Insoddisfazione dei Clienti di un Provider Telefonico

Progetto II per il corso di Statistica

Daniele Giaquinta
Corso di laurea in Artificial Intelligence and Data Engineering

Indice

1 Abstract

Lo scopo di questa relazione è studiare l'incidenza di alcuni fattori sulla scelta di alcuni clienti anonimi di un provider telefonico di disdire il servizio. I fattori presi in analisi sono dati sul servizio a loro erogato e sull'assistenza ricevuta.

L'utilità di tale ricerca è ovvia; potrebbe permettere a tali aziende di migliorare il servizio offerto imparando a comprendere le situazioni che generano malcontento nella clientela.

2 Dati

La tabella di dati è reperibile all'indirizzo https://data.world/bob-wakefield/call-center-data; gli attributi della tabella sono i seguenti:

- ~~**recordId**: Chiave primaria della tupla~~
- ~~**state**: Stato (americano) di provenienza del cliente~~

- **account_length**: Età dell'account del cliente in mesi
- ~~**area_code**: Codice identificativo dell'area~~
- **international_plan**: Il cliente ha o meno una tariffa internazionale
- **voice_mail_plan**: Il cliente ha o meno una tariffa vmail
- **number_vmail_messages**: Numero di vmail del cliente presenti sul server
- **total_day_minutes**: Minuti di chiamata durante il giorno
- **total_day_calls**: Numero di chiamate durante il giorno
- **total_day_charge**: Spesa del cliente i servizi durante il giorno
- **total_eve_minutes**: Come sopra, durante la sera
- **total_eve_calls**: Come sopra, durante la sera
- **total_eve_charge**: Come sopra, durante la sera
- **total_night_minutes**: Come sopra, durante la notte
- **total_night_calls**: Come sopra, durante la notte
- **total_night_charge**: Come sopra, durante la notte
- **total_intl_minutes**: Come sopra, riferito a chiamate internazionali
- **total_intl_calls**: Come sopra, riferito a chiamate internazionali
- **total_intl_charge**: Come sopra, riferito a chiamate internazionali
- **number_customer_service_calls**: Numero di chiamate al servizio clienti
- **churn**: Il cliente ha o meno reciso il contratto
- ~~**customer_id**: Identificativo del cliente~~

Gli attributi non considerati per la classificazione sono barrati; rappresentano identificativi o indicazioni geografiche e pertanto ritenute non di rilevanza per classificazione. In rosso è evidenziato l'attributo target che è binario e indica se il cliente abbia o meno disdetto il contratto con la compagnia telefonica.

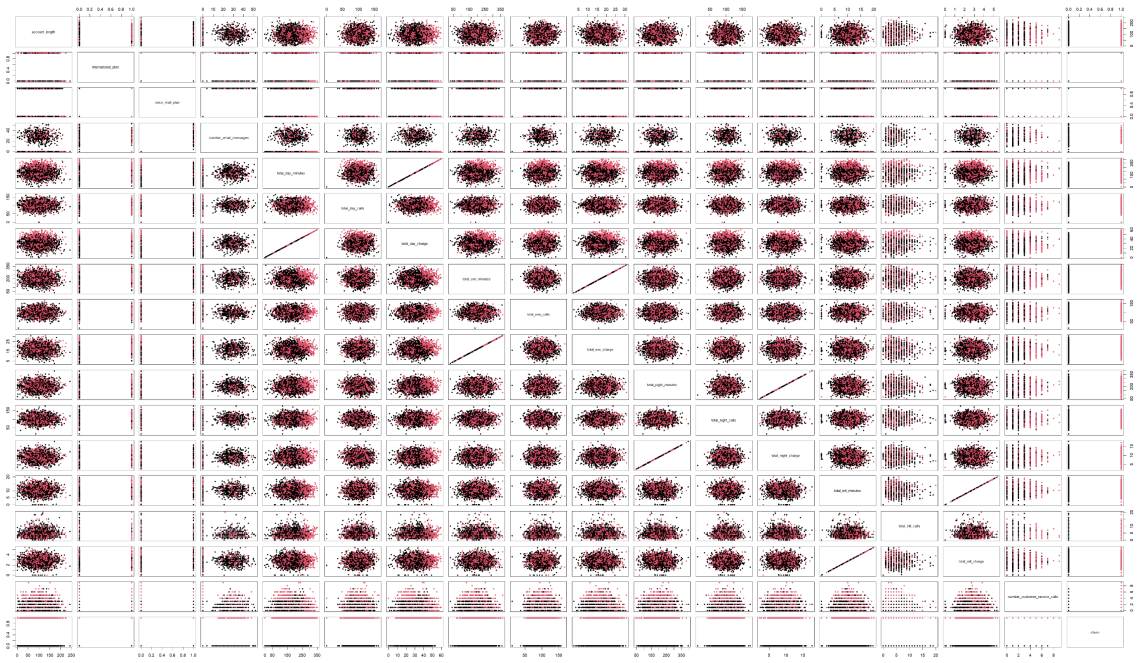

Figura 1: Diagramma di dispersione

2.1 Preprocessing dei Dati

Si è iniziato osservando il numero di occorrenze delle due classi e **il dataset risulta sbilanciato**; sono contenuti infatti 11043 esiti negativi e solo 1813 esiti positivi. Per prima cosa quindi è stato effettuato un sample degli esiti negativi di dimensione paragonabile e questo è stato unito poi agli esiti positivi in modo da bilanciare il dataset. Questa fase iniziale è stata svolta in *MySQL* e il codice si trova in appendice.

Purtroppo dal grafico di dispersione Figura 1 e dalla matrice delle correlazioni (omessa) non si traggono informazioni particolarmente utili, ma si potranno fare considerazioni sugli attributi maggiormente incidenti nella classificazione una volta osservati i resoconti delle classificazioni attraverso regressione.

3 Classificazione

L'analisi di classificazione è stata svolta mediante la creazione di 4 distinti modelli basati su regressione lineare, logistica, analisi del discriminante lineare e quadratica.

3.1 Classificazione con Regressione Lineare

Per prima cosa è stato costruito il modello di regressione lineare con tutti gli attributi; si nota dal sommario l'importanza delle feature *international_plan* e *voice_mail_plan*, prima di investigare sarà presentato anche il modello di classificazione con regressione logistica per confermare tale riscontro. Non verrà invece approfondita l'importanza dell'attributo *number_customer_service_calls* in quanto risulta

ovvia; un cliente insoddisfatto o che riscontra dei problemi col servizio chiamerà verosimilmente più spesso l'assistenza rispetto ad altri clienti. **Questo modello ha ottenuto una accuratezza di circa 77% e si nota dalla matrice di confusione che l'errore è circa equamente distribuito tra le classi** Figura 2.

```
Call:                lm(formula = churn ~ ., data = cust)

Residuals:              Min       1Q   Median       3Q      Max
                  -1.21115 -0.33931 -0.01892  0.36486  1.12488

Coefficients:                    Estimate  Std.Error t-value Pr(>|t|)
(Intercept)                   -0.6893491  0.0826418  -8.341  < 2e-16 ***
account_length                 0.0003254  0.0001784   1.823 0.068336 .
international_plan             0.3936843  0.0186926  21.061  < 2e-16 ***
voice_mail_plan               -0.3473642  0.0582792  -5.960 2.76e-09 ***
number_vmail_messages          0.0056391  0.0018744   3.009 0.002643 **
total_day_minutes              0.6143665  0.4151194   1.480 0.138967
total_day_calls                0.0002248  0.0003394   0.662 0.507916
total_day_charge              -3.6007607  2.4418793  -1.475 0.140411
total_eve_minutes             -0.0925161  0.2055965  -0.450 0.652745
total_eve_calls                0.0004002  0.0003562   1.123 0.261350
total_eve_charge               1.1017828  2.4188178   0.456 0.648774
total_night_minutes           -0.0589286  0.1082886  -0.544 0.586351
total_night_calls             -0.0003291  0.0003497  -0.941 0.346631
total_night_charge             1.3261346  2.4063610   0.551 0.581602
total_intl_minutes            -0.6050489  0.6624177  -0.913 0.361096
total_intl_calls              -0.0095808  0.0028179  -3.400 0.000681 **
total_intl_charge              2.2864163  2.4534643   0.932 0.351443
number_customer_service_calls  0.1064587  0.0044975  23.671  < 2e-16 ***
```

Accuracy: 0.7687449

	actual 1	actual 0
predicted 1	1376	405
predicted 0	437	1423

Figura 2: Sommario del modello lm, accuratezza e matrice di confusione

3.2 Classificazione con Regressione Logistica

Si procede con la classificazione tramite regressione logistica, con la quale i risultati sono molto simili (come si vedrà anche in fase di predizione). Figura 3

```
Call:          glm(formula = churn ~ ., family = binomial, data = cust)

Deviance Residuals:          Min       1Q   Median       3Q      Max
                         -2.9122  -0.8118  -0.1540   0.8586   2.7152

Coefficients:                     Estimate  Std.Error z-value Pr(>|z|)
(Intercept)                     -6.875985   0.504851 -13.620  < 2e-16 ***
account_length                   0.001905   0.001014   1.878  0.06038 .
international_plan               2.315158   0.125821  18.400  < 2e-16 ***
voice_mail_plan                 -2.039602   0.368854  -5.530 3.21e-08 ***
number_vmail_messages            0.033612   0.011560   2.908  0.00364 **
total_day_minutes                3.161767   2.384910   1.326  0.18493
total_day_calls                  0.001029   0.001972   0.522  0.60173
total_day_charge               -18.522551  14.028871  -1.320  0.18673
total_eve_minutes               -0.319656   1.190411  -0.269  0.78829
total_eve_calls                  0.002108   0.002038   1.034  0.30111
total_eve_charge                 3.840031  14.004983   0.274  0.78394
total_night_minutes             -0.182862   0.624266  -0.293  0.76958
total_night_calls               -0.002169   0.002020  -1.074  0.28298
total_night_charge               4.159760  13.872533   0.300  0.76429
total_intl_minutes              -2.374650   3.813671  -0.623  0.53350
total_intl_calls                -0.053156   0.016225  -3.276  0.00105 **
total_intl_charge                9.067644  14.125467   0.642  0.52091
number_customer_service_calls    0.623645   0.029972  20.808  < 2e-16 ***
```

Accuracy: 0.7681956

	actual 1	actual 0
predicted 1	1381	412
predicted 0	432	1416

Figura 3: Sommario del modello glm, accuratezza e matrice di confusione

L'accuratezza è leggermente inferiore (ma comunque di un valore trascurabile) e l'errore è leggermente più distribuito sui clienti che non disdicono rispetto a quelli insoddisfatti, su cui la precisione è di poco migliorata. Ciò che si nota è la confermata importanza degli attributi *international_plan* e *voi-*

ce_ mail_ plan, quindi sarà investigata in breve la variazione delle classi al variare di questi due attributi binari Figura 4:

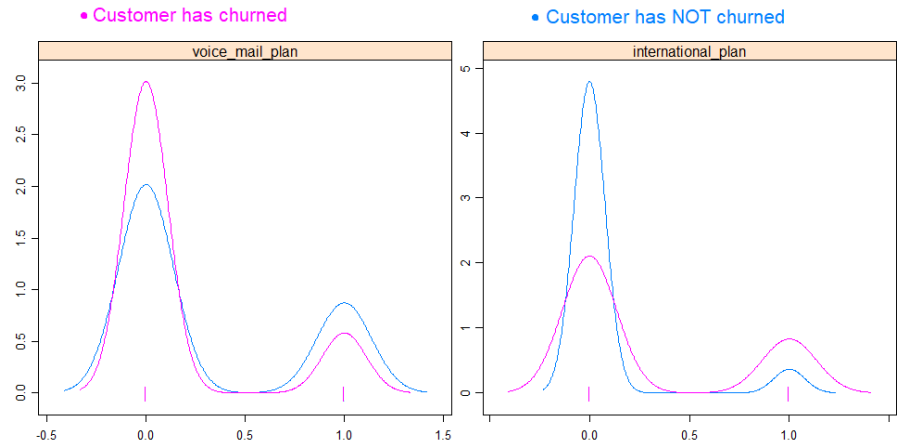

Figura 4: Densità delle classi in base ai valori dei due attributi binari indicati

Si nota nei due plot una tendenza opposta, infatti tra i clienti con un piano che comprende i voice-mail (ovvero messaggi di segreteria telefonica) sono più quelli che hanno disdetto il servizio di quelli che non lo hanno fatto; tra coloro che non possedevano tale piano tariffario la tendenza è opposta. Per quanto riguarda il piano che comprende accordi sulle chiamate internazionali sono molti i clienti che hanno deciso di non lasciare il servizio; tra coloro che non possedevano tale piano sono più numerosi i clienti che hanno disdetto. **È presumibile pensare che il piano voice-mail non sia molto conveniente secondo i clienti e che, al contrario, quello internazionale lo sia.**

3.3 Classificazione con Analisi del Discriminante Lineare

Questo modello si basa sull'ipotesi che le covarianze degli attributi siano molto paragonabili, il che probabilmente è falso considerando la compresenza di attributi binari e numerici, ma riportiamo comunque il modello per notare la differenza con quello quadratico Figura 5.

Accuracy: 0.7687449

	actual 1	actual 0
predicted 1	1376	405
predicted 0	437	1423

Figura 5: Accuratezza e matrice di confusione (lda)

I valori numerici di sensibilità, specificità e accuratezza sono esattamente gli stessi del modello con regressione lineare, anche se come si vedrà dalla curva ROC e dall'area sotto la curva i modelli non si comportano esattamente nello stesso modo.

3.4 Classificazione con Analisi del Discriminante Quadratica

Vengono infine presentati i risultati nel caso della qda che ottiene risultati leggermente migliori Figura 6.

Accuracy: 0.8305411

	actual 1	actual 0
predicted 1	1559	363
predicted 0	254	1465

Figura 6: Accuratezza e matrice di confusione (qda)

Con una **accuratezza più alta di poco più del 6%** rispetto all'analisi lineare si è deciso di rigettare l'ipotesi di similitudine delle covarianze degli attributi; questo punteggio è ritenuto finalmente soddisfacente.

4 Confronto dei Modelli e Predizione

Adesso andranno tratte conclusioni sul modello migliore sfruttando la curva ROC e la validazione, benché il vincitore risulta già chiaro.

4.1 Curve ROC e valori AUC

Sono state prodotte le quattro curve dei modelli e si nota chiaramente che la migliore sia quella ottenuta con analisi del discriminante quadratica Figura 7. Si nota inoltre che benché le altre curve sembrino quasi uguali, il percorso che attraversano non è il medesimo e se ne ha riscontro nei valori AUC.

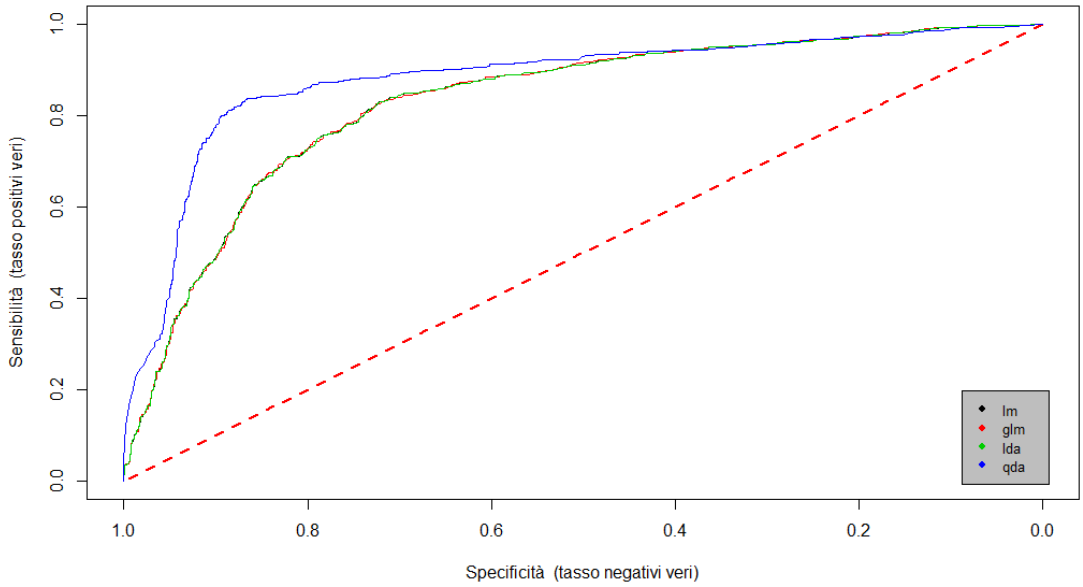

Figura 7: Curve ROC dei quattro modelli

	AUC
lm	0.7734487
glm	0.8303226
lda	0.8301046
qda	0.8814913

Il modello migliore risulta in modo evidente quello che sfrutta l'analisi del discriminante quadratica, e questo verrà confermato nella validazione.

4.2 Validazione

Si è deciso di dividere il dataframe in 1/2 e 1/2 per training e testing. La validazione è stata ripetuta 100 volte effettuando sempre sampling diversi; sono state quindi ottenute informazioni sulle accuratezze e sulle deviazioni standard degli errori dei modelli in un contesto più simile alla realtà.

	Accuratezza media	Deviazione standard errore
lm	0.763214	0.007895
glm	0.763643	0.007631
lda	0.763209	0.007898
qda	0.819022	0.008423

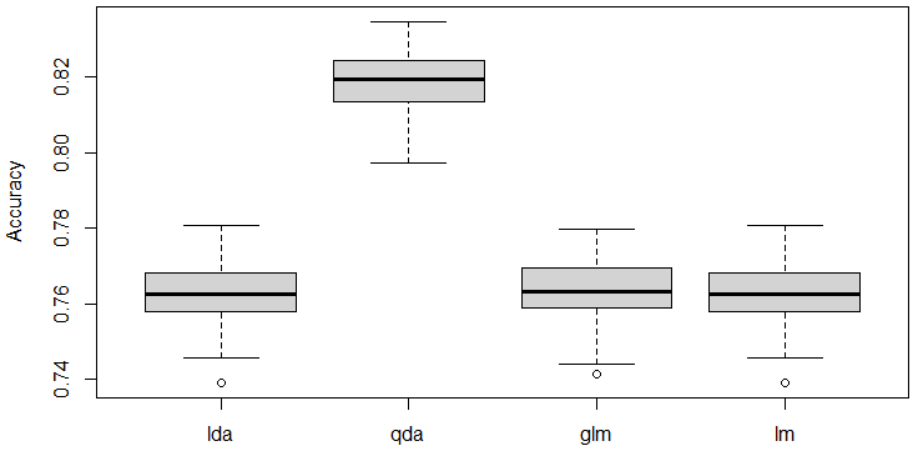

Figura 8: Boxplot delle accuratezze dei modelli

I risultati confermano quanto detto e sebbene il modello con analisi del discriminante quadratica abbia una deviazione standard dell'errore leggermente superiore agli altri modelli, offre una accuratezza media, minima e massima Figura 8 decisamente superiori a quelle dei modelli concorrenti.

4.3 Robustezza dei Dati

Inoltre è stato eseguito il test per misurare la robustezza dei dati, invertendo la classe a sempre più tuple (con un passo di 1) per osservare l'effetto sull'accuratezza dei modelli Figura 9.

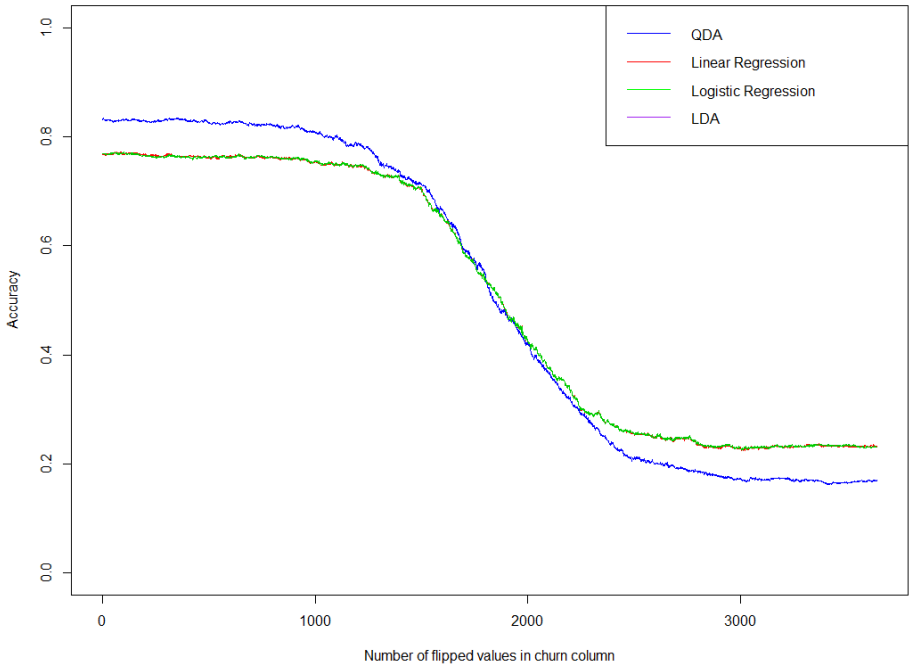

Figura 9: Esito del test di robustezza

Il modello più robusto risulta ancora una volta quello ottenuto con analisi del discriminante quadratica in quanto l'andamento osservato e la pendenza della discesa dell'accuratezza risultano circa i medesimi in tutti i modelli ma partendo da un valore di accuratezza più alto per il modello scelto.

5 Conclusioni

Rimangono pochi dubbi sul modello da scegliere, quindi traiamo delle considerazioni generali: l'accuratezza ottenuta è soddisfacente, non ottima considerando che si tratta di una classificazione binaria. È probabile che non tutti gli attributi siano influenti e forse si potrebbe effettuare una riduzione per ridurre il carico computazionale, ma in questo contesto difficilmente questo potrebbe risultare problematico. Si possono trarre informazioni utili (nel caso specifico di questa azienda) relative alla sconvenienza della tariffa vmail e alla convenienza della tariffa internazionale; questo indica su cosa "giocare" per minimizzare il numero di clienti persi (migliorando ad esempio il primo) e massimizzare il profitto (peggiorando ad esempio il secondo).

Appendice

Codice MySQL

```
#LO SPAZIO DI LAVORO DOVE E' STATA IMPORTATA LA TABELLA
USE rstudio;

#QUERY CHE RESTITUISCE IL NUMERO DI UNI E DI ZERI NELL'
   ATTRIBUTO CHURN
SELECT
  SUM(CASE WHEN churn = 0 THEN 1 ELSE 0 END) as num_zeros,
  SUM(CASE WHEN churn = 1 THEN 1 ELSE 0 END) as num_ones
FROM cust_data;

#CREAZIONE DELLA TABELLA BILANCIATA CHE EFFETTUA UN SAMPLING
   DELLE TUPLE CON CHURN=0
CREATE TABLE cust_reduced1 AS
  SELECT * FROM cust_data WHERE churn = 1
  UNION ALL
  SELECT * FROM cust_data WHERE churn = 0 AND RAND() < 0.16;

#EXPORT IN CSV
SELECT * FROM cust_reduced INTO OUTFILE '/path/to/export/
   cust_reduced.csv'
FIELDS TERMINATED BY ',' ENCLOSED BY '"'
LINES TERMINATED BY '\n';
```

Codice R

```
library(corrplot)
library(MASS)
library(readr)
library(caret)
#CAMBIARE IL PATH PER RUNNARE IL CODICE
source("C:/Users/deft5/Desktop/s2_helper.r")
#IMPORTARE QUI LA TABELLA
cust <- read_csv("tabella.csv", col_types = cols(recordID =
   col_skip(),
                                                    state = col_
   skip(), area_code = col_skip(),

   international_plan = col_double(),
                                                    customer_id
   = col_skip()))
View(cust)
plot(cust,pch=20,col=1+cust$churn)
corrplot.mixed(cor(cust))
```

```
#CLASSIFICAZIONE CON REGRESSIONE LINEARE
cust.lm=lm(churn~.,data=cust)
summary(cust.lm)
#vediamo l'accuratezza del modello e la matrice di confusione
cust.lm.p=predict(cust.lm,type="response")
sum((cust.lm.p>0.5)==(cust$churn>0.5))/length(cust$churn)
s2_confusion(cust$churn,cust.lm.p)
#creiamo la variabile per produrre successivamente la curva
   ROC
cust.lm.roc=s2_roc(cust$churn,cust.lm.p)



#CLASSIFICAZIONE CON REGRESSIONE LOGISTICA
cust.glm=glm(churn~.,family=binomial,data=cust)
summary(cust.glm)
#vediamo l'accuratezza del modello e la matrice di confusione
cust.glm.p=predict(cust.glm,type="response")
sum((cust.glm.p>0.5)==(cust$churn>0.5))/length(cust$churn)
s2_confusion(cust$churn,cust.glm.p)
#creiamo la variabile per produrre successivamente la curva
   ROC
cust.glm.roc=s2_roc(cust$churn,cust.glm.p)

#grafico delle densit  delle classi rispetto a due attributi
   significativi nella regressione
featurePlot(x = cust[, c("voice_mail_plan", "international_
   plan")],
            y = cut(cust$churn,2),
            plot = "density",
            scales = list(x = list(relation = "free"),
                          y = list(relation = "free")),
            adjust = 1.5,
            pch = "|",
            layout = c(2, 1),
            auto.key = list(columns = 2))



#ANALISI DISCRIMINANTE LINEARE
cust.lda=lda(churn~.,data=cust,CV=F)
#vediamo l'efficacia del modello con la predizione
cust.lda.p=predict(cust.lda)
#ricaviamo anche le probabilit  (a posteriori) di essere nell
   'una o nell'altra classe
cust.lda.post=cust.lda.p$posterior[,2]
#confronto per calcolare la percentuale di accuratezza
sum((cust.lda.post>0.5)==(cust$churn>0.5))/length(cust$churn)
#vediamo la matrice di confusione
s2_confusion(cust$churn,cust.lda.post)
```

```
#variabile per la curva ROC
cust.lda.roc=s2_roc(cust$churn,cust.lda.post)



#ANALISI DISCRIMINANTE QUADRATICA
cust.qda=qda(churn~.,data=cust,CV=F)
cust.qda.p=predict(cust.qda)
cust.qda.post=cust.qda.p$posterior[,2]
#ipotesi di similarit  delle covarianze rifiutata
sum((cust.qda.post>0.5)==(cust$churn>0.5))/length(cust$churn)
s2_confusion(cust$churn,cust.qda.post)
#variabile per la curva ROC
cust.qda.roc=s2_roc(cust$churn,cust.qda.post)


#CONFRONTIAMO LE CURVE ROC
plot = s2_roc.plot(cust.lm.roc,col="black")
s2_roc.lines(cust.glm.roc,col="red")
s2_roc.lines(cust.lda.roc,col="green3")
s2_roc.lines(cust.qda.roc,col="blue")
legend("bottomright",inset=0.03,c("lm","glm", "lda", "qda"),
   col=c("black","red", "green3", "blue"),pch=c(19,19),bg="
   gray",cex=.8)
#area sotto la curva
s2_auc(cust.lm.roc)
s2_auc(cust.glm.roc)
s2_auc(cust.lda.roc)
s2_auc(cust.qda.roc)



#STABILIAMO IL MIGLIOR MODELLO TRA QUESTI 4 TRAMITE LA STIMA
   DELL'ERRORE, CIOE' LA VALIDAZIONE
l=length(cust$churn)
it=100 # numero iterazioni
nt=round(l/2) # numerosit  del test set

acc=matrix(0,it,4)
for(i in 1:it){
  idx=sample(l,nt)
  custcv=cust[-idx,]
  cust.lda=lda(churn~.,data=custcv)
  cust.lda.p=predict(cust.lda,cust[idx,])$posterior[,2]
  acc[i,1]=sum((cust.lda.p>0.5)==(cust$churn[idx]>0.5))/nt
  cust.qda=qda(churn~.,data=custcv)
  cust.qda.p=predict(cust.qda,cust[idx,])$posterior[,2]
  acc[i,2]=sum((cust.qda.p>0.5)==(cust$churn[idx]>0.5))/nt
  cust.glm=glm(churn~.,family=binomial,data=custcv)
  cust.glm.p=predict(cust.glm,cust[idx,],type="response")
  acc[i,3]=sum((cust.glm.p>0.5)==(cust$churn[idx]>0.5))/nt
  cust.lm=lm(churn~.,data=custcv)
```

```
  cust.lm.p=predict(cust.lm,cust[idx,],type="response")
  acc[i,4]=sum((cust.lm.p>0.5)==(cust$churn[idx]>0.5))/nt
}

#anche l'errore    quasi uguale in tutti e tre i modelli, un
   modello vale l'altro
#LE DIFFERENZE DEVONO ESSERE SIGNIFICATIVE E NON COME QUESTE
   PER TRARRE DELLE CONCLUSIONI SUL MIGLIOR MODELLO
paste("reg. lin.: ",round(mean(acc[,4]),6)," con sd=",round(sd
   (acc[,4]),6))
paste("reg. log.: ",round(mean(acc[,3]),6)," con sd=",round(sd
   (acc[,3]),6))
paste("lda:       ",round(mean(acc[,1]),6)," con sd=",round(sd
   (acc[,1]),6))
paste("qda:       ",round(mean(acc[,2]),6)," con sd=",round(sd
   (acc[,2]),6))

#finestra per il boxplot delle accuratezze
windows()
# boxplot delle accuratezze
boxplot(acc, main = "Boxplot of Accuracy Values", xlab = "
   Model", ylab = "Accuracy",
        names = c("lda", "qda", "glm", "lm"))



#studiamo la robustezza dei modelli permutando i dati
idx <- sample(3641, 3641)
acc_qda <- acc_lm <- acc_lr <- acc_lda <- rep(0, 3641)

for(i in 1:3641){
  custf <- cust
  custf$churn[idx[1:i]] <- ifelse(custf$churn[idx[1:i]] == 0,
   1, 0)

  # Linear regression
  custf.lm <- lm(churn~., data=custf)
  custf.lm.p <- predict(custf.lm)
  acc_lm[i] <- sum((custf.lm.p > 0.5) == (cust$churn > 0.5)) /
    length(cust$churn)

  # Logistic regression
  custf.lr <- glm(churn~., data=custf, family="binomial")
  custf.lr.p <- predict(custf.lr, type="response")
  acc_lr[i] <- sum((custf.lr.p > 0.5) == (cust$churn > 0.5)) /
    length(cust$churn)

  # Linear discriminant analysis
  custf.lda <- lda(churn~., data=custf)
  custf.lda.p <- predict(custf.lda)$posterior[,2]
  acc_lda[i] <- sum((custf.lda.p > 0.5) == (cust$churn > 0.5))
    / length(cust$churn)
```

```
  # Quadratic discriminant analysis
  custf.qda <- qda(churn~., data=custf, CV=F)
  custf.qda.p <- predict(custf.qda)$posterior[,2]
  acc_qda[i] <- sum((custf.qda.p > 0.5) == (cust$churn > 0.5))
    / length(cust$churn)
}

# plot dell'accuratezza
plot(acc_qda, type="l", col="blue", ylim=c(0,1), xlab="Number
   of flipped values in churn column", ylab="Accuracy")
lines(acc_lm, type="l", col="black")
lines(acc_lr, type="l", col="red")
lines(acc_lda, type="l", col="green3")
legend("topright", legend=c("qda", "Lin", "Logistic Regression
   ", "LDA"), col=c("blue", "red", "green", "purple"), lty=1)
```