come la customer satisfaction sia decisamente bassa tra quanti utilizzano ancora l'ADSL rispetto ai clienti fibra. La propensione a disdire tra coloro i quali hanno l'ADSL è 3.02 volte maggiore l'analoga tendenza di coloro i quali navigano utilizzando la fibra ottica. Con un intervallo di confidenza che va da 2.61 a 3.49. **La propensione alla rescissione del contratto aumenta ancora di più tra coloro i quali sono sprovvisti di una connessione rete rispetto a coloro che utilizzano la fibra.** Infatti, la propensione a disattivare tra coloro che non usufruiscono della connessione telecom è 9.79 volte maggiore la medesima tendenza di quanti navigano con la connesione fibra di ultima generazione. Con un intervallo di confidenza al 95% che va da 7.75 a 12.38.

I servizi di attenzione al cliente non sembrano incrementare la soddisfazione dello user. La propensione a disdire tra coloro che non usufruiscono del servizio di assistenza tecnica è 0.55 volte inferiore rispetto a quanti utilizzano tale servizio. In termini percentuali la propensione a disdire per il primo gruppo è 45% inferiore l'analoga tendenza per il secondo gruppo. Con un stima intervallare compresa tra 0.47 e 0.65.

Emerge l'importanza dei nuovi servizi di intrattenimento in streaming. La propensione al churn tra coloro i quali non hanno attivato il servizio di film in streaming è 1.48 volte la tendenza dei customer che invece usufruiscono del servizio. Con un intervallo di confidenza al 95% che va da 1.28 e 1.70. In altre parole, la propensione a disdire per i clienti privi di servizio film in streaming è almeno il 28% superiore e al massimo il 70% superiore rispetto a coloro che invece ne usufruiscono.

Le vecchie forme di pagamento rendono assai più insidiosa l'esperienza del cliente. La tendenza alla disattivazione da parte dei clienti che non effettuano pagamento in forma elettronica è 1.61 volte l'analoga propensione di coloro che invece pagano in forma digitale. Con un intervallo di confidenza che va da 1.40 a 1.86 si può affermare che la tendenza alla disdetta tra quanti non utilizzano pagamento elettronico è almeno il 40% e al massimo l'86% maggiore dell'analoga propensione tra quanti invece pagano in forma elettronica.

Col passare del tempo il cliente mostra segnali di insoddisfazione. La propensione a disdire tra coloro che sono clienti da un mese in più è 1.04 volte l'analoga propensione di quelli che sono customer da un mese in meno. Con un intervallo di confidenza che va da 1.04 a 1.05 si può dimostrare come la tendenza al churn di coloro che sono clienti da un mese in più è almeno del 4% e al massimo del 5% maggiore la medesima propensione tra quanti sono clienti da un mese in meno.

E’ possibile ottenere ulteriori spunti informativi dal modello stimato esprimendo i risultati in termini probabilistici ed adattandoli ai diversi casi.
Si parte dalla profilazione del **worst-case scenario** ossia un cliente con persone a carico, senza connessione internet, che usufruisce del servizio di assistenza tecnica, che non ha attivato il pacchetto di film in streaming , che paga in forma tradizionale (non digitale) e che è legato alla società contrattualmente da un anno. **Ebbene in questo caso la probabilità di disdire l’abbonamento è di circa il 96%** (almeno del 94% e al massimo del 97%).

```
> #caso peggiore dopo un anno
> prob1<-predict.glm(logistic11, type="response", newdata = data.frame(personeacarico="con",internet="no", assistenzatecnica="si",streamingfilm="no",pag
amentoelettronico="no",permanenza=12), se=TRUE)
> prob1
$fit
       1
0.9560828

$se.fit
         1
0.005934783

$residual.scale
[1] 1

> L1<-0.9560828-(1.96*0.005934783 );L1
[1] 0.9444506
> U1<-0.9560828+(1.96*0.005934783 );U1
[1] 0.967715
```

Situazione che peggiora col passare dei mesi, infatti, a distanza di due anni la probabilità di churn si attesta attorno al 97% (almeno il 97% al massimo il 98%)

```
> #caso peggiore dopo 2 anni
> prob2<-predict.glm(logistic11, type="response", newdata = data.frame(personeacarico="con",internet="no", assistenzatecnica="si",streamingfilm="no",pag
amentoelettronico="no",permanenza=24), se=TRUE)
> prob2
$fit
       1
0.9738419

$se.fit
          1
0.003556078

$residual.scale
[1] 1

> L2<-0.9738419 -(1.96*0.003556078  );L2
[1] 0.966872
> U2<-0.9738419 +(1.96*0.003556078  );U2
[1] 0.9808118
```

L’altro lato della medaglia è invece il **best-case scenario** ovvero il cliente senza persone a suo carico, con connessione fibra, che non usufruisce dell’assistenza tecnica, che gode del servizio film in streaming , che paga in forma elettronica e che è presente all’interno della base clienti da un anno. **Questa situazione coincide con un’elevata customer satisfaction dal momento che la probabilità di disdire il contratto di abbonamento è solo del 27%** (almeno del 25% al massimo del 30%).

```
> #profilo young fibra dopo un anno (caso migliore)
> prob3<-predict.glm(logistic11, type="response", newdata = data.frame(personeacarico="senza",internet="fibra", assistenzatecnica="no",streamingfilm="si",p
agamentoelettronico="si",permanenza=12), se=TRUE)
> prob3
$fit
        1
0.2731791

$se.fit
         1
0.01363701

$residual.scale
[1] 1

> L3<-0.2731791-(1.96*0.01363701);L3
[1] 0.2464506
> U3<-0.2731791+(1.96*0.0136370);U3
[1] 0.2999076
```

Tuttavia a distanza di due anni la medesima probabilità si aggira attorno il 39% (almeno del 36% al massimo del 42%).

Ad ogni modo **la connessione fibra, lo scarso utilizzo dei servizi di assistenza tecnica e la facilità dei pagamenti elettronici hanno nel complesso un impatto evidente.**
Infatti la probabilità dello user di abbandonare i servizi telecom a distanza di un anno cresce al 77% per coloro i quali hanno connessione adsl, hanno usufruito dell'assistenza e che pagano in forma tradizionale.

```
> #profilo young adsl, con assistenza tecnica e senza pagamento elettronico dopo un anno
> prob5<-predict.glm(logistic11, type="response", newdata = data.frame(personeacarico="senza",internet="adsl", assistenzatecnica="si",streamingfilm="si",pa
gamentoelettronico="no",permanenza=12), se=TRUE)
> prob5
$fit
        1
0.7675702

$se.fit
         1
0.01897669

$residual.scale
[1] 1

> L5<-0.7675702-(1.96*0.01897669);L5
[1] 0.7303759
> U5<-0.7675702+(1.96*0.01897669);U5
[1] 0.8047645
```

L'analisi dei dati ha messo in risalto queste tendenze è ora il momento di valutare possibili soluzioni.

Soluzioni proposte e churn prevention

Il modello realizzato ha offerto una prospettiva interessante in termini di churn prediction ed indica con chiarezza la direzione verso la quale dirigere le risorse aziendali e nel breve e nel lungo periodo. Di seguito vengono avanzate agli stakeholders una serie di proposte coerenti con i risultati dell'analisi effettuata.
Le soluzioni individuate possono essere sintetizzate nei seguenti punti:

- Ridefinizione del key product aziendale: servizi di connessione rete internet veloce;
- Focus su profili young e family;
- Investire sulle piattaforme di intrattenimento multimediale;
- Conversione tempestiva dei customer ADSL e privi di connessione;
- Customer Success;
- Promuovere e facilitare forme di pagamento elettronico.

In primo luogo occorre **porre al centro del core business i servizi di connessione rapida** dal momento che sono questi quelli che maggiormente incidono sulla customer experience e in definitiva sulla decisione del cliente di continuare ad usufruire dei servizi telecom. Inoltre, bisogna **continuare a scommettere sui profili young** dal momento che per le nuove generazioni la connessione rete è ora più che mai un bene essenziale. A tal proposito, è evidente **l'importanza delle piattaforme di intrattenimento multimediale** e in quest'ottica occore investire sui servizi di film e tv in streaming. Tuttavia, **è necessario un focus particolare sulle famiglie** e sviluppare servizi e piani tariffari adeguati. Le famiglie sono una compononente strategica della base cliente dal momento che in questi gruppi la decisione di un membro può avere un'impatto sugli altri componenti che può rivelarsi positivo o negativo per l'impresa.
E' necessario **promuovere con rapidità la conversione alla rete fibra di tutta la customer base** dal momento che coloro i quali ne sono sprovvisti sono maggiormente esposti al rischio churn.
La tempestività è giustificata dall'agguerrita concorrenza sui servizi di connessione rete veloce.
C'è ampio margine di miglioramento per i servizi di assistenza tecnica e assistenza cliente dal momento che i servizi attuali non sembrano favorire l'experience dei clienti. Per il coseguimento di tale scopo si richiede un cambio di filosofia del tradizionale customer care in un'ottica indirizzata verso il cosiddetto "**customer success"** che si basa sulla tempestività e la pianificazione nel lungo termine: giocare sulla previsione e soddisfare il cliente PRIMA che esso si metta in contatto con l'azienda. Lo spostamento verso servizi di natura digitale renderà questa soluzione assai più coerente. In definitiva, la customer experience risulterà essere assai più funzionale se accompagnata da dinamiche customer-business più smart, a cominciare dalla **promozione e la facilitazione all'uso di forme di pagamento elettroniche.**