Federico Luo 882912

Nicoleta Mihalachi 875130

**LOAN DEFAULTS**

*Dati anonimizzati di Loan Default per prevedere se un prestito sarà insolvente o meno*

MATERIALE:

* Il dataset utilizzato, reperito da Kaggle, contiene molte osservazioni : [LOAN DEFAULTS](https://www.kaggle.com/datasets/joebeachcapital/loan-default)
* Descrizione variabili : **`id`**: Un identificatore unico per ciascun prestito, **`member\_id`**: Un identificatore unico per ciascun membro,`**loan\_amnt**`: L'ammontare del prestito richiesto,`**funded\_amnt**`:L'ammontare effettivamente finanziato,`**funded\_amnt\_inv**`: L'ammontare effettivamente investito nel prestito,`**term**`: La durata del prestito in mesi (ad esempio, 36 o 60 mesi),`**int\_rate**`: Il tasso di interesse del prestito,`**installment**`: L'importo delle rate mensili del prestito,`**emp\_length**`: L'anzianità lavorativa dell'individuo (ad esempio, "5 years"),`**home\_ownership**`: Il tipo di proprietà dell'individuo (ad esempio, "own", "mortgage", "rent", ecc.),`**annual\_inc**`: Il reddito annuale dell'individuo,`**verification\_status**`: Lo stato di verifica del reddito (ad esempio, "verified" o "not verified"),`**issue\_d**`: La data di emissione del prestito,`**loan\_status**`: Lo stato attuale del prestito (ad esempio, "Fully Paid", "Charged Off", ecc.),`**purpose**`: Lo scopo del prestito (ad esempio, "debt\_consolidation", "credit\_card", ecc.),`**zip\_code**`: Il codice postale dell'individuo, `**addr\_state**`: Lo stato in cui risiede l'individuo, `**dti**`: Il rapporto debito/reddito dell'individuo,`**delinq\_2yrs**`: Il numero di pagamenti in ritardo negli ultimi 2 anni,`**earliest\_cr\_line**`: La data di apertura del primo conto di credito,**`inq\_last\_6mths`**: Il numero di richieste di informazioni del credito negli ultimi 6 mesi,`**mths\_since\_last\_delinq`**: Il numero di mesi dall'ultimo pagamento in ritardo, **`open\_acc**`: Il numero di conti aperti, **`pub\_rec**`: Il numero di record pubblici, `**revol\_bal**`: Il saldo totale delle carte di credito revolving, `**revol\_util**`: L'utilizzo del credito sulle carte revolving, `**total\_acc**`: Il numero totale di conti,`**total\_pymnt**`: L'importo totale pagato,`**total\_pymnt\_inv**`: L'importo totale pagato dagli investitori,`**total\_rec\_prncp**`: L'importo totale del capitale ricevuto,`**total\_rec\_int**`: L'importo totale degli interessi ricevuti, `**last\_pymnt\_d**`: La data dell'ultimo pagamento,`**last\_pymnt\_amnt**`: L'importo dell'ultimo pagamento, `**next\_pymnt\_d**`: La data del prossimo pagamento,`**last\_credit\_pull\_d**`: La data dell'ultima richiesta di informazioni del credito, `**repay\_fail**`: Una variabile che rappresenta il fallimento del rimborso (una variabile binaria che indica se un prestito è stato rimborsato o meno).

IDEA DELLO STUDIO E DESCRIZIONE DEL DATASET:

Il nostro obiettivo è sviluppare un modello di machine learning in grado di classificare efficacemente la variabile **"repay-fail",** consentendoci di prendere decisioni informate sul concedere o meno un prestito. L'obiettivo è quindi stimare accuratamente i clienti che potrebbero avere difficoltà nel restituire il finanziamento.

Il dataset che utilizzeremo per questo studio contiene una serie di variabili che riguardano i clienti e i prestiti concessi, come ad esempio l'ammontare del prestito, il reddito del cliente, la sua storia creditizia, la durata del prestito, e così via. Ogni record nel dataset rappresenta un prestito precedentemente concesso, con informazioni associate sul suo successo o fallimento nel rimborso. Utilizzeremo queste informazioni per addestrare il nostro modello di machine learning, in modo da identificare i pattern che portano al successo o al fallimento nel rimborso del prestito.

TARGET

**repay\_fail**: Una variabile che rappresenta il fallimento del rimborso (una variabile binaria che indica se un prestito è stato rimborsato o meno).



Questo dataset sembra rappresentativo della popolazione , e quindi non consideriamo il fatto di bilanciare.

METRICA PIÙ IMPORTANTE DELL’ANALISI:

**MAX SENSITIVITY**

Massimizzare la sensibilità è particolarmente importante quando vuoi ridurre al minimo i falsi negativi (ossia, non prevedere correttamente un cliente che andrà in default). Riducendo i falsi negativi, si cerca di garantire che la banca identifichi correttamente la maggior parte degli individui che effettivamente andranno in default, consentendo di adottare misure preventive o gestire in modo più efficace i rischi.

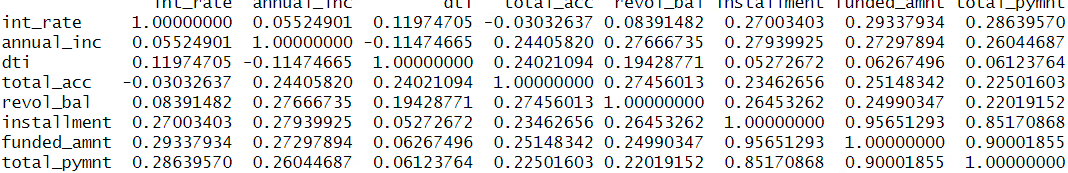
PRE-PROCESSING GLOBALE: IMPUTAZIONE E COLLINEARITÀ e ZV

**SI ESEGUE IL PRE-PROCESSING :**

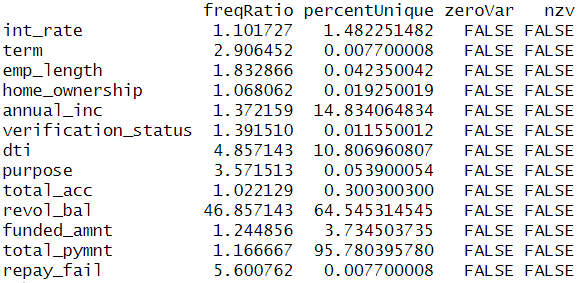
SI CREANO DIVERSI DATASET CON DIVERSI PREPROCESSING

-1 DATASET CON CONTROLLO DI VALORI MANCANTI (ELIMINAZIONE DELLE COLONNE CHE CONTENGONO PIU’ del 20 % MISSING VALUE)

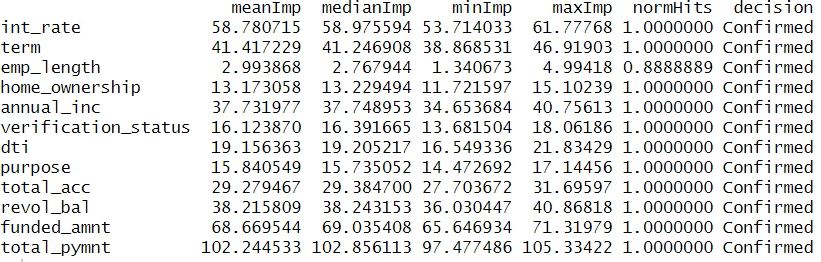
-2 DATASET CON COLLINEARITA (ELIMINATO INSTALLMENT)



-3 PREDITTORI CON VARIANZA ZERO E NEAR ZERO VARIANCE



-Model selection con Boruta(usato solo per algoritmi che ne hano bisogno)



PARTIZIONE DEI DATI (STRATIFICAZIONE PER TARGET)

* Dataset di score : 10% dei dati globali vengono messi apparte



* HOLDOUT method: Il restante dataset viene suddiviso nel seguente modo:

Data di Training 75% ( e un dataset di Training con model selection)

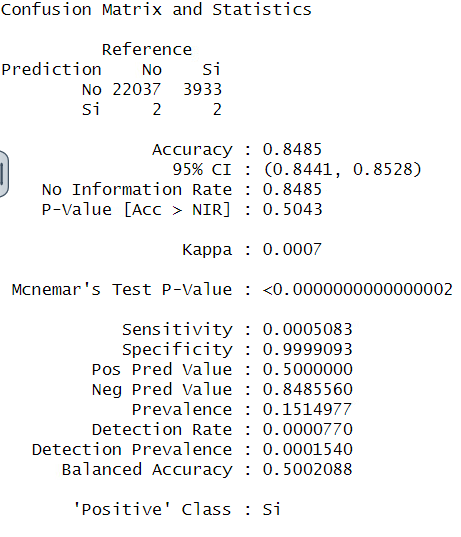
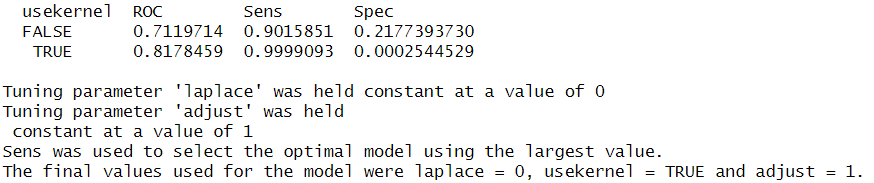


Data di Validation 25%

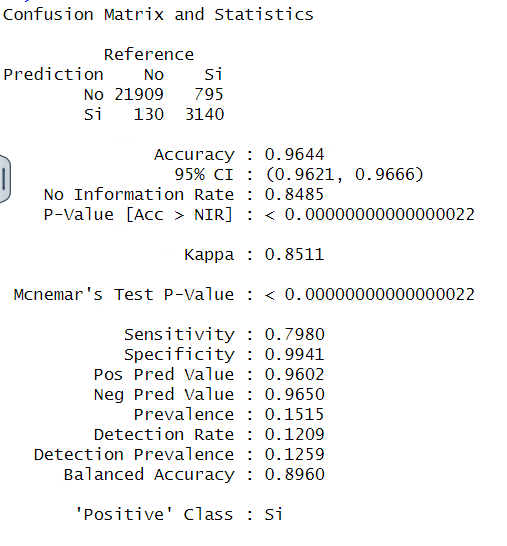


STEP 1 – FITTING DI K MODELLI CLASSIFICATIVI

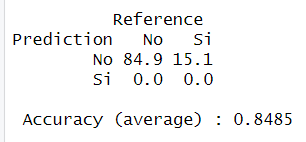
NAIVE BAYES

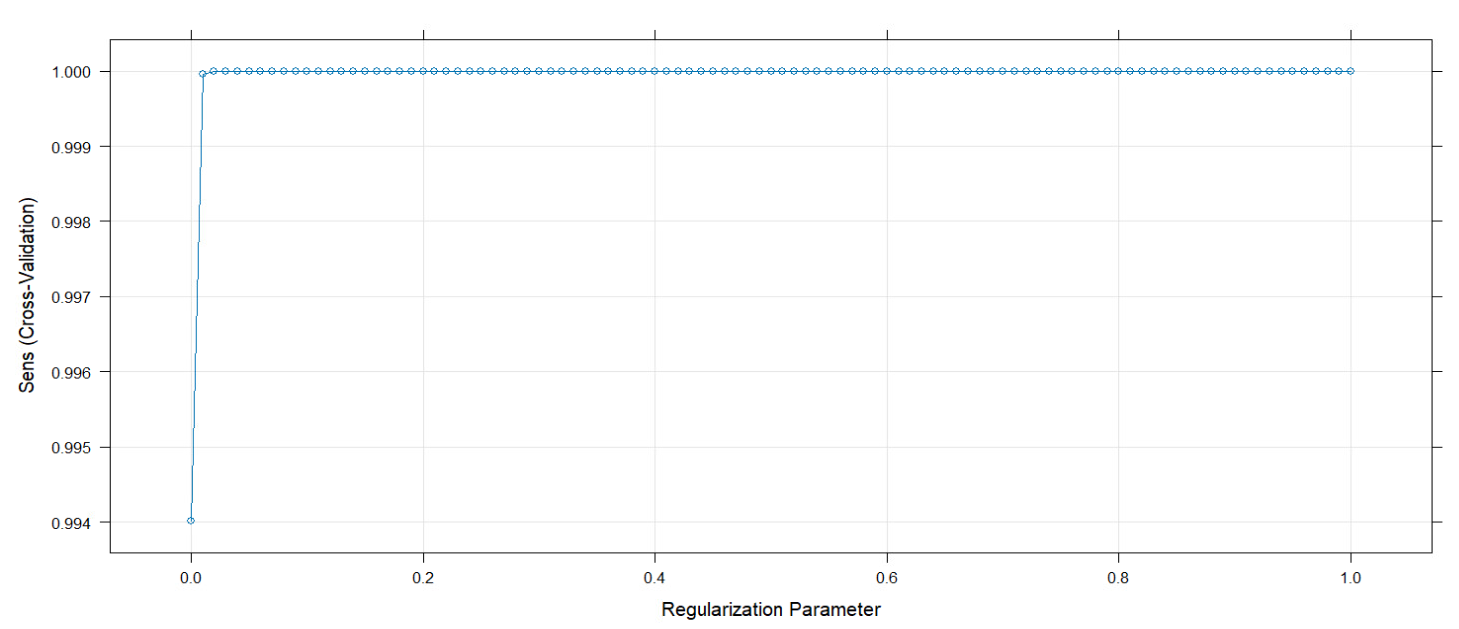
****

GLM

****

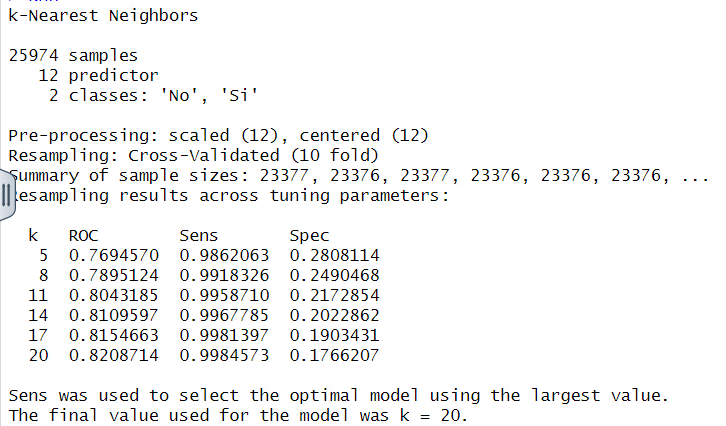
**LASSO**

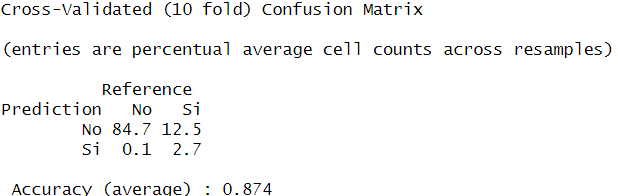


****

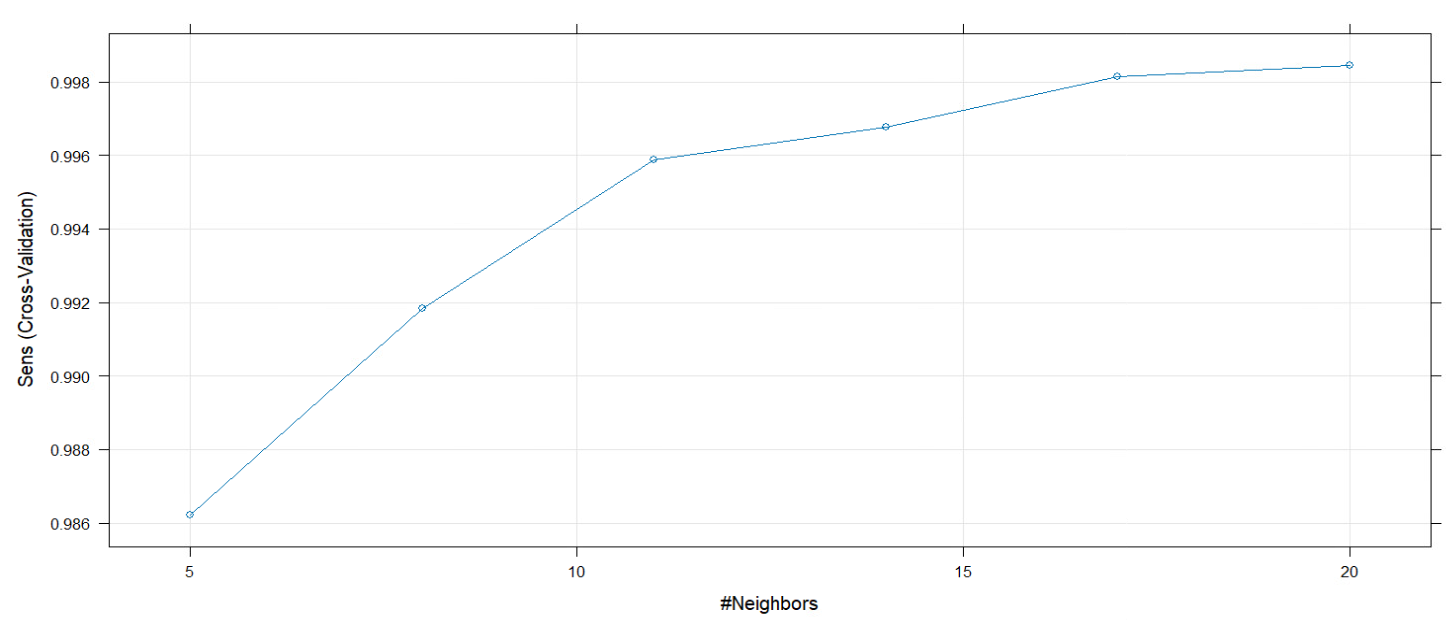
**KNN**

Pre-Process Specifico : model selection, input transformation (z, or 0-1), variabili fattoriali rese continue.

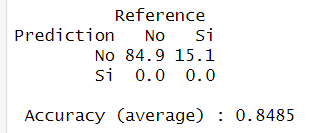


****

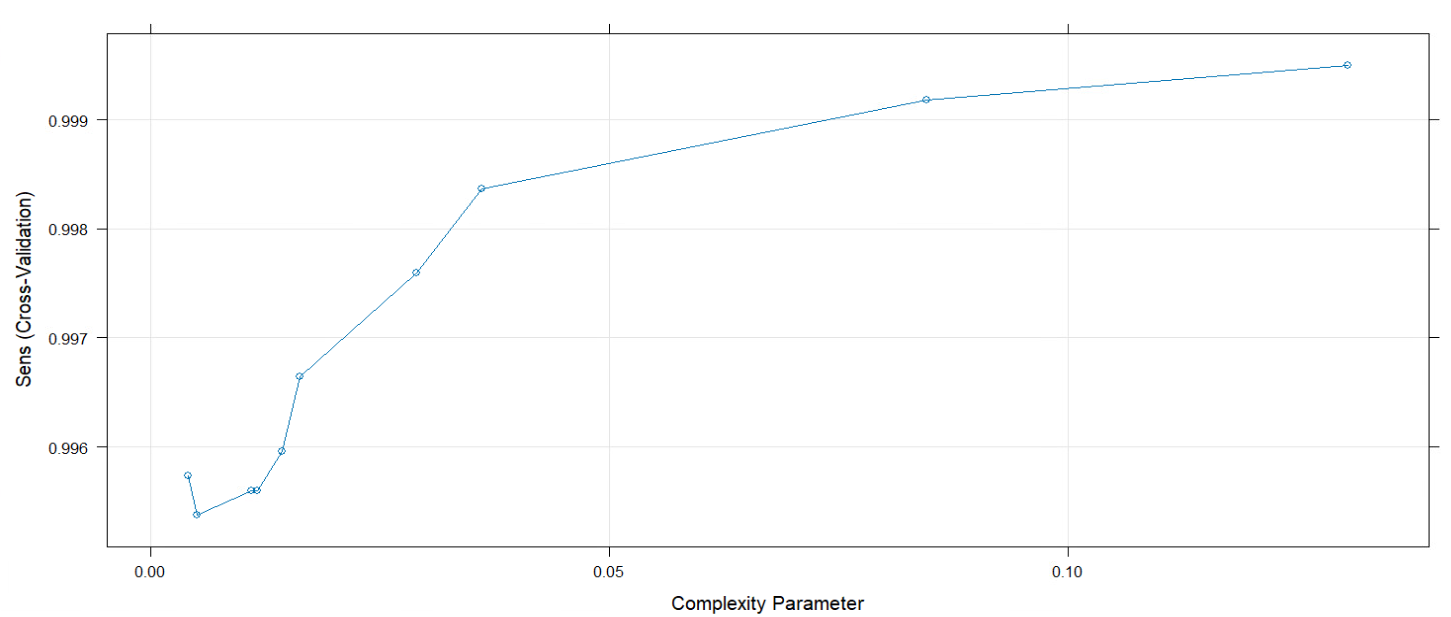
****

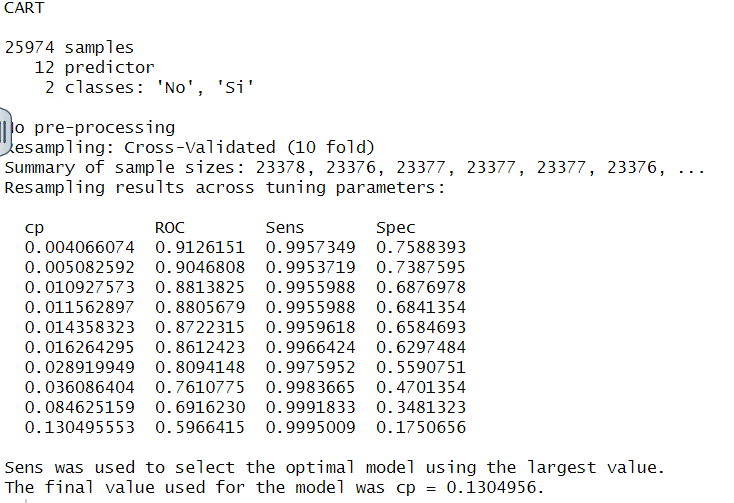
****

**PLS**

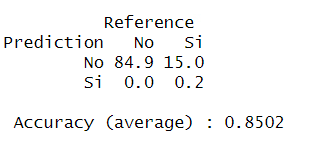
****

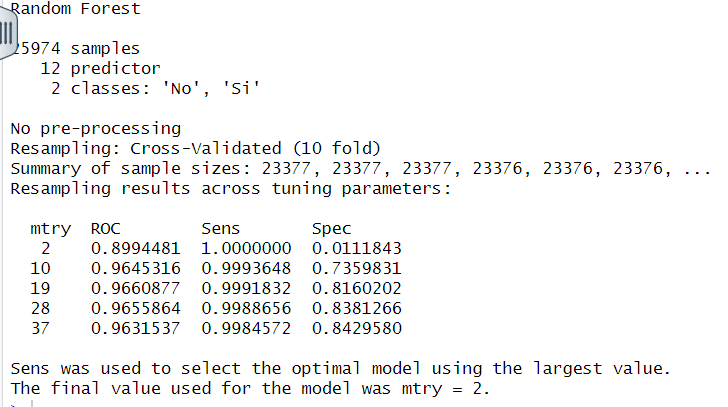
**TREE**

****

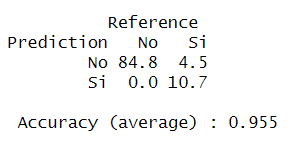
****

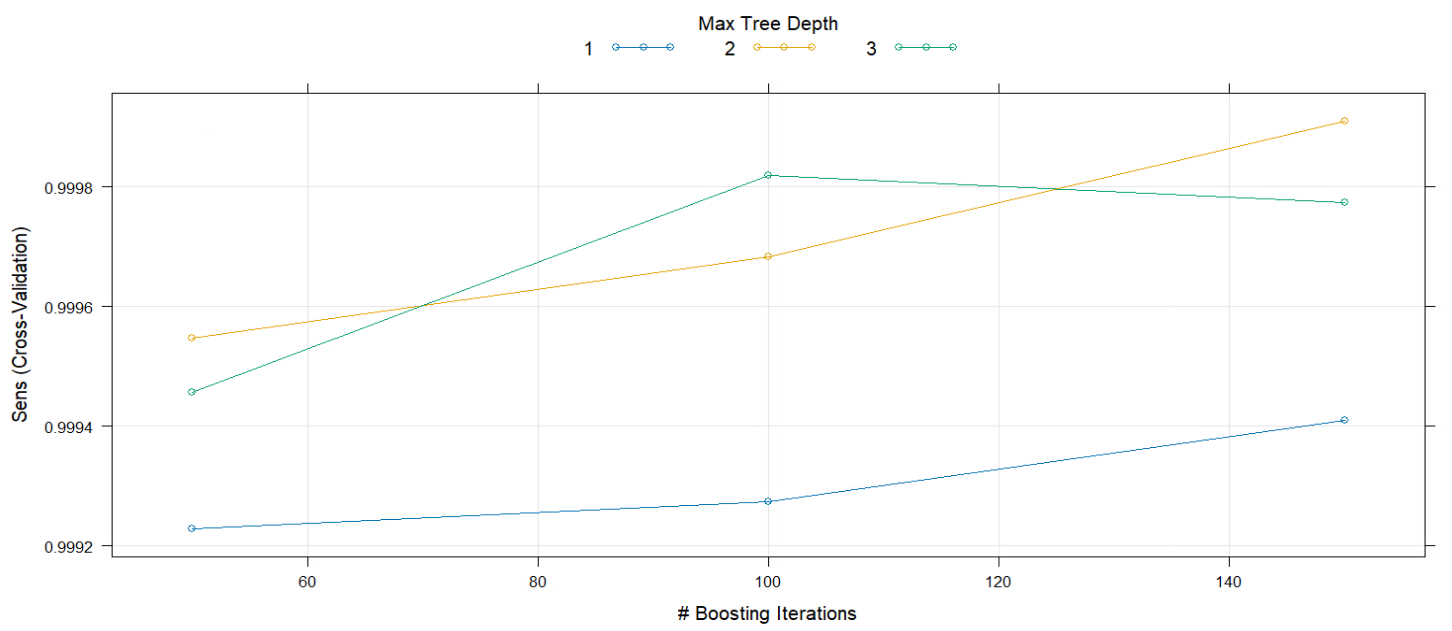
**RANDOM FOREST**

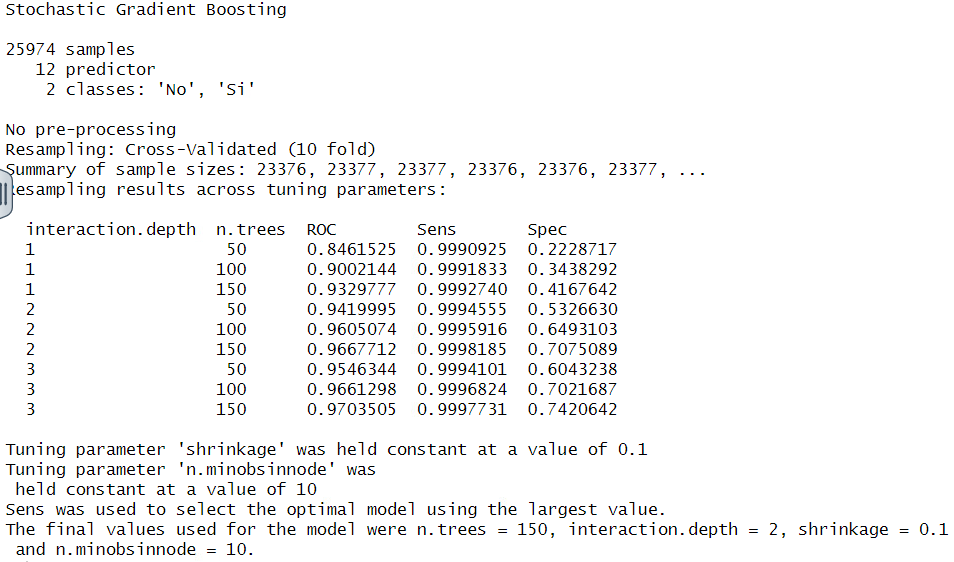
****

****

**GRADIENT BOOSTING**

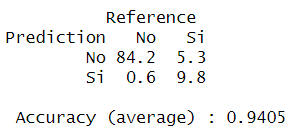
****

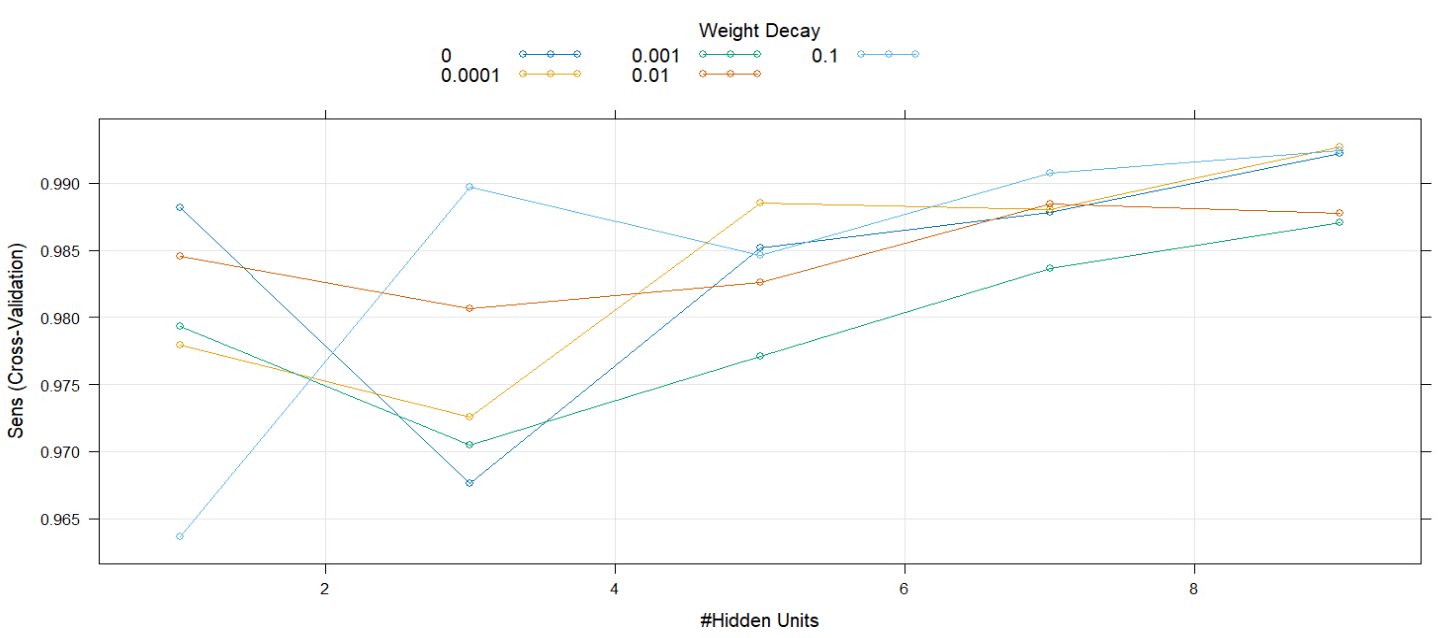
****

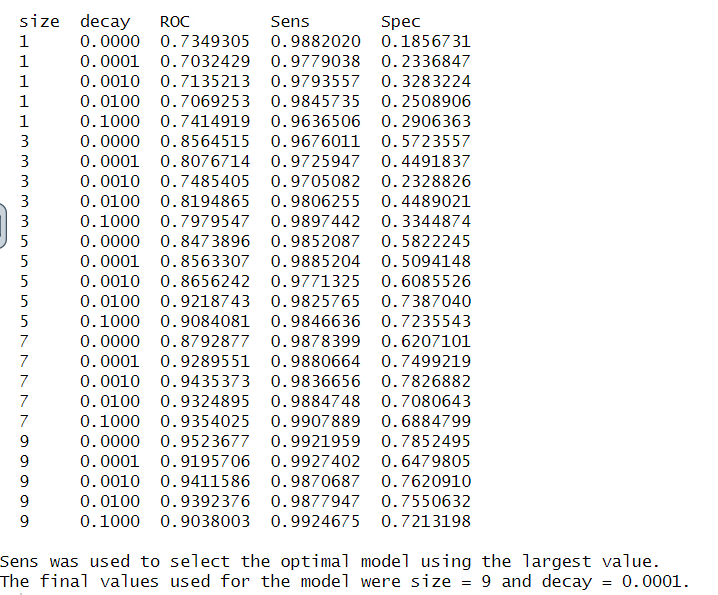
****

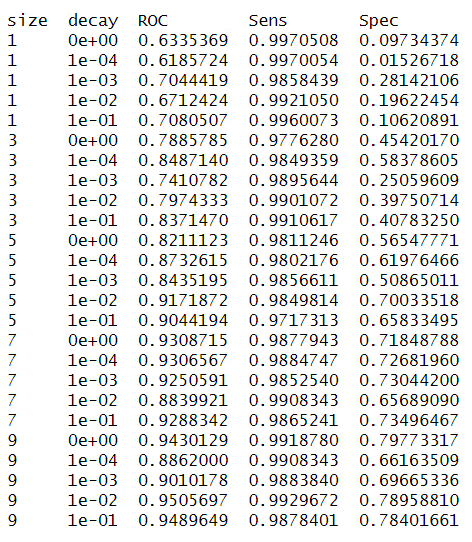
**NN**

**Preprocess: variabili standardizzate, non collineari e near zero variance**

****

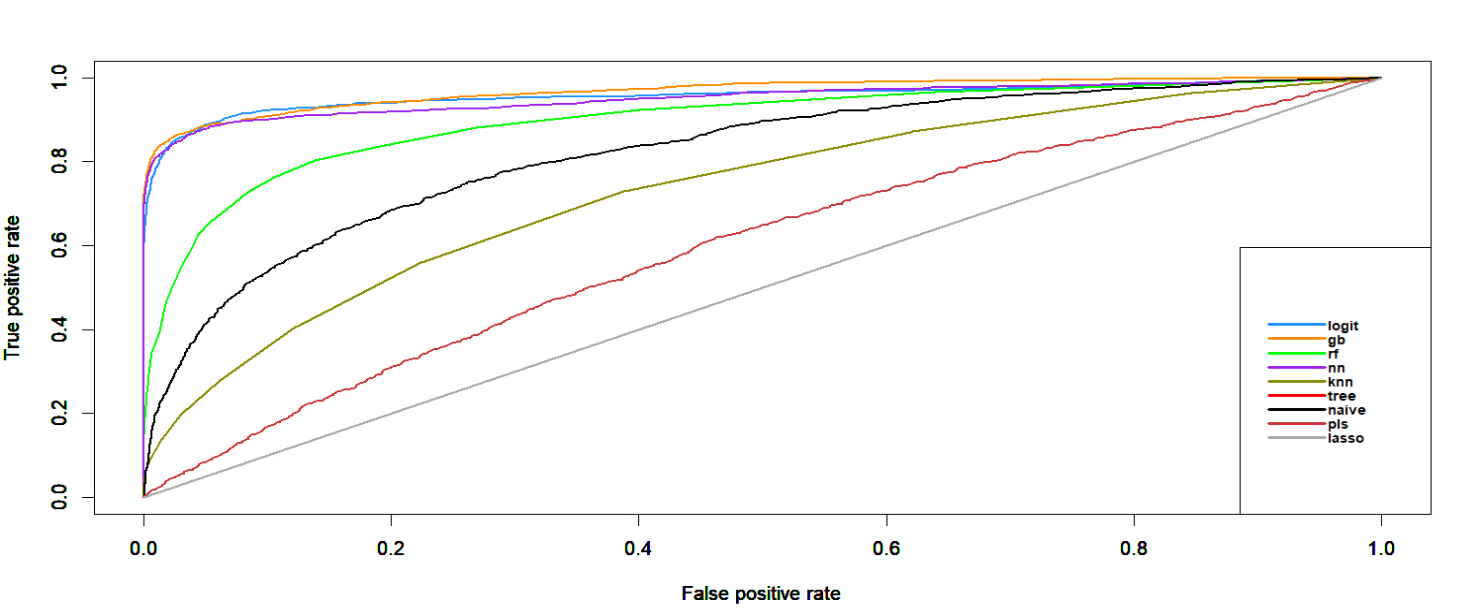
****

****

****

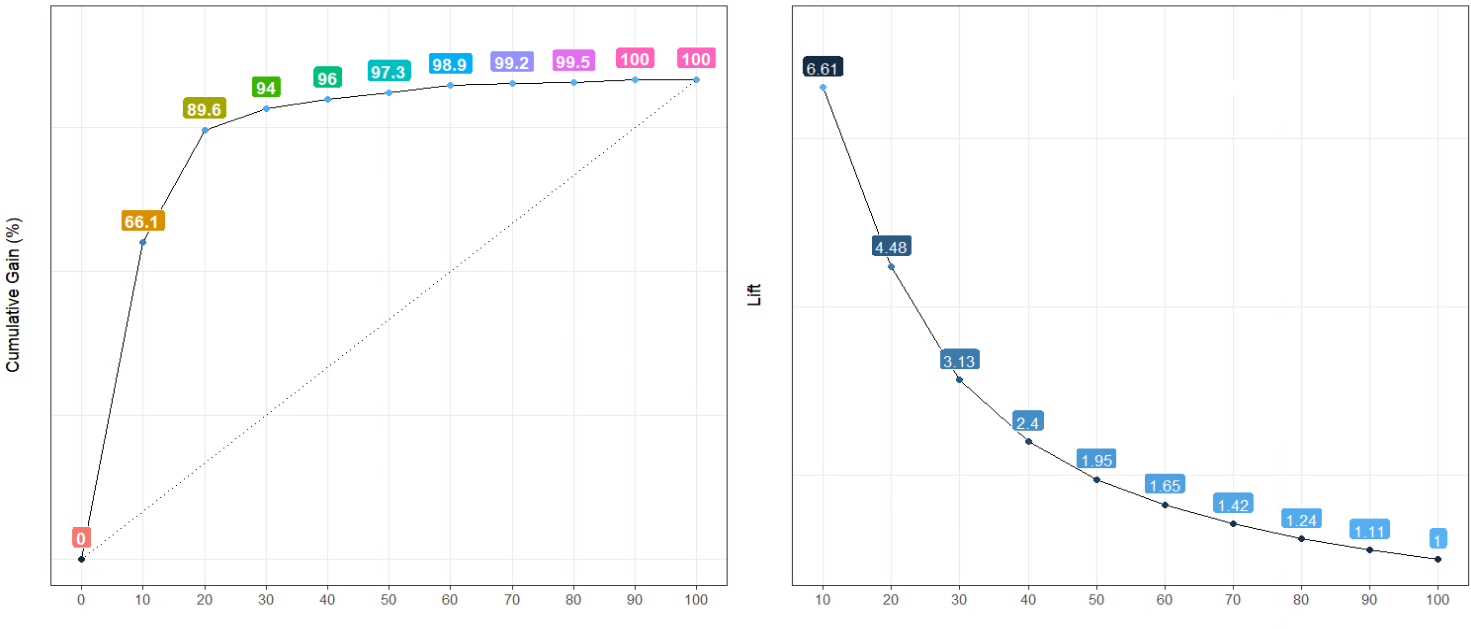
ASSESSMENT

**ROC**

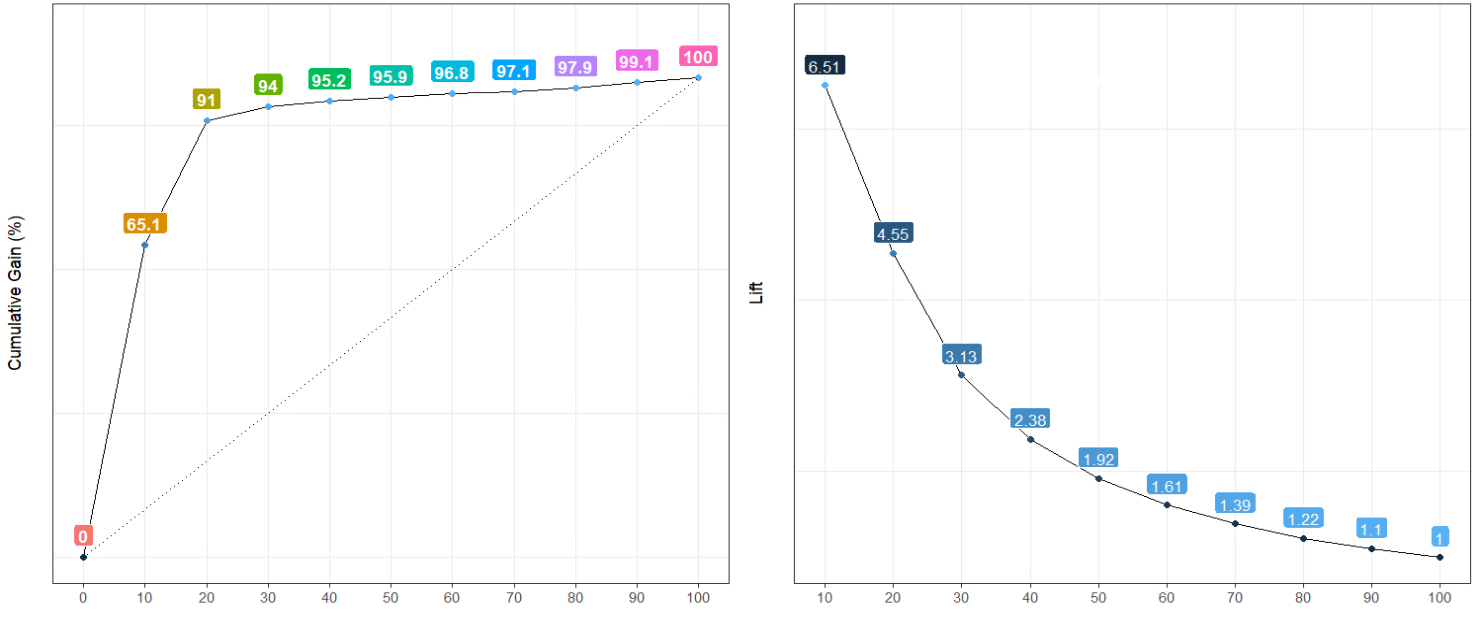
****

**GB,NN;LOGIT**  risultano quelli migliori , ora attraverso le lift verifichiamo meglio

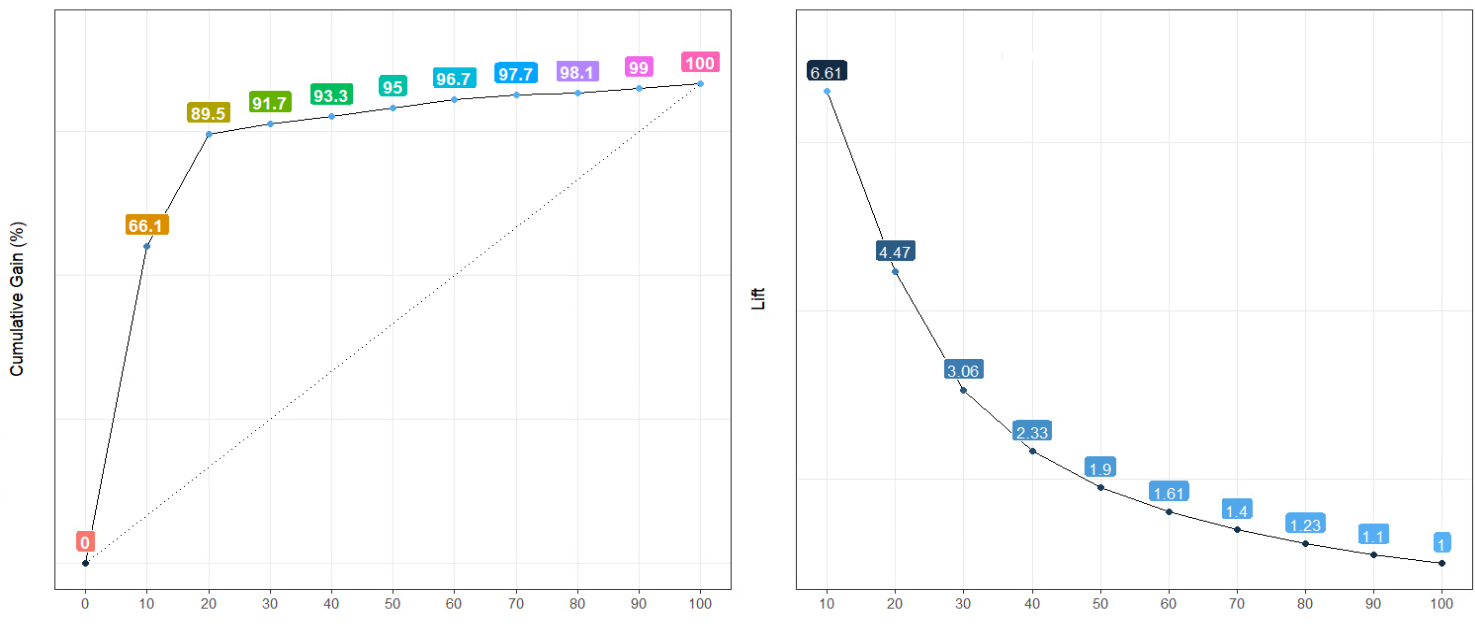
**LIFT GRADIENT BOOSTING**

****

**LIFT GLM**

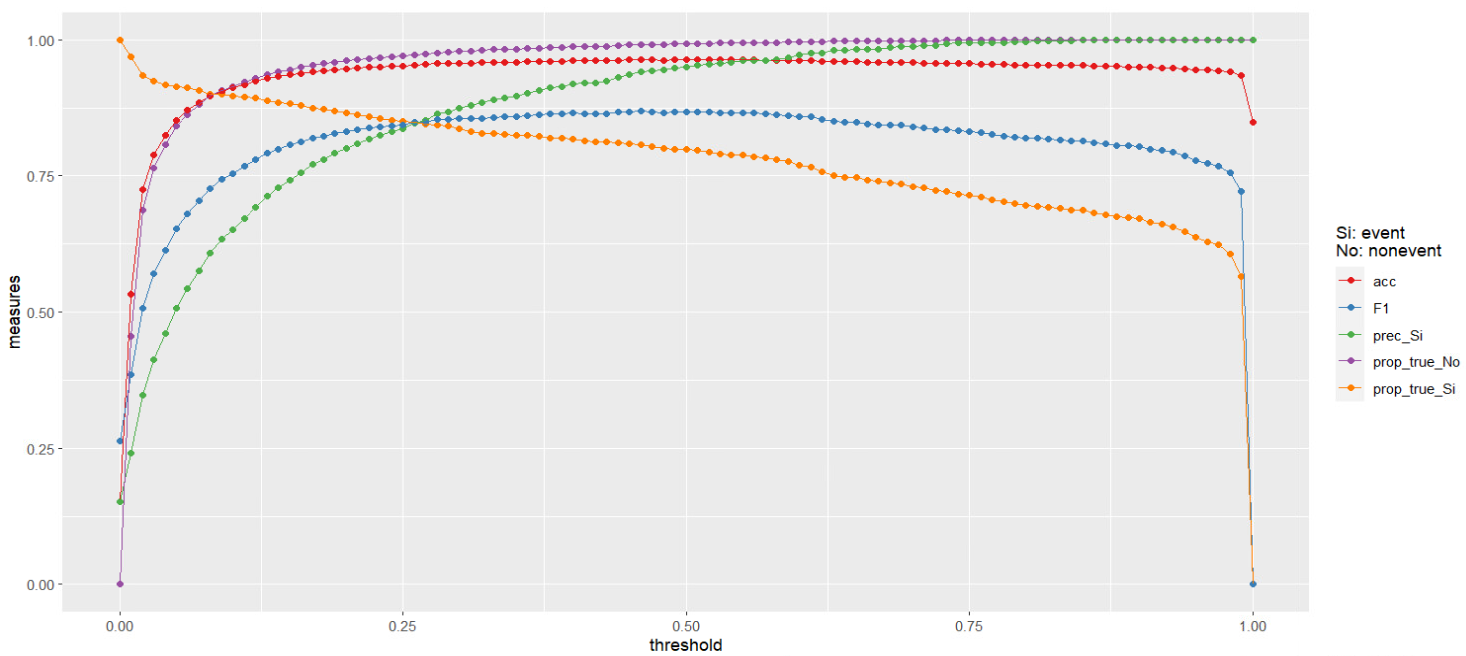
****

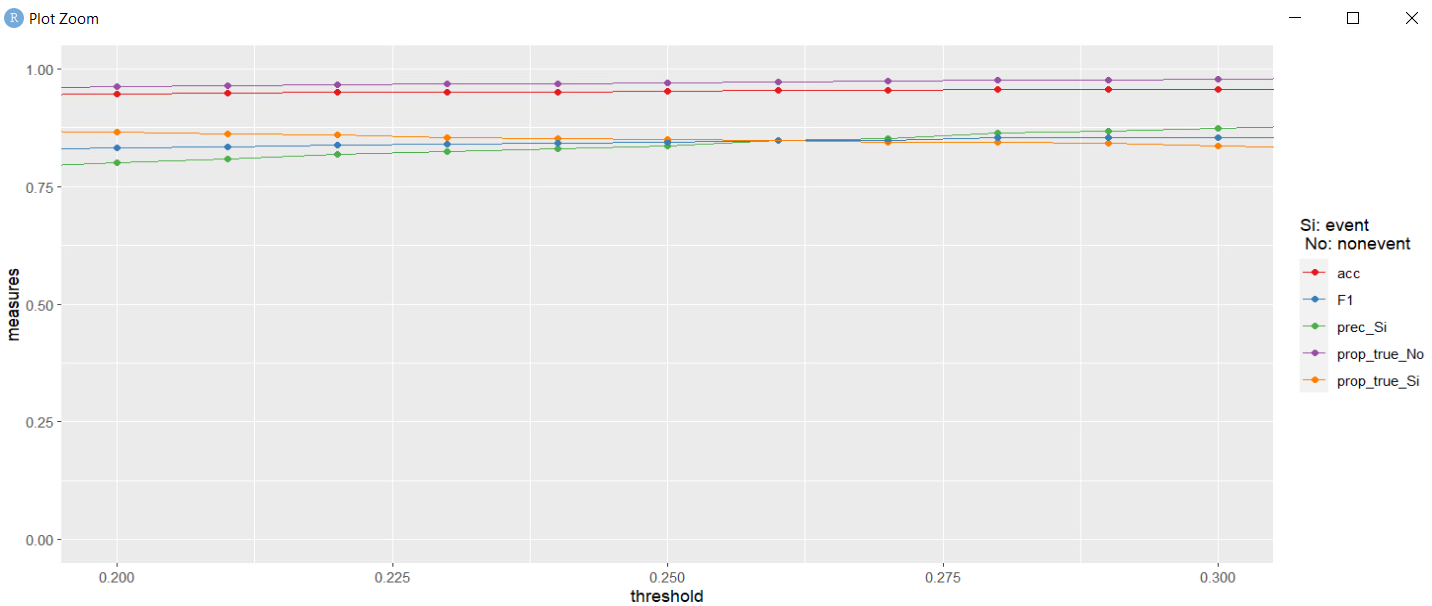
**LIFT NN**

****

**MODELLO MIGLIORE GRADIENT BOOSTING!**

SCELTA DELLA SOGLIA



****

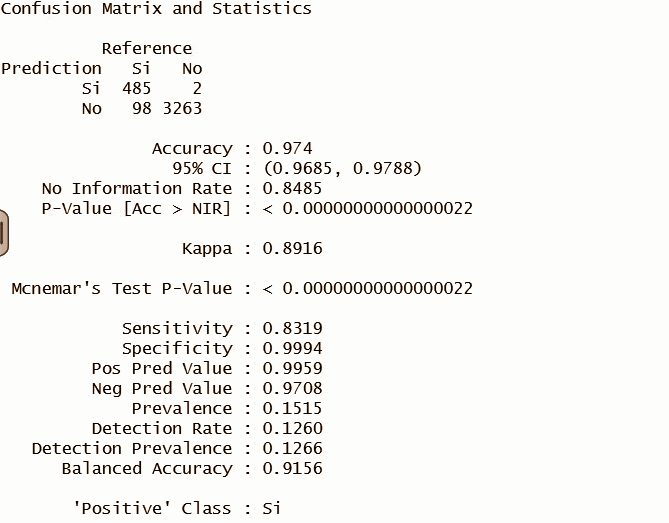
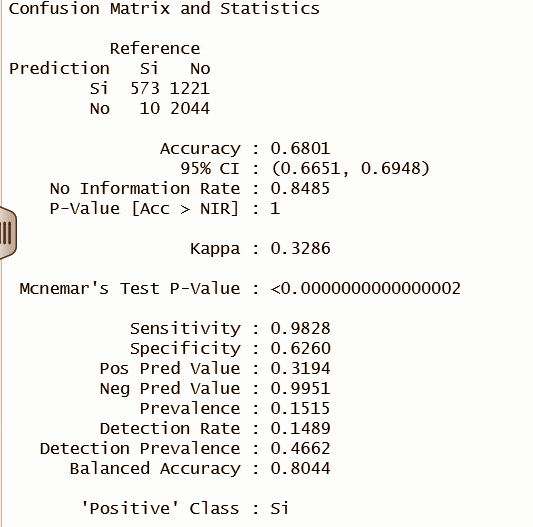
**1 CASO ->SOGLIA 0,05 !**

**2 CASO ->Scegliamo la soglia 0.27!**

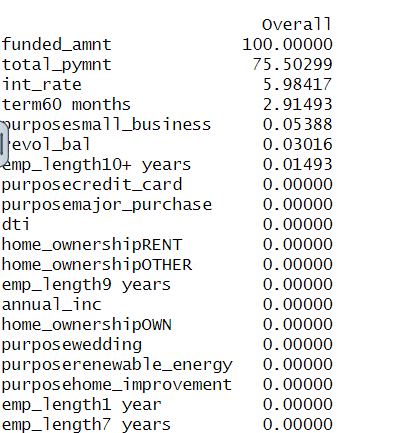
**METRICE USATA PER LA SOGLIA :** TRADE OFF FRA PRECISION E SENSITIVITY

SCORE:

**GRADIENT BOOSTING:**

**soglia 0.05 soglia 0.27**

**soglia 0.27**



**variable importance of GB:**

**CONCLUSIONE…. /DIBATTITO**