UNIVERSITÀ DI PISA

SECONDO PROGETTO DEL CORSO DI STATISTICA

MODELLO DI CLASSIFICAZIONE PER GLI ACADEMY AWARDS

ANNO ACCADEMICO 2019-2020

STEFANO PETROCCHI

SOMMARIO

INTRODUZIONE

SCOPO DELL'ANALISI

Si ipotizza che l'analisi sia stata commissionata dagli Academy Awards, come strumento per un'individuazione automatica delle pellicole adatte per una *nomination*:

- Lo scopo dell'analisi è ottenere un modello che permetta la classificazione di film in: *adatti ad una nomination* e *non adatti per una nomination* agli oscar.
- Il classificatore dovrà avere un'*accuratezza* migliore possibile preferendo *sensibilità* a *specificità*, in modo da non trascurare film degni di nota per le nominations.

DATASET

DATASET

L'analisi è stata svolta su un dataset costituito da *326* osservazioni (film), *17* fattori numerici e un fattore contenente le etichette per la classificazione (1/3 con nomination, 2/3 senza nomination).

Il dataset è frutto di *fuzzy (con fattore 1) inner join senza duplicati* effettuati tramite la funzione *Combina Query* della sezione *Dati* di *Excel* tra varie tabelle i cui link di riferimento sono allegati in Link.txt.

FATTORI

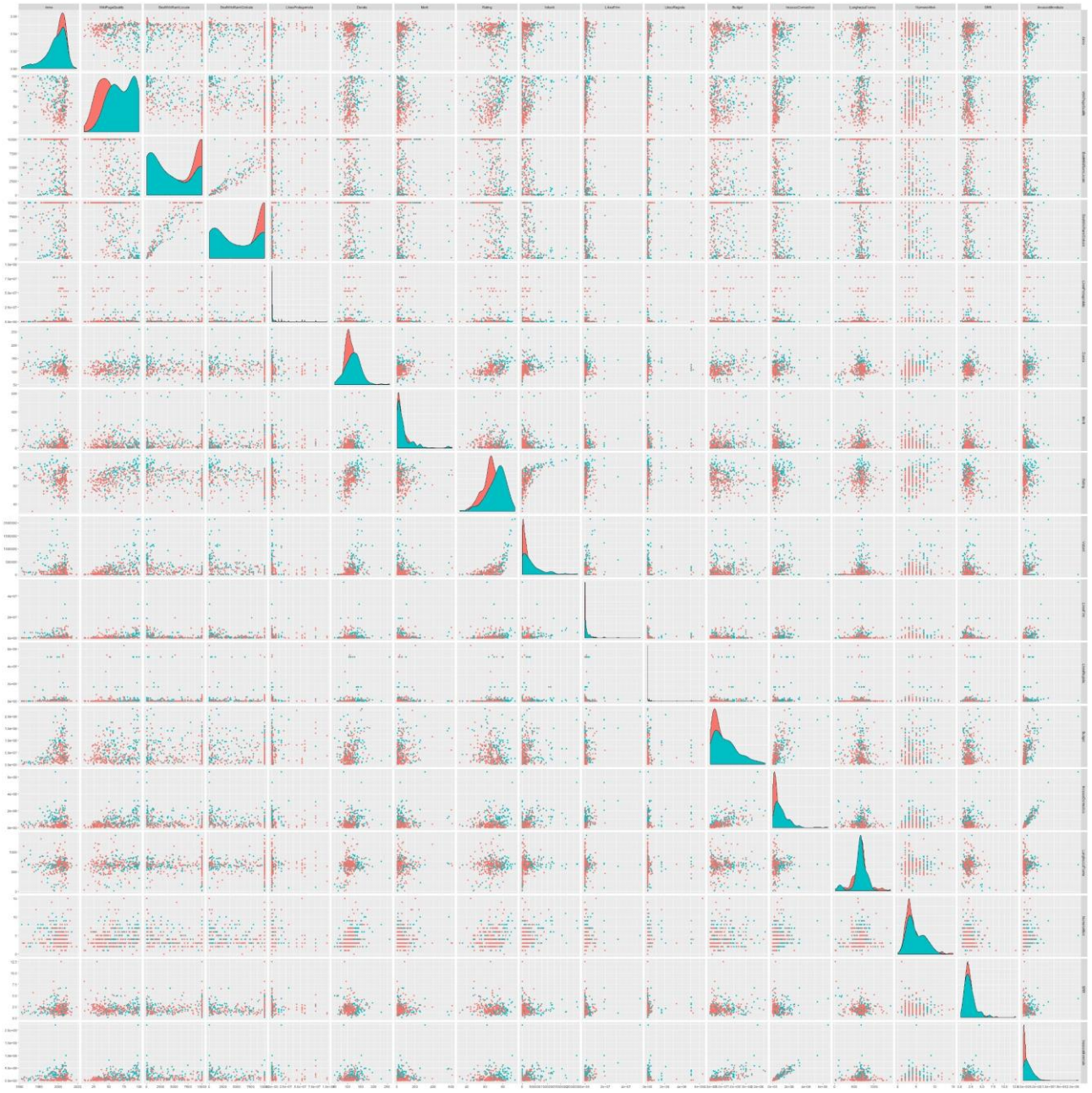
Il dataset è composto dai seguenti fattori:

- **Anno:** anno di pubblicazione del film (intero positivo).
- **Durata:** durata, espressa in minuti, del film (intero positivo).
- **Rating:** voto medio espresso sul sito IMDB da critica e utenti in una scala da 0 a 100 per il film (intero positivo).
- **Votanti:** numero di voti attraverso il quale è stato calcolato il rating del film (intero positivo).
- **LikesFilm:** numero di "likes" della pagina Facebook ufficiale o della "fanpage" con più likes del film (intero positivo).
- **LikesRegista:** numero di "likes" della pagina Facebook ufficiale o della "fanpage" con più likes del regista del film (intero positivo).
- **LikesProtagonista:** numero di "likes" della pagina Facebook ufficiale o della "fanpage" con più likes del protagonista del film (intero positivo).
- **Budget:** budget, espresso in dollari americani, ricavato da The Numbers, del film (intero positivo).
- **IncassoDomestico:** ricavi, espressi in dollari americani, nel paese d'origine del film da parte dello stesso, prelevati da The Numbers (intero positivo).
- **IncassoMondiale:** ricavi, espressi in dollari americani, ottenuti in tutto il mondo dal film, prelevati da The Numbers (intero positivo).
- **LunghezzaTrama:** numero di parole contenute della trama del film, prelevata da Wikipedia (intero positivo).
- **NumeroAttori:** numero di attori principali del film (intero positivo).
- **Morti:** numero di corpi senza vita preseti nel film, ricavati da Movie Body Counts (intero positivo).
- **SRR:** Sexual References Ratio è il numero di parole più comini riguardanti gli schemi sessuali (selezionate in questa ricerca) contenute nella trama del film, divise per la lunghezza del film, moltiplicate per 100 in modo da ottenerne la percentuale (decimale positivo).
- **WikiPageQuality:** qualità della pagina Wikipedia del film, prelevata da wikirank (decimale positivo).
- **BestWikiRankLocale:** posizione massima storica, della pagina del film, nel ranking delle pagine Wikipedia della nazione d'origine, prelevata da wikirank (intero positivo, valore massimo *9999°*).
- **BestWikiRankGlobale:** posizione massima storica globale nel ranking di Wikipedia della pagina del film, prelevata da wikirank (intero positivo, valore massimo *9999°*).
- **Nomination:** **1** se ha ricevuto una *nomination* agli Oscar, **0** se non ha mai ricevuto *nomination*.

ANALISI

ANALISI PRELIMINARE

Dallo *scatterplot* si può osservare come non sia evidente una distinzione tra film che hanno ricevuto una nomination e non, dunque il problema non risulta banale. È però da notare come la distribuzione dei film con nomination sia diversa da quelli senza nomination per alcuni fattori che potrebbero essere la chiave per un'opportuna classificazione.



CONFRONTO TRA METODI DI CLASSIFICAZIONE

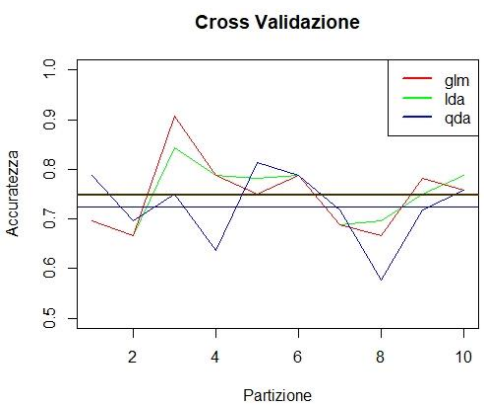
I diversi metodi di classificazione sono stati testati senza modifiche al dataset in modo da verificare se esistano differenze significative tra di essi. A tale scopo una *10-fold cross validation* permette di verificare le prestazioni dei modelli evitando problemi di **overfitting**, ma anche di *campionamento asimmetrico*. Vengono create 10 partizioni complementari del dataset con circa lo stesso numero di osservazioni. Ognuna delle partizioni costituisce un *testset* per un modello costruito tramite un *trainingset* composto dalle rimanenti osservazioni. La convalida incrociata permette infine di avere una previsione per ogni campione, rendendo i modelli confrontabili come se fosse stato utilizzato l'intero dataset come *testset* (la convalida incrociata è stata eseguita "a mano" in modo da avere maggior controllo rispetto all'utilizzo dell'apposito comando).

ANALISI DEL DISCRIMINANTE LINEARE

L'analisi porta ad un'**accuratezza** del *74,86%* con la seguente matrice di confusione:

	actual 1	actual 0
predicted 1	59	30
predicted 0	52	185

La ***specificità*** (*86%*) risulta molto maggiore della ***sensibilità*** (*52,2%*) è perciò da valutare l'utilizzo di un margine per la classificazione più basso per aumentare il numero di predizioni positive.



ANALISI DEL DISCRIMINANTE QUADRATICA

L'analisi porta ad un'**accuratezza** del *72,4%*, minore del modello lineare, con la seguente matrice di confusione:

```
            actual 1 actual 0
predicted 1       56       35
predicted 0       55      180
```

La ***specificità*** (*83,7%*) risulta molto maggiore della ***sensibilità*** (*50,5%*) è perciò da valutare anche in questo caso l'utilizzo di un margine per la classificazione diverso.

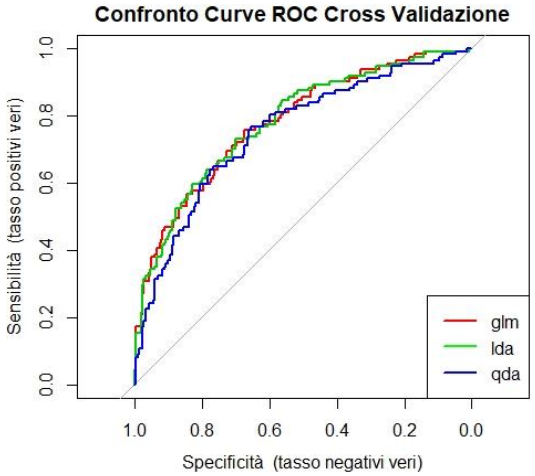
REGRESSIONE LINEARE GENERALIZZATA

L'analisi porta ad un'**accuratezza** del *74,88%* e la seguente matrice di confusione:

```
            actual 1 actual 0
predicted 1       59       32
predicted 0       52      183
```

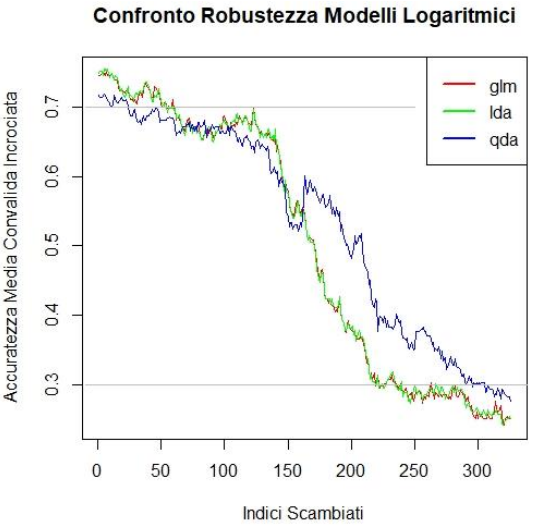
La ***specificità*** (*85,1%*) risulta anche in questo caso molto maggiore della ***sensibilità*** (*52,2%*).

Non emergono pertanto distinzioni significative tra i modelli in termini di *accuratezza* e *sensibilità*.



CONFRONTO ROC

Il confronto tra le curve **ROC** conferma la similarità tra i modelli: quello con *discriminante quadratico* (AUC *75,18%*) è quello con un risultato lievemente inferiore, rispetto al modello con *regressione generalizzata* (AUC *78,01%*) e al modello con *discriminante lineare* (AUC *78,14%*).



CONFRONTO ROBUSTEZZA MODELLI

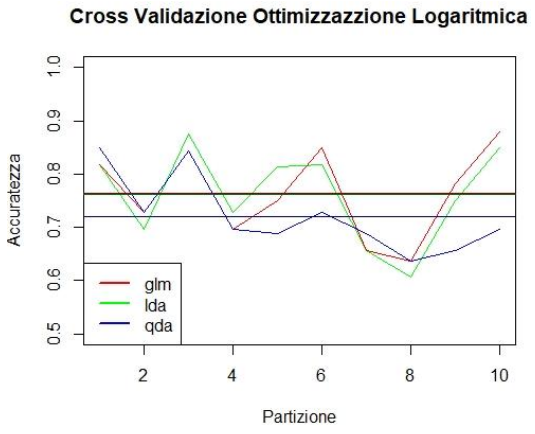
Eseguendo un test sulla **robustezza** dei modelli, che consiste nello scambiare l'etichetta di indici casuali via via crescenti, è emerso come il modello con **discriminante quadratico** si confermi essere quello con l'andamento peggiore.

A seguito dei vari test non risulta perciò prevalere nettamente nessun modello: quello con **discriminante quadratico** possiede caratteristiche leggermente inferiori, mentre i modelli di **regressione generalizzata** e con **discriminante lineare** si equivalgono.

OTTIMIZZAZZIONE E SEMPLIFICAZIONE DEI MODELLI

OTTIMIZZAZZIONE LOGARITMICA

Dati i buoni risultati dei modelli lineari sul dataset non modificato, i fattori che presentavano andamenti non lineari sono stati linearizzati applicando il logaritmo ai rispettivi valori delle osservazioni. I fattori sono stati linearizzati in modo da massimizzare la correlazione totale con *Nomination*.



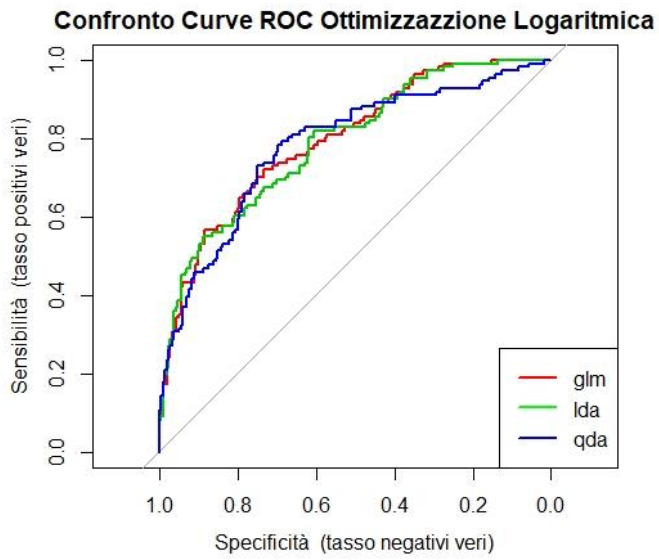
RISULTATI CONVALIDA INCROCIATA

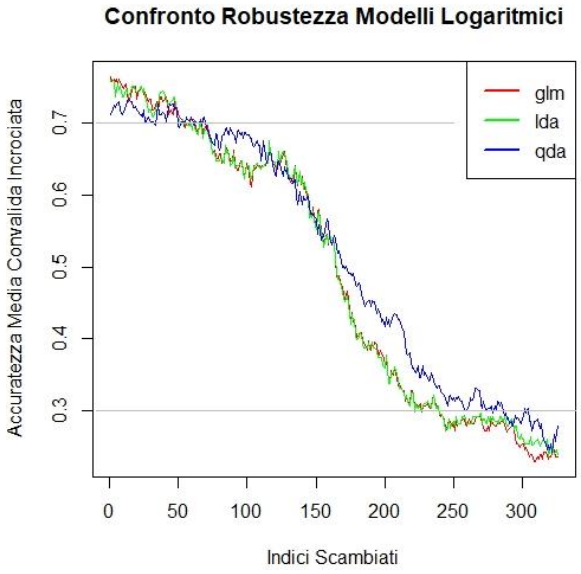
Come prevedibile i modelli lineari presentano un'**accuratezza** migliore a seguito dell'ottimizzazione: passando dal *74,86%* al *76%* per il modello con **discriminante lineare** e dal *74,88%* al *76,3%* per il modello di **regressione generalizzata**. Il modello con **discriminante quadratico** è invece sceso lievemente dal *72,4%* al *72,1%* di accuratezza.

Per quanto riguarda le curve **ROC** si può osservare un andamento meno lineare rispetto al precedente e un aumento delle **AUC**: il modello con **discriminante quadratico** è passato dal *75,18%* al *78,31%*; il

modello con **regressione generalizzata** è passato dal *78,01%* al *79,58%* e il modello con **discriminante lineare** è passato dal *78,14%* al *79,02%*.

Il profilo della **robustezza** è rimasto pressoché invariato se non per un andamento più lineare del modello con **discriminante quadratico**.

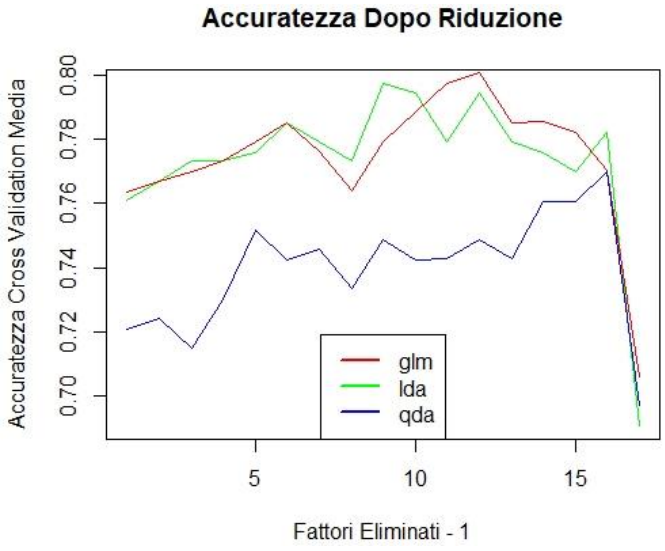




RIDUZIONE DEI MODELLI

Al fine di ottimizzare ulteriormente le previsioni, i modelli sono stati ridotti eliminando fattore per fattore quelli che catturavano la *minor variabilità* (nell'ordine: *LikesFilm, SRR, BestWikiRankGlobale, Durata, NumeroAttori, Morti, BestWikiRankLocale, WikiPageQuality, Votanti, Budget, LunghezzaTrama, LikesProtagonista, IncassoDomestico, Anno, LikesRegista, Rating, IncassoMondiale*).

A seguito dei test è stata prelevata la combinazione di fattori eliminati con *accuratezza* maggiore e *numero di fattori* minore (11 per glm, 11 per lda e 16 per qda).



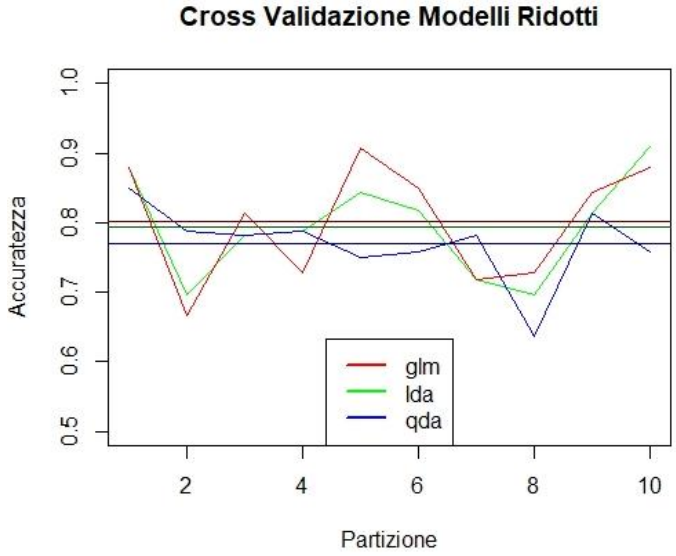
RISULTATI CONVALIDA INCROCIATA

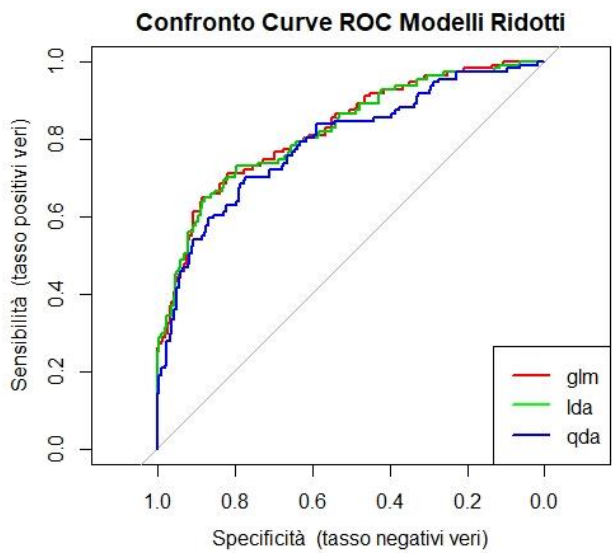
I modelli ridotti presentano un notevole aumento dell'accuratezza: passando al *79,4%* per il modello con **discriminante lineare**, all'*80%* per il modello di **regressione generalizzata** e al *77%* per il modello con **discriminante quadratico** .

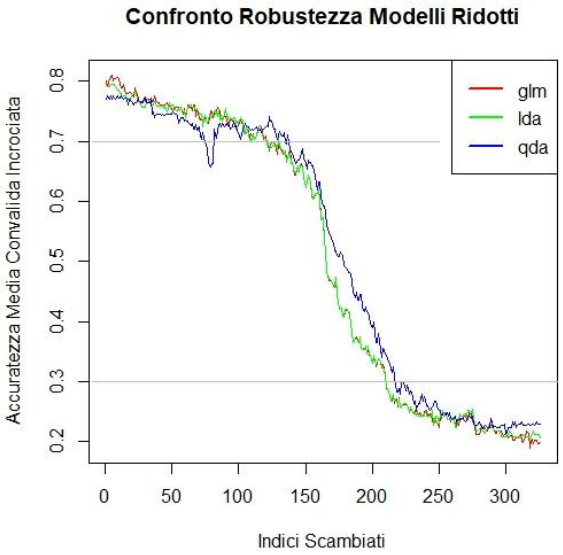
Le curve **ROC** presentano un andamento simile a quello precedente ma con valori **AUC** più elevati: il modello con **discriminante quadratico** è passato al *79,16%*; il modello con **regressione generalizzata** è passato all'*82,39%* e il modello con **discriminante lineare** è passato all'*82,08%*.

Il profilo della **robustezza** è sensibilmente migliorato, confermando la bontà dei nuovi modelli.

Seppur migliorata, la **sensibilità** dei modelli risulta comunque troppo bassa per il tipo di problema, è preferibile segnalare come una possibile nomination una pellicola senza le caratteristiche adatte, rispetto che ignorare dei film con le giuste caratteristiche. È necessario dunque trovare un equilibrio tra *sensibilità* e *accuratezza* dei modelli.







SENSIBILITÀ

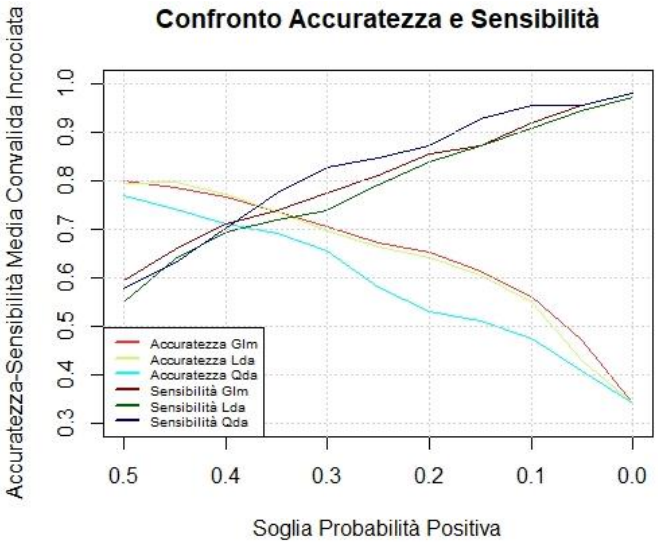
Al fine di ottenere un equilibrio tra **sensibilità** e **accuratezza** è stata eseguita una serie di *cross validation* con una probabilità minima per classificare un'osservazione come positiva (con *nomination*) via via decrescente.

Il miglior compromesso sembra verificarsi con il modello di **regressione generalizzata**, utilizzando una **soglia di probabilità** pari al *35%*.

Di seguito è riportata la matrice di confusione:

	actual 1	actual 0
predicted 1	82	51
predicted 0	29	164

La ***specificità*** (*76,3%*) seppur ridotta risulta comunque buona, mentre la ***sensibilità*** (*73,9%*) è ora notevolmente cresciuta. L'***accuratezza*** rimane accettabile al *75,5%*.



PARTIZIONI

Seppur appurato essere il modello con **regressione generalizzata** quello con le migliori caratteristiche, è interessante osservare come il modello con **discriminante quadratico,** utilizzando come fattori solamente *IncassoMondiale* e *Rating*, riesca ad ottenere degli ottimi risultati. Nel grafico a destra sono evidenziati i confini di classificazione dello spazio tra *Rating* e *IncassoMondiale* (viola classificato come possibile *nomination* e azzurro no).