PROGETTO MACHINE LEARNING

Millone A. (mat. 846588), Rossi S. (mat. 857183), Università degli Studi di Milano-Bicocca,

Laurea Magistrale in Biostatistica, Machine Learning

Contesto

Il dataset scelto proviene dalla banca di dati della University of California (UCI), contiene 2111 osservazioni analizzate su 17 variabili, questo set di dati include dati per la stima dei livelli di obesità negli individui provenienti dai paesi del Messico, Perù e Colombia, in base alle loro abitudini alimentari e alle condizioni fisiche.

I record sono etichettati con la variabile di classe NObesity (Livello Obesità), che consente la classificazione dei dati utilizzando i valori di Peso Insufficiente, Peso Normale, Sovrappeso Livello I, Sovrappeso Livello II, Obesità Tipo I , Obesità di tipo II e Obesità di tipo III. Il 77% dei dati è stato generato sinteticamente utilizzando lo strumento Weka e il filtro SMOTE, il 23% dei dati è stato raccolto direttamente dagli utenti attraverso una piattaforma web. Questi dati possono essere utilizzati per generare strumenti computazionali intelligenti per identificare il livello di obesità di un individuo e per costruire sistemi di raccomandazione che monitorino i livelli di obesità.

Le 17 variabili presenti nel dataset sono divise in due gruppi :

Gli attributi legati alle abitudini alimentari:

1. Consumo frequente di alimenti ad alto contenuto calorico (FAVC,categoriale 2 livelli),
2. Frequenza di consumo di verdure (FCVC,continua)
3. Numero di pasti principali (NCP,continua)
4. Consumo di alimenti fuori pasto (CAEC,categoriale)
5. Consumo di acqua giornaliero (CH20,continua)
6. Consumo di alcol (CALC,categoriale)

Gli attributi relativi alla condizione fisica :

7. Monitoraggio del consumo di calorie (SCC,categoriale)
8. Frequenza dell'attività fisica (FAF,continua)
9. Tempo di utilizzo di dispositivi tecnologici (TUE,discreta)
10. Fumatore ? (SMOKE,categoriale)
11. Trasporti utilizzati (MTRANS,categoriale)
12. Storia familiare con sovrappeso (family history with overweight,categoriale)
13. Sesso (Gender,categoriale 1=M;2=F)
14. Età (Age,discreta)
15. Altezza (Height,continua)
16. Peso (Weight,continua)
17. NObesity (Livello Obesità,categoriale con 7 livelli)

L'obesità è associata a una serie di problemi di salute: malattie cardiache, ictus e altro. Riconoscerla fin da subito può ridurre il rischio di insorgenza di ulteriori patologie e, in generale, migliora la vita.
Per i professionisti sanitari identificare l'obesità permette di implementare strategie preventive per gestire la patologia stessa e prevenire potenziali complicazioni mediche associate.
Si rende, quindi, necessario prevenire l'obesità con strategie che siano in accordo con lo stile di vita del paziente. L'obesità è anche associata a costi sanitari significativi, sia per il paziente che per il sistema sanitario nel suo complesso.
Il nostro obiettivo è individuare il miglior modello classificatore che identifichi correttamente le persone che sono effettivamente obese e che permetta successivamente di classificare correttamente nuove osservazioni.

Il target del modello è la variabile NObesity, per motivi pratici l'abbiamo ricodificata da categoriale a sette livelli ad una nuova variabile binaria (lv_bmi): X0 = non obeso e X1 = obeso .
Dalle statistiche descrittive e dalle distribuzioni di frequenza si nota che le variabili numeriche sono tutte asimmetricamente positive (dato che la media è maggiore della mediana), ciò vuol dire che la maggior parte delle osservazioni è concentrata intorno al primo quartile. Inoltre tutte presentano dei range molto ampi di minimo e massimo, questo è dovuto al fatto che il 77% delle osservazioni sono state create sinteticamente, quindi possono presentare valori anomali.

I partecipanti presentano un'età media di 24 anni e un peso medio di 84 kg. Analizzando le deviazioni standard di tutte le variabili non sorge il problema della near-zero variance, ossia quando una variabile nel dataset ha una varianza molto bassa o quasi nulla. Questo può causare problemi durante l'addestramento di alcuni modelli di machine learning, in quanto la variabile non fornisce informazioni significative per fare previsioni accurate.

	Media	1° quartile	Mediana	3° quartile	sd	min	Max
Age	24.391	19.12	22.185	26.00	19.830	1.63	477.06
Height	13.598	1.633	1.710	1.786	43.286	1.450	187.407
Weight	84.21	60.00	80.726	105.03	53.328	1.03	860.80
FCVC	11.439	2.000	2.497	3.000	45.186	1.000	299.448
NCP	9.407	2.764	3.000	3.000	42.308	1.000	398.955
CH20	11.850	1.634	2.000	2.619	44.966	1.000	295.311
FAF	6.557	0.124	1.000	1.800	30.530	0.000	289.118
TUE	3.036	0.000	0.625	1.000	19.447	0.000	199.219

Tab. 1 statistiche descrittive

Per quanto riguarda le variabili categoriali le persone si dividono:

- Per il sesso (Gender): il 49% sono femmine e il 51% sonomaschi;
- Per famigliarità con l'obesità (family history with overweight): l'82% ha suscettibilità familiare e il 18% no ;
- Per consumo frequente di alimenti ad alto contenuto calorico (FAVC): l'88% ha un consumo elevato e il restante 12% no;
- Per consumo di alimenti fuori pasto (CAEC): l'83% dei soggetti lo fa qualche volta, l'11 lo fa invece frequentemente;
- Per chi è fumatore (SMOKE); il 97% non è fumatrice;
- Per Monitoraggio del consumo di calorie (SCC): solo il 5% svolge questo controllo ;
- Per Consumo di alcolici (CALC): il 66% della popolazione qualche volta consuma alcolici, il 30% è astemia;
- Per trasporti pubblici utilizzati (MTRANS): il 74% usa trasporti pubblici, il 21% usa l'automobile, infine con percentuali basse (intorno al 2%) i soggetti si muovono o con bicicletta o con moto o camminando;

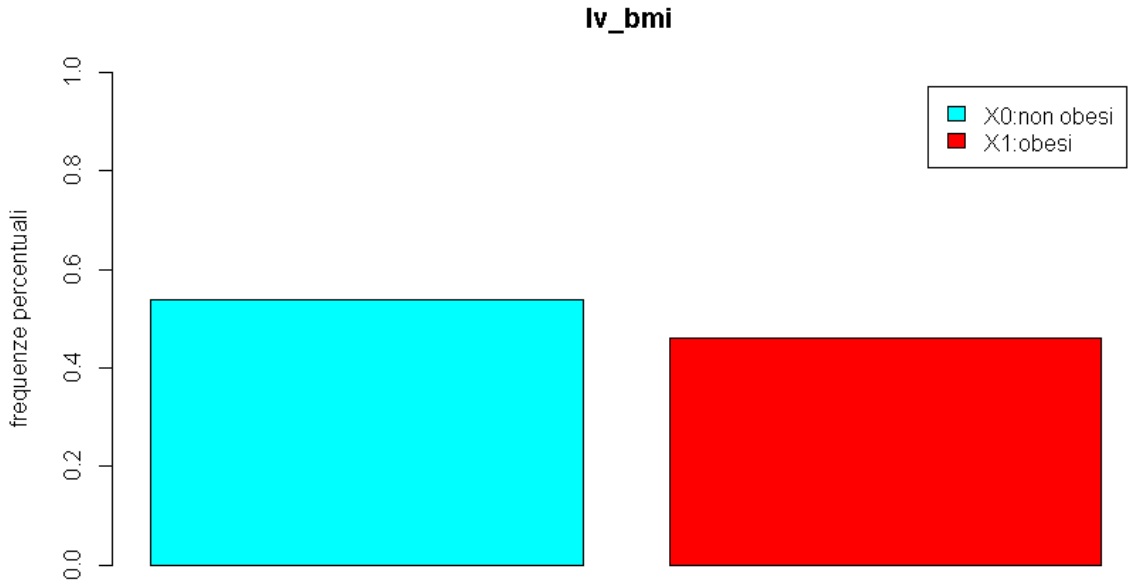

Figura 1 grafico delle percentuali di classe di lv_bmi

Il grafico mostra che per la variabile lv_bmi i due livelli X0 e X1 sono abbastanza bilanciati (XO pari al 53%, X1 al 47%).

Osserviamo le distribuzioni univariate rispetto alle classi del target (rosa = obeso, blu = non obeso) e notiamo che solo la variabile Weight (peso in kg) sembra discriminare le osservazioni tra le classi del target. Le altre variabili discriminano poco le classi del target.

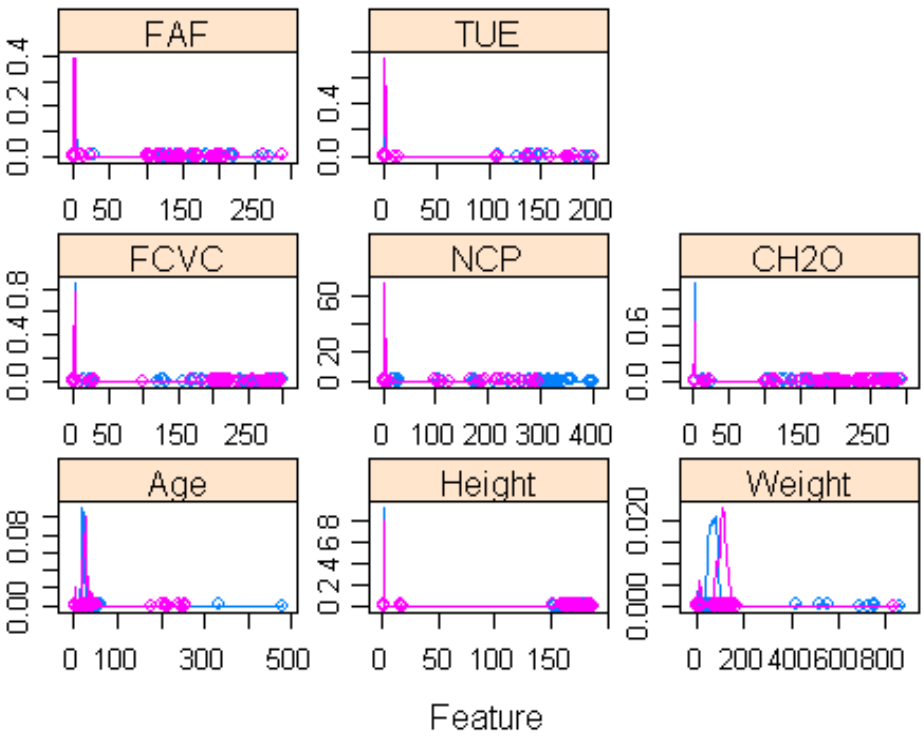

Figura 2 Grafici univariati delle variabili continue rispetto alle classi della target

Di seguito i 4 step principali da svolgere quando si vuole creare, allenare e valutare un modello di machine learning:

1. Preprocessing e training del modello si fanno tanti modelli, si scelgono i parametri di tuning del modello e si creano i primi modelli;
2. Valutazione dei modelli migliori sui dati di validation si valutano performance del modello su dati indipendenti (metriche di assesment, metriche robuste)
3. Valutazione performance di classificazione del modello migliore scelto nello step 2
4. Si applicare il miglior modello ai dati nuovi e si sceglie la soglia definitiva; infine si utilizza la nuova soglia nel dataset di score.

Preprocessing: Analisi dati mancanti

Dai grafici sottostanti si può osservare che nessuna variabile presenta valori mancanti, non è necessario utilizzare alcun metodo di imputazione.

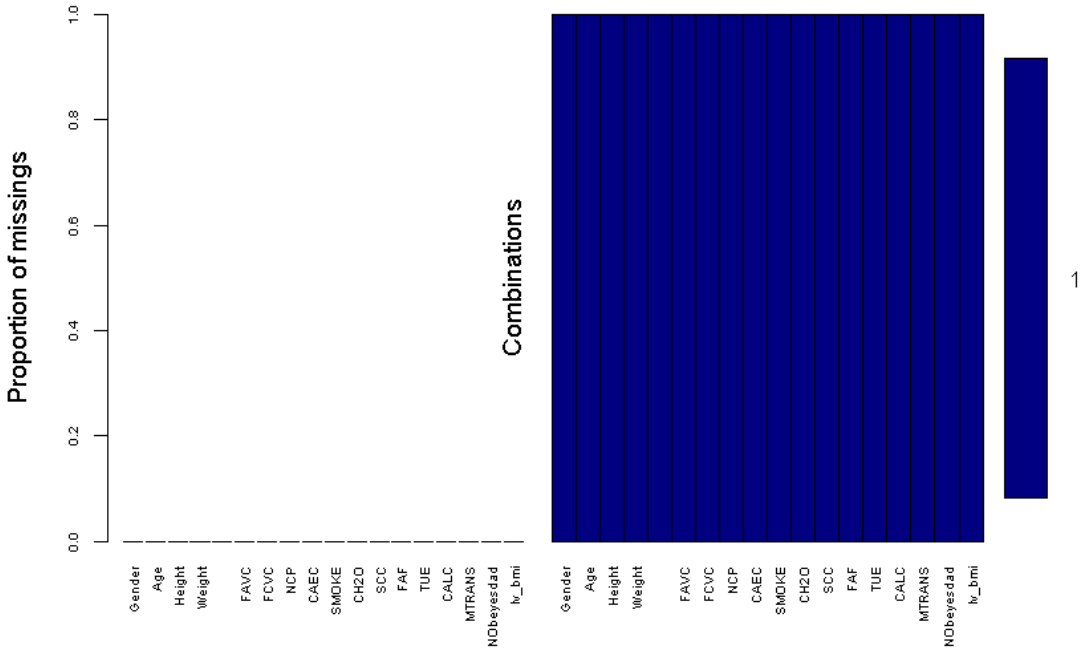

Figura 3 Grafico dei missing values

Preprocessing: Collinearità

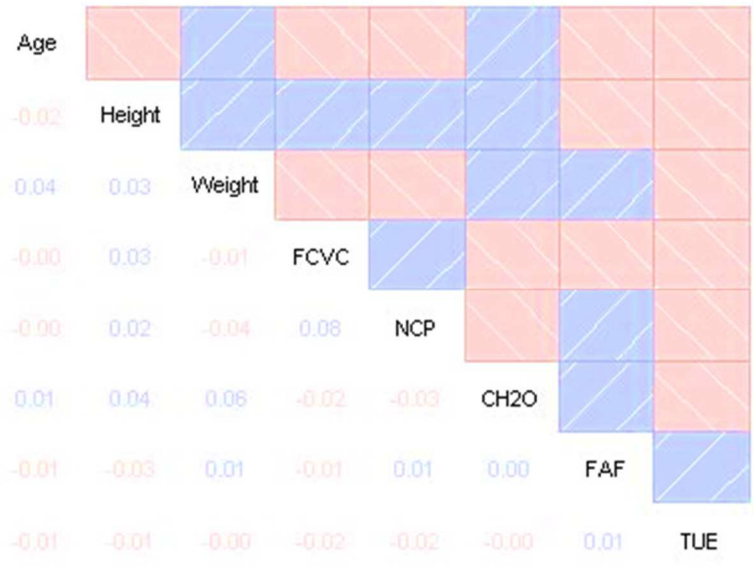

Figura 4 grafico delle correlazioni

Dalla matrice di correlazione non si evidenziano nessuna correlazione bivariata forte, i valori non superano il 10%. Possiamo, quindi, affermare l'assenza del "problema" della collinearità.

Dopo aver tolto variabili inutili o confondenti ,controllato i missing data,controllato la collinearità e la near 0 variance, dividiamo il dataset in dati di training (60%), dati di validation (30%) e dati di test(10%).

STEP 1: COSTRUZIONE dei MODELLI DI CLASSIFICATIVI

Per selezionare le covariate si usa un metodo non parametrico, un modello ad albero, ossia un metodo non analitico che permette di individuare le variabili più importanti ed eliminare eventualmente il problema della collinearità e 0 variance.
Con questa tecnica, l'importanza di una variabile è data dalla somma dei $\Delta Gini$ di ogni singola variabile, ossia la somma dei decrementi di "impurità" che la variabile genera nei vari piani dell'albero.
Il grafico mostra le variabili più importanti nel nostro studio:

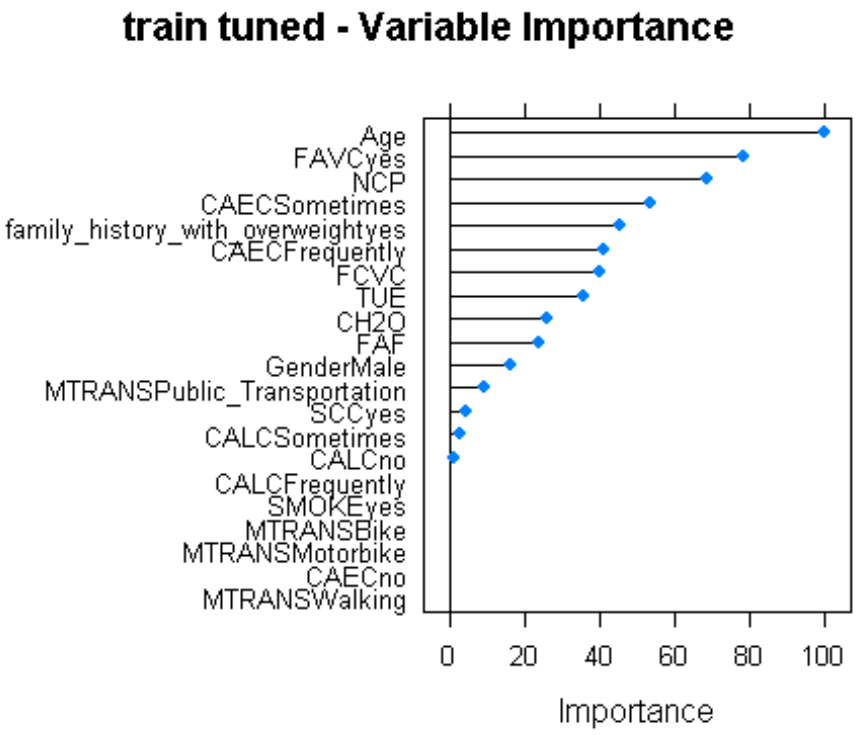

Figura 5 Grafico importanza variabili

Con questa procedura abbiamo selezionato le variabili maggiormente importanti, prendiamo queste variabili e creiamo un nuovo dataset
che servirà per i modelli non robusti.

Adesso procediamo a svolgere i vari algoritmi di classificazione, ogni algoritmo è stato sviluppato con la tecnica *Cross Validation a 10 Fold*, questa procedura consiste nel dividere il dataset in uso in 10 parti uguali: 9 parti saranno utilizzate come training (quindi per allenare il modello) e una parte come validation test, in questo modo si ottengono valori previsti del target non distorti/robusti. Per scegliere i parametri di tuning dei modelli si utilizza la metrica ROC.
I modelli calcolati sono: modello logistico, LDA, QDA, Pls, Lasso, Ridge Regression, albero "classico", Random Forest , Gradient Boosting, Xgboost, Reti Neurali, Knn, Naive Bayes e SVM.
Per ogni modello è stata utilizzata una griglia di parametri di tuning specifici, al fine di trovare il modello che classifica al meglio le due classi della variabile target.

Di seguito si riporta il grafico *BW PLOT* con le seguenti metriche: curve ROC, sensitivity e specificity.

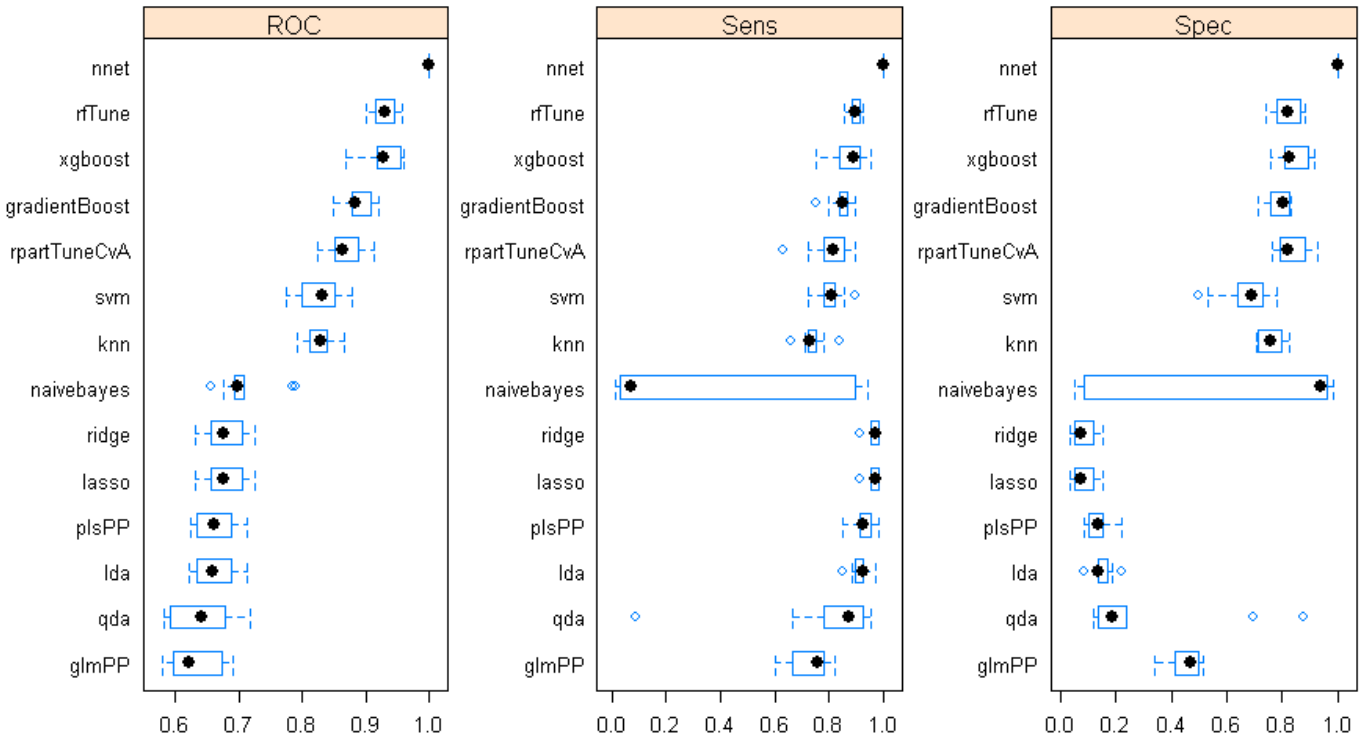

Figura 6 Valori delle metriche ROC sensitivity e specificity dei modelli

Da una prima analisi dei risultati dei modelli, quelli che performano meglio su tutte e tre le metriche mostrate sono i modelli basati sulle reti neurali e sugli alberi (random forest, gradient boosting xgboost, albero "classico"), invece i restanti , come ad esempio il pls, non ottengono performance buone in tutte le metriche.
Da notare che tutti i modelli (tranne il naive bayes) ottengono valori di sensibilità superiori a 0.75, cioè individuano bene la quota di soggetti obesi che sono effettivamente obesi, i cosiddetti *veri positivi*; mentre ottengono valori più contenuti di specificità cioè la quota di soggetti non obesi che sono effettivamente non obesi (*veri negativi*), solo i modelli basati sugli alberi e le reti neurali ottengono valori di specificità superiori a 0.80.
Per valutare se ci sono differenze significative tra i modelli, è stata applicato la statistica t-test (corretta con metodo Bonferroni), da cui risulta che:

- per quanto riguarda la sensitività: non ci sono differenze significative tra i diversi modelli;
- il modello logistico, PLS, Lasso, Ridge, LDA e QDA hanno la specificità significativamente più bassa rispetto ai modelli Random Forest, Gradient Boosting, XGBoost, Reti Neurali e Knn.

STEP 2: VALUTAZIONE DELLE PERFORMANCE CLASSIFICATIVE DEI VARI MODELLI

Dopo aver creato i modelli analizziamo i risultati di quest'ultimi sui dati di validation al fine di verificare la loro capacità classificativa, in termini di sensibilità e sensibilità su dati indipendenti.
Sulla base dei risultati ottenuti (confusion matrix), si costruiscono dei grafici che mostrano le curve ROC, che non sono altro che una rappresentazione grafica che si basa sulla combinazione delle seguenti metriche: sensitività e complemento a uno della specificità.
Quest'ultima è la metrica che si preferisce perché valuta la bontà classificativa di ogni modello per tutte le possibili soglie. La curva ROC mostra come variano specificità e sensitività, cioè come varia la percentuale della corretta classificazione degli eventi al variare del tasso di misclassificazione.

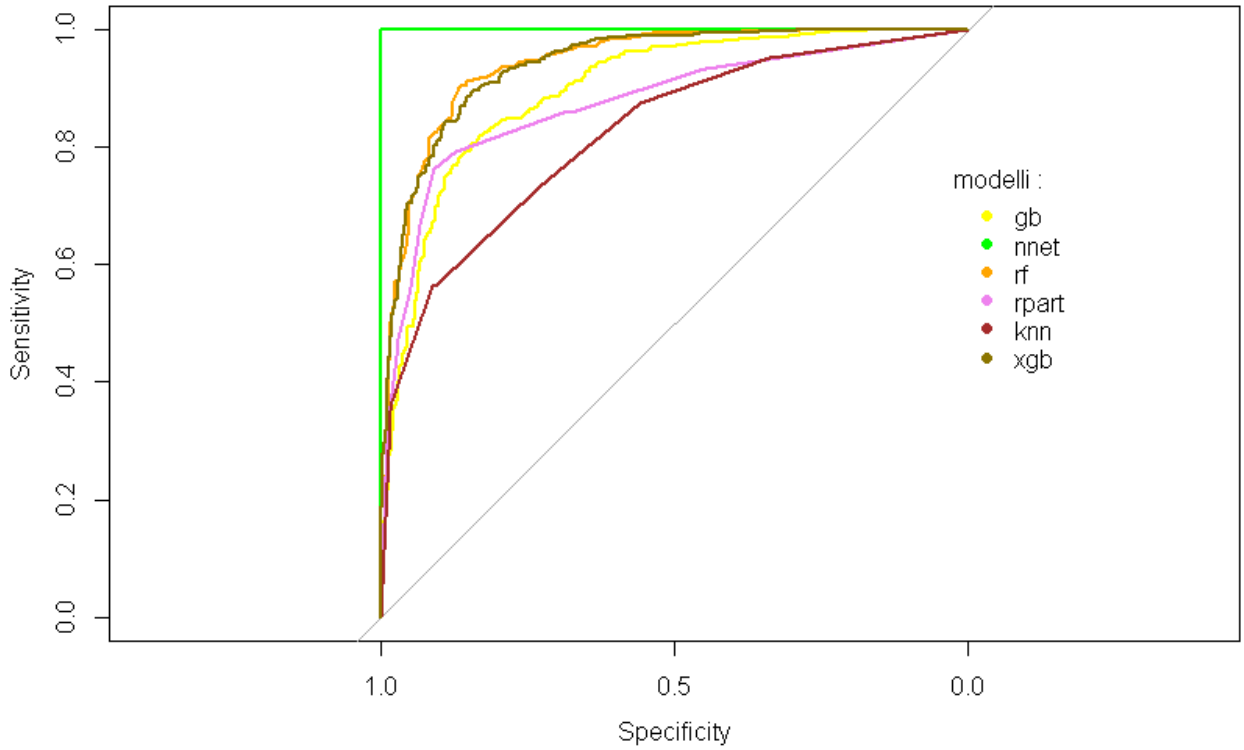

Figura 7 Curve ROC dei modelli performanti

Dalle curve Roc si conferma che i migliori modelli performanti sono quelli basati sulle reti neurali e alberi, infatti se la curva Roc cresce in modo repentino, come nel caso delle Reti Neurali, Random Forest, Gradient Boosting, Xgboost significa che c'è una buona capacità di classificazione e basso errore di misclassificazione anche all'aumentare della soglia (in breve quando la curva ROC sta sull'asse Y modello migliore).
Quando la curva ROC sta sulla diagonale, il modello a cui corrisponde è il peggiore .
Si calcola specificity e sensitivity per ogni soglia, la ROC è la rappresentazione di questi due valori al variare della soglia. Il modello deve lavorare bene con soglie alte, in questo caso è un buon modello.
Ogni curva Roc ha un'area sottostante chiamata AUC (Area Under the Curve) che rappresenta quanto il modello performa bene, più l'area è grande più performa meglio, ossia classifica bene i nostri dati in studio.

Le Curve Gain sono la metrica più interessante per scegliere il modello migliore, è la percentuale di risposte catturate nei primi *x* decili. Più la curva della cumulata cresce, più il modello avrà delle ottime performance classificative. Solitamente si scelgono i primi decili.
In questo caso, le Curve Lift indicano la percentuale di corretti obesi (evento principale dello studio in generale) per ogni porzione di popolazione scelta. Il grafico qui a fianco è relativo ai modelli basati su alberi e reti neurali.

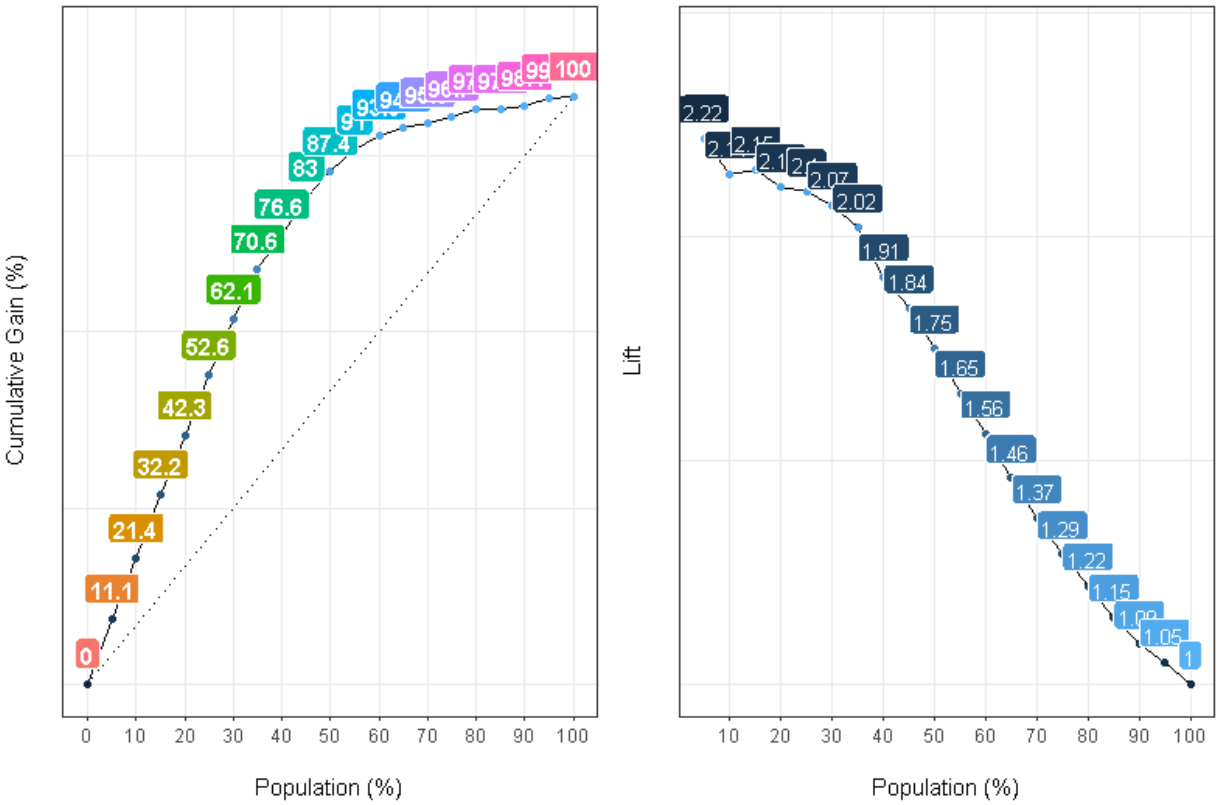

Figura 8 curve lift del modello xgboost

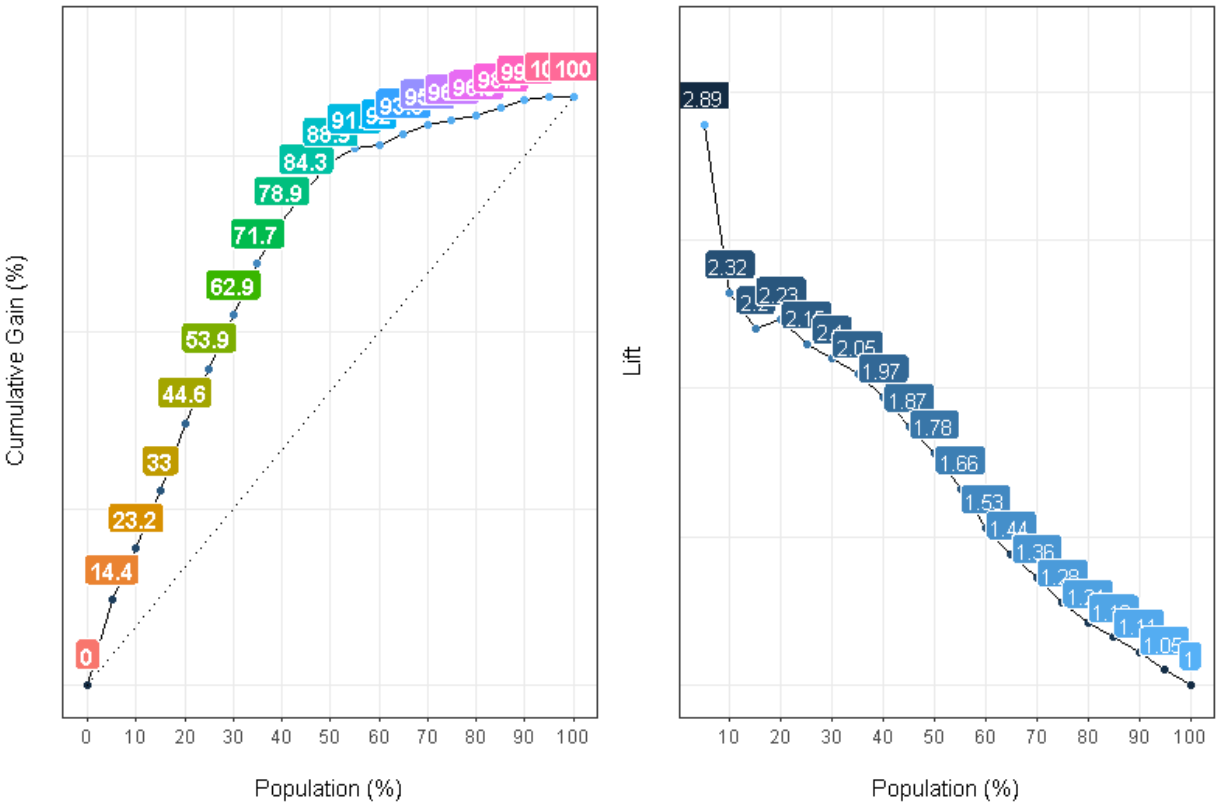

Figura 9 curve lift del modello random forest

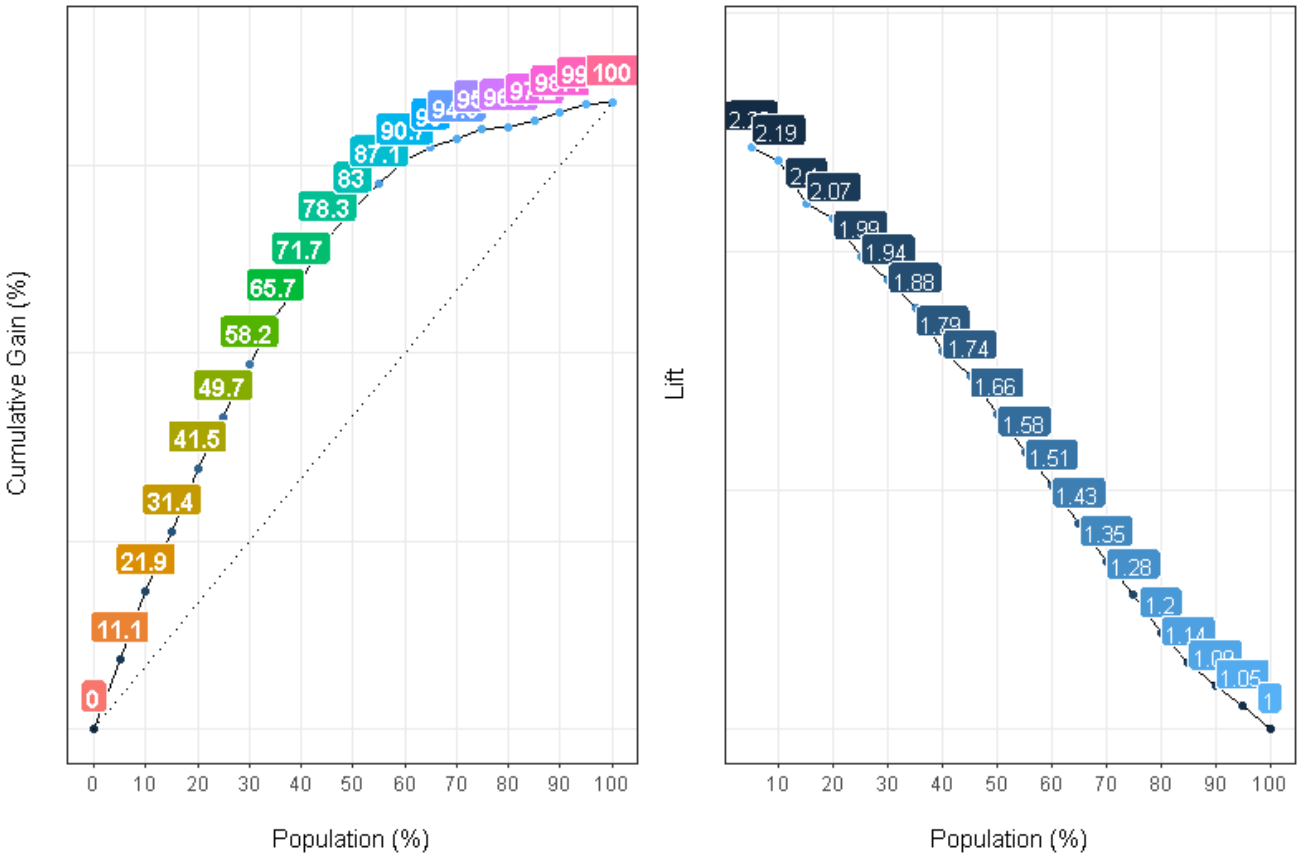

Figura 10 curve lift del modello gradient boosting

Per confrontare le curve lift e decidere il miglior modello, si prende una porzione di popolazione (secondo o terzo decile) e si vede la percentuale di Gain (% risposte catturate veri positivi), il miglior modello sarà quello con il valore di Gain maggiore (maggiore sensibilità).

Modelli	Popolazione	Gain	Score point
Xgboost	20/100	42.53	0.923
Random forest	20/100	43.30	0.850
Gradient boosting	20/100	41.49	0.803

Tab. 2 dei valori delle lift

Nella tabella sono riportati i valori numerici delle lift relativi al 20% della popolazione. Se prendessimo un modello Xgboost riusciremmo ad individuare il 42.53% dei soggetti obesi, la performance salirebbe al 43.30% se utilizzassimo la Random Forest e al 41.49% se utilizzassimo il Gradient Boosting.
In base ai valori delle lift, i modelli migliori sono Xgboost, Gradient Boosting o Random Forest.

Verifichiamo se i modelli Gradient Boosting , Random Forest e Xgboost soffrono di overfitting: calcoliamo le curve Roc del medesimo modello, prima sui dati di train e poi su quelli di validation, per vedere se esse sono uguali o diverse; se esse risultassero sovrapposte non ci sarebbe overfitting.

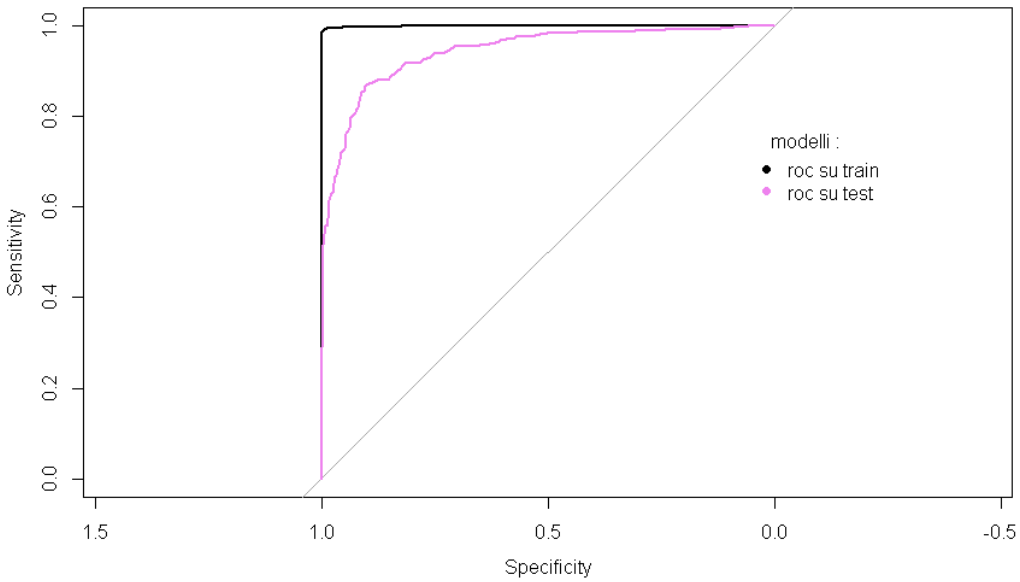

Figura 11 confront curve ROC random forest

Graficamente si può vedere che le curve Roc basate sul modello Random Forest non sono sovrapposte, presentano differenze in termini di distanza e hanno valori molto diversi. Numericamente questo si può notare anche dal valore dell'AUC: la Random Forest su dati di train mostra una AUC = 0.9985, mentre sui dati di test mostra una AUC pari a 0.9286

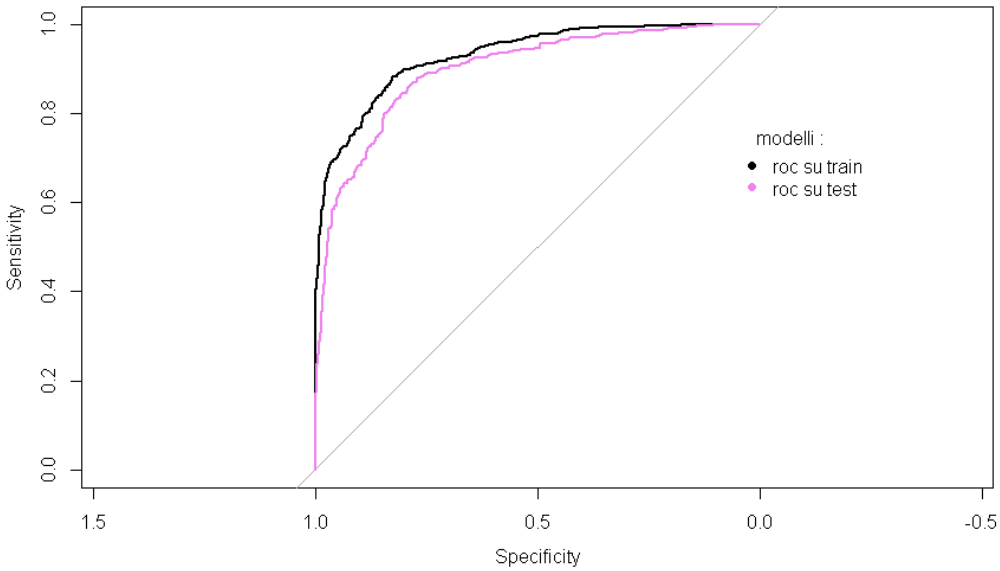

Figura 12 confronto curve ROC Gradient Boosting

Nelle due curve Roc del Gradient Boosting c'è una leggera differenza graficamente, sono tuttavia quasi sovrapposte e non presentano grosse differenze in termini di distanza. Il valore di AUC sui dati di train è pari a 0.9351, mentre sui dati di test è pari a 0.8827.

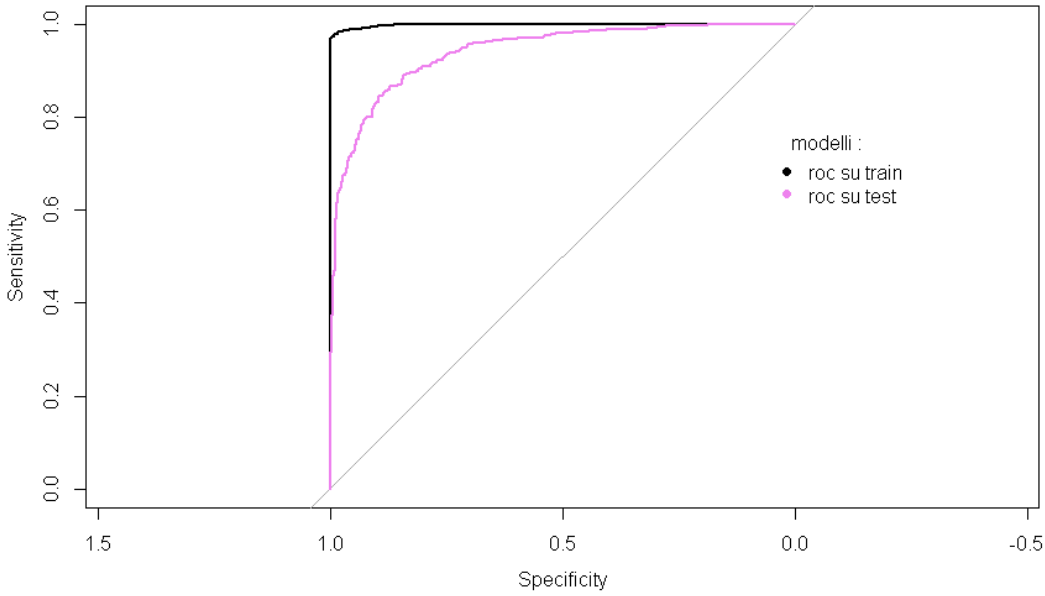

Figura 13 confronto curve ROC Xgboost

Graficamente si può vedere che le curve Roc basate sul modello Xgboost non sono sovrapposte, presentano differenze in termini di distanza e hanno valori diversi.Il valore dell'AUC sui dati di train è pari a 0.9962, mentre sui dati di test è pari a 0.9308.
Si decide che il modello migliore è il Gradient Boosting poiché non presenta il fenomeno dell'overfitting, quindi garantisce una maggiore generalizzabilità e replicabilità, non garantite dai modelli Random Forest e Xgboost. Inoltre il Gradient Boosting ha tempi computazionali ristretti e garantisce risparmio di tempo e risorse

STEP 3: VALUTARE LA PERFORMANCE CLASSIFICATIVA DEL MODELLO SCELTO (GRADIENT BOOSTING)

Ora valutiamo la soglia ottimale per il modello *Gradient Boosting*, consigliata dalle lift utilizzando il criterio statistico che confronta diverse metriche su diverse soglie.

La scelta delle metriche dipende dalle specifiche esigenze del problema e dalla sua natura. Come nel nostro esempio in cui è più importante evitare diagnosi errate di obesità (ridurre i *falsi positivi*), la precisione e la specificità possono essere metriche cruciali. Tuttavia, in situazioni in cui è fondamentale identificare tutti i casi di obesità (ridurre i *falsi negativi*), la sensibilità può essere la metrica principale. In genere, è importante considerare più metriche insieme per valutare congiuntamente le prestazioni del modello.

Si sceglie la soglia in modo che il modello soddisfi le nostre necessità, nel nostro caso si deve scegliere la soglia ottimale a cui corrisponde la migliore sensitivity ma al contempo vogliamo avere anche una buona specificity.
Per il modello Gradient Boosting, la soglia che ottimizza quanto detto sopra è 0.375, come mostrato dal grafico seguente.

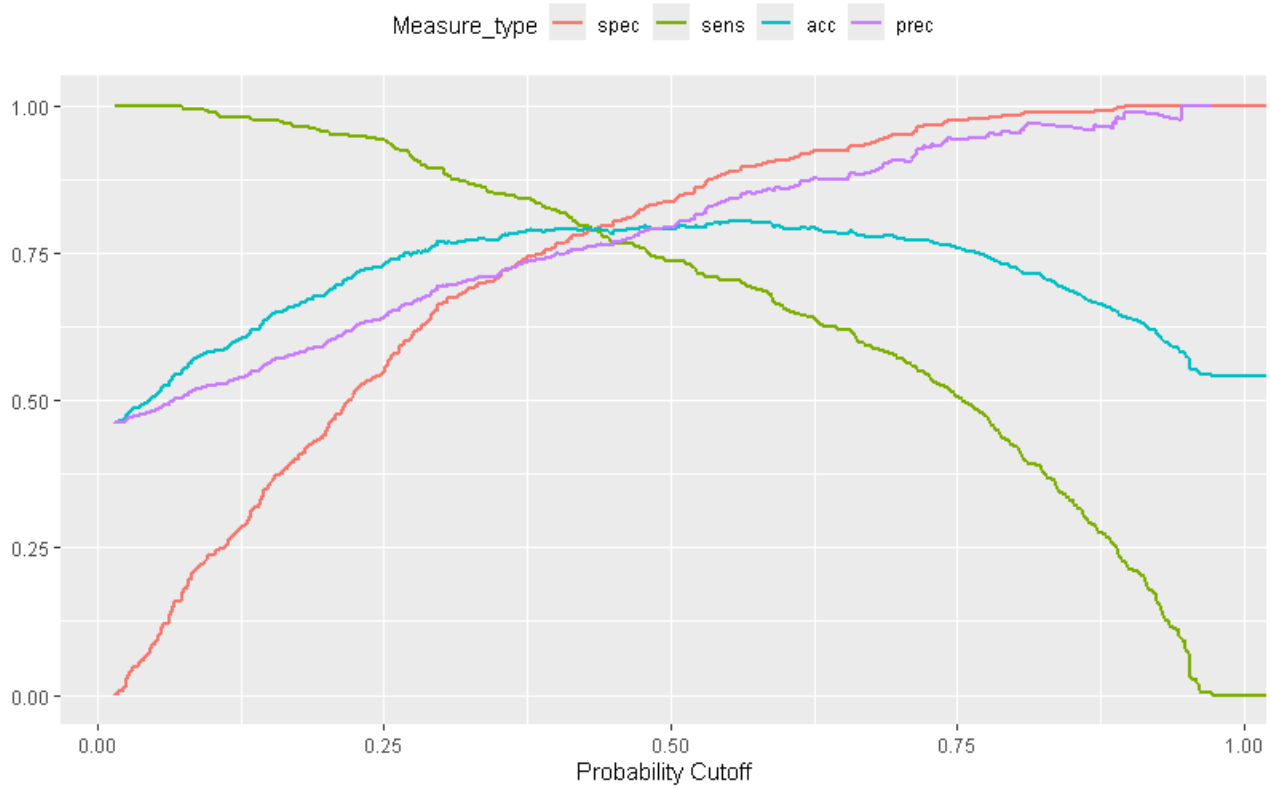

Figura 14 grafico valori metriche in relazione alla soglia(thresold)

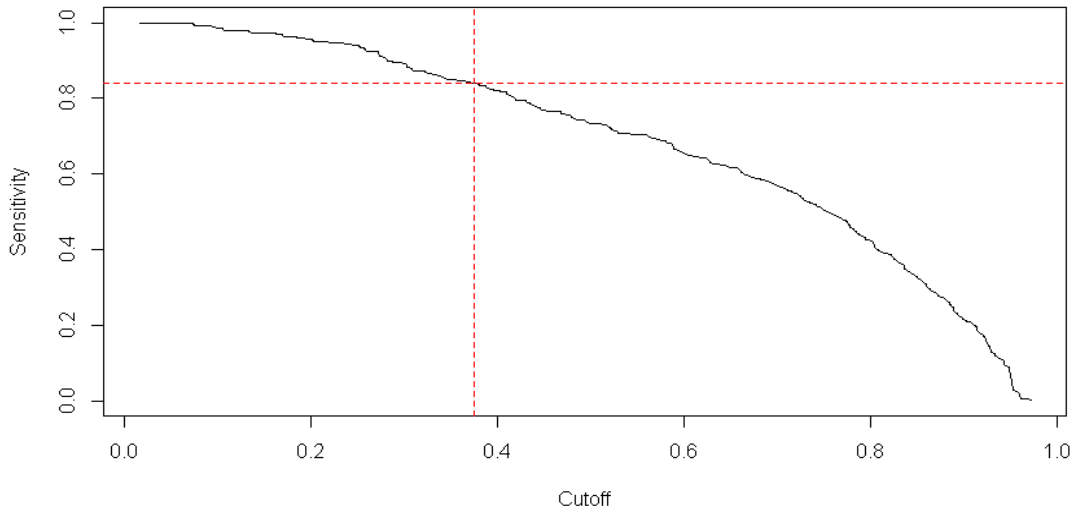

Figura 15 .Grafico sensitivity in relazione alla soglia (thresold=0.375)

STEP 4: SCORE NUOVI DATI

Nel quarto step applichiamo il miglior modello a nuovi dati per valutare le sue performance. Il dataset di score che useremo contiene circa 100 osservazioni prese randomicamente. Replicando il modello su nuovi dati di score 22 soggetti su 100 vengono missclassificati, rispetto al nostro obiettivo principale ossia di massimizzare i veri positivi, vengono classificati correttamente 32 soggetti con una sensibilità pari a 0.86 e le persone identificate come non obese, i veri negativi, sono 46 con una specificità del 0.73

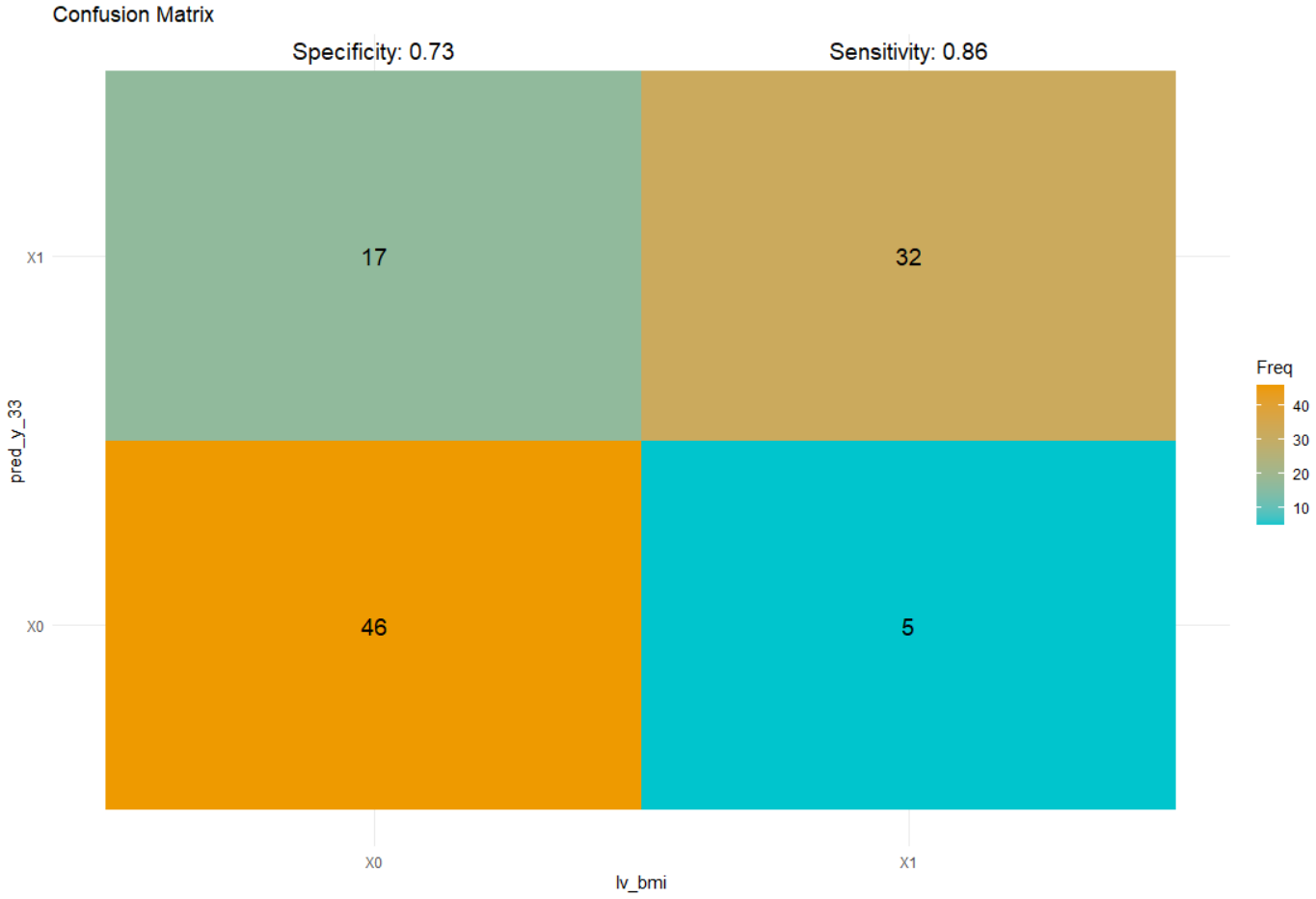

Figura 16 confusion matrix sui dati di test con nuova soglia scelta nello step 3