Progetto Machine Learning in Economics and Business.

Realizzato da :

- Mariangela Tafuri
- Vincenzo Picarelli
- Simari Paolo

Descrizione Dataset.

Il set di dati contiene informazioni raccolte dal Servizio di censimento degli Stati Uniti in merito agli alloggi nell'area di Boston Mass.

Obiettivo principale dell’analisi è quello di prevedere il valore mediano del prezzo delle case (medv) sulla base di una serie di caratteristiche descritte dalle variabili di seguito riportate.

- **Attività di selezione del modello migliore**: in base all'insieme delle variabili, l'attività consiste nel identificare la funzione migliore, in termini di accuratezza, che prevede il prezzo delle abitazioni.

Features

CRIM - tasso di criminalità pro capite per città

ZN - proporzione di terreno residenziale suddiviso in zone per lotti superiori a 25.000 piedi quadrati.

INDUS - Percentuale di acri di attività non al dettaglio per città

CHAS - Charles River variabile fittizia (= 1 se il tratto delimita il fiume; 0 altrimenti)

NOX - Concentrazione di ossidi di azoto (parti per 10 milioni)

RM - numero medio di stanze per abitazione

AGE - Proporzione delle unità occupate dai proprietari costruite prima del 1940

DIS - Distanze ponderate per cinque centri per l'impiego di Boston

RAD - Indice di accessibilità alle autostrade

TASSA- aliquota dell'imposta sulla proprietà a valore pieno per $ 10.000

PTRATIO - rapporto alunni-insegnanti per città

B - 1000(Bk - 0,63)^2 dove Bk è la proporzione di neri per città

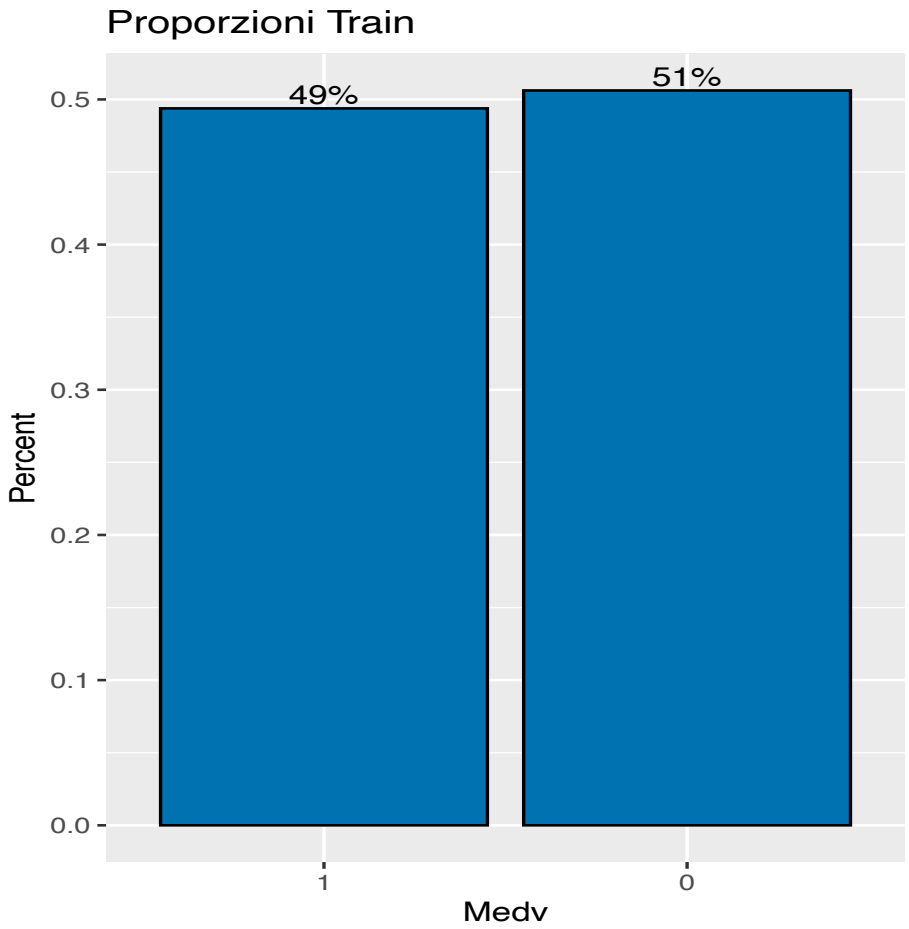
LSTAT - % di popolazione sotto la soglia di povertà

MEDV - Valore mediano delle case occupate dai proprietari in $ 1000

Set.

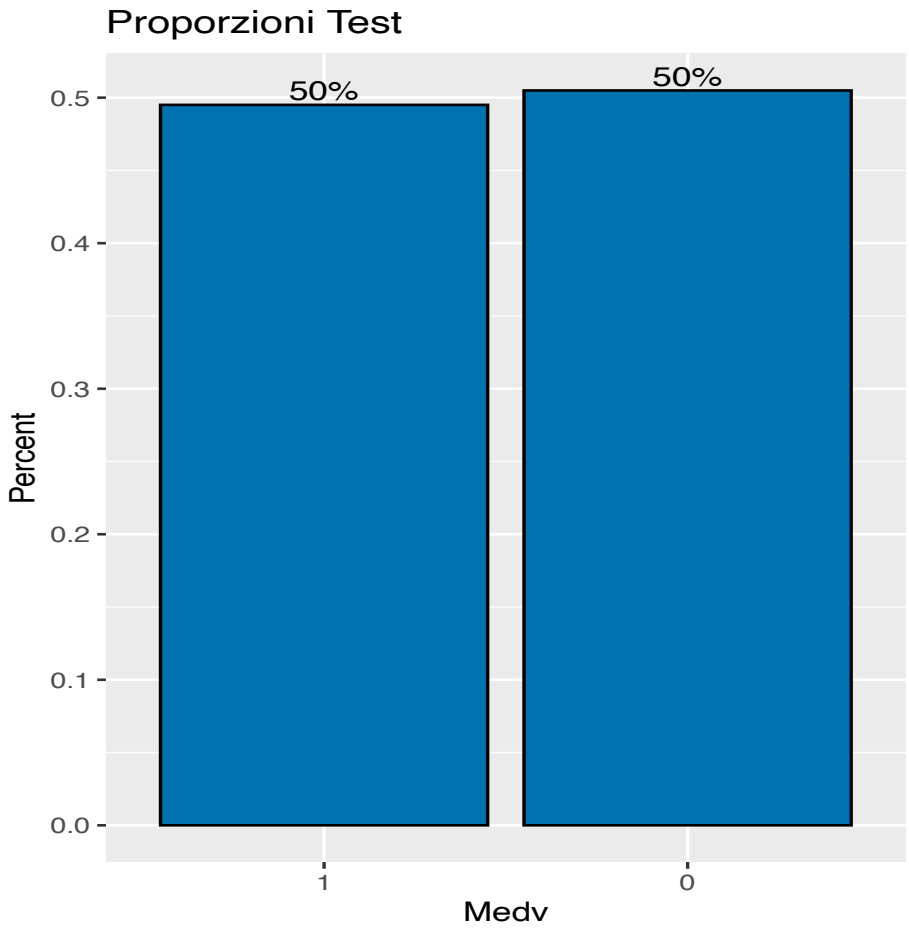
➢ *Training Set* :

Set di dati su cui viene costruito e messo a punto il modello.

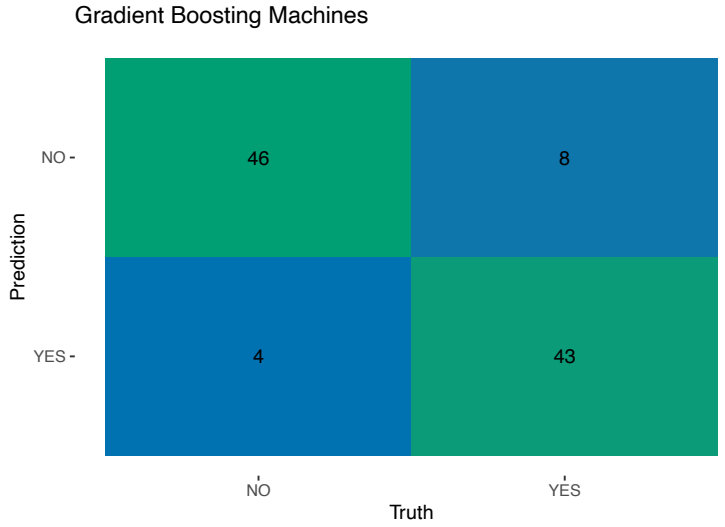


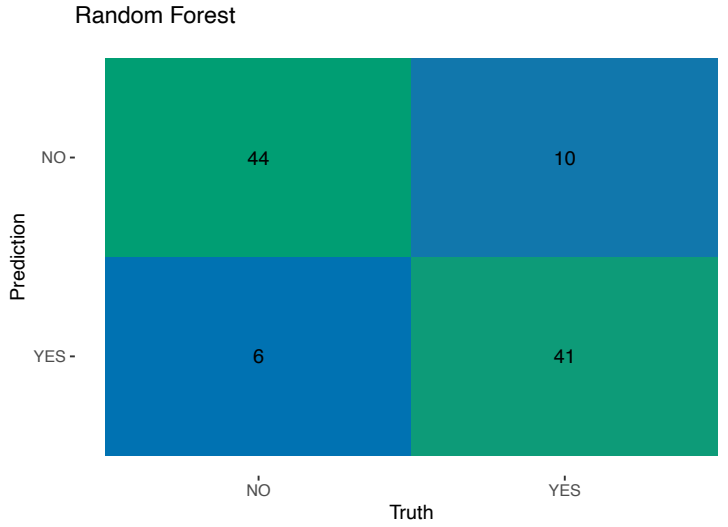
➢ *Testing Set* :

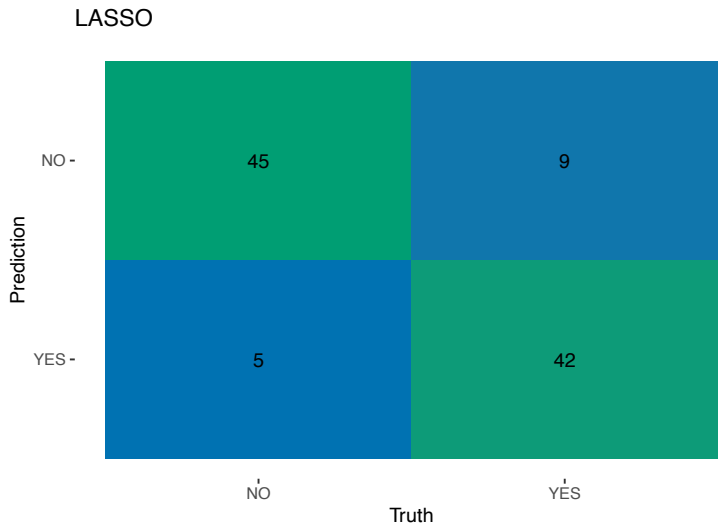
Set di dati su cui viene valutato il potere predittivo del modello.

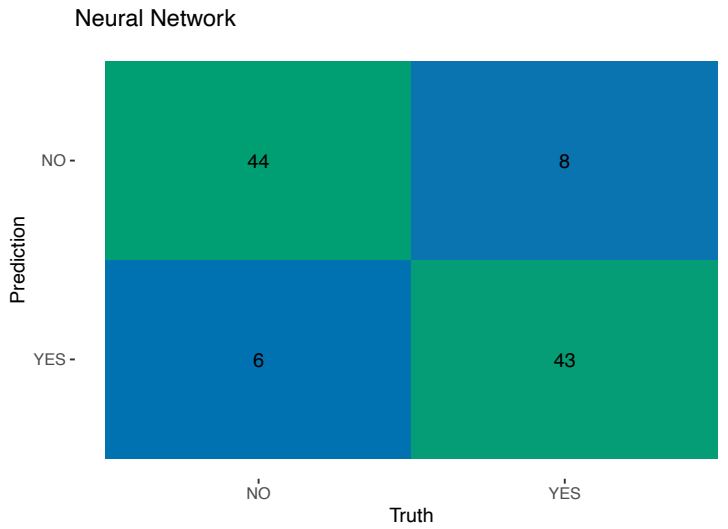


Confusion Matrix









Valutazioni modelli.

- Frazione di accuracy delle previsioni che l'algoritmo/modello ottiene correttamente:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + FN} =$$

- Sensibilità, T P R : è la probabilità che il prezzo di un'abitazione sia maggiore al valore mediano e sia classificato correttamente.

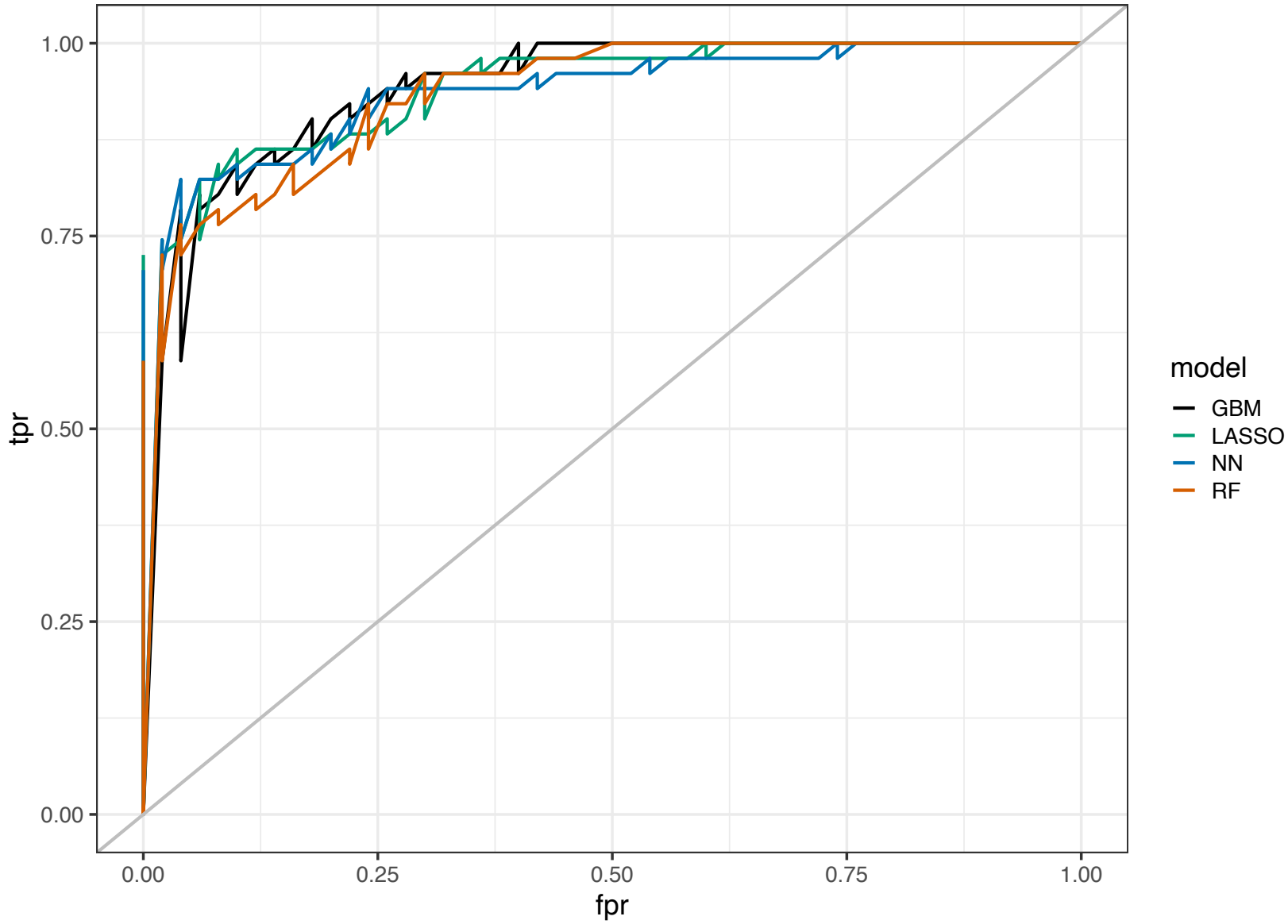
$$TPR = \frac{TP}{TP + FN} =$$

- Specificità, 1 − F P R : è la probabilità che il prezzo di un'abitazione sia maggiore al valore mediano e sia correttamente sia classificato come tale.

$$FPR = 1 - \frac{TN}{TN + FP} = \frac{FP}{FP + TN}$$

Risultati

- **Gradient Boosting Machines,** Area under the curve: ***0.9588***
- **LASSO**, Area under the curve: ***0.9475***
- **Random Forest**, Area under the curve: ***0.9508***
- **Neural Network**, Area under the curve: ***0.9302***



Perfomance

Boroughs :	GBM	RF	LASSO	NN
Accuracy :	***0.8812***	0.8416	0.8614	0.8614
95% CI :	(0.8017, 0.9371)	(0.7555, 0.9067)	(0.7784, 0.9221)	(0.7784, 0.9221)
No Information Rate:	0.505	0.505	0.505	0.505
P-Value [Acc > NIR] :	1.158e-15	1.158e-12	<4.937e-14	4.937e-14
Kappa :	0.7625	0.6834	0.723	0.7229
Mcnemar's Test P-Value :	0.3865	0.4533	0.4227	0.7893
Sensitivity :	0.92	0.88	0.9000	0.8800
Specificity :	0.8431	0.8039	0.8235	0.8431
Pos Pred Value :	0.8519	0.8148	0.8333	0.8462
Neg Pred Value :	0.9149	0.8723	0.8936	0.8776
Prevalence :	0.4950	0.4950	0.495	0.4950
Detection Rate :	0.4554	0.4356	0.4455	0.4356
Detection Prevalence :	0.5347	0.5347	0.5347	0.5149
Balanced Accuracy :	0.8816	0.8420	0.8618	0.8616
'Positive' Class :	NO	NO	NO	NO

- Le stime si basano sull'algoritmo di cross validation: addestra e testa il modello mettendo a punto i parametri con l'obiettivo di massimizzare la curva ROC.
- I migliori modelli in termini di AUC sono RF (***0.9508***) e GBM (***0.9588***) mentre LASSO (***0.9475***) e NN (***0.9302***) mostrano prestazioni inferiori.
- Il Migliore è il **Gradient Boosting Machines** poiché ha l'*accuracy* maggiore. E' opportuno guardare anche la non information rate - misura della prevalenze della classe meno prevalente – poiché se le due misure dovessero coincidere vi sarebbe un problema nella classificazione. Tuttavia è possibile osservare che l' *accuracy* è statisticamente superiore al tasso di *No Information Rate* per ciascuno dei modelli presentati.

Best tune

GBM	
Parametri	Valori
	9
N.trees :	150
Interaction.depth:	3
Shrinkage :	0.1
N.Minobsinnode :	10

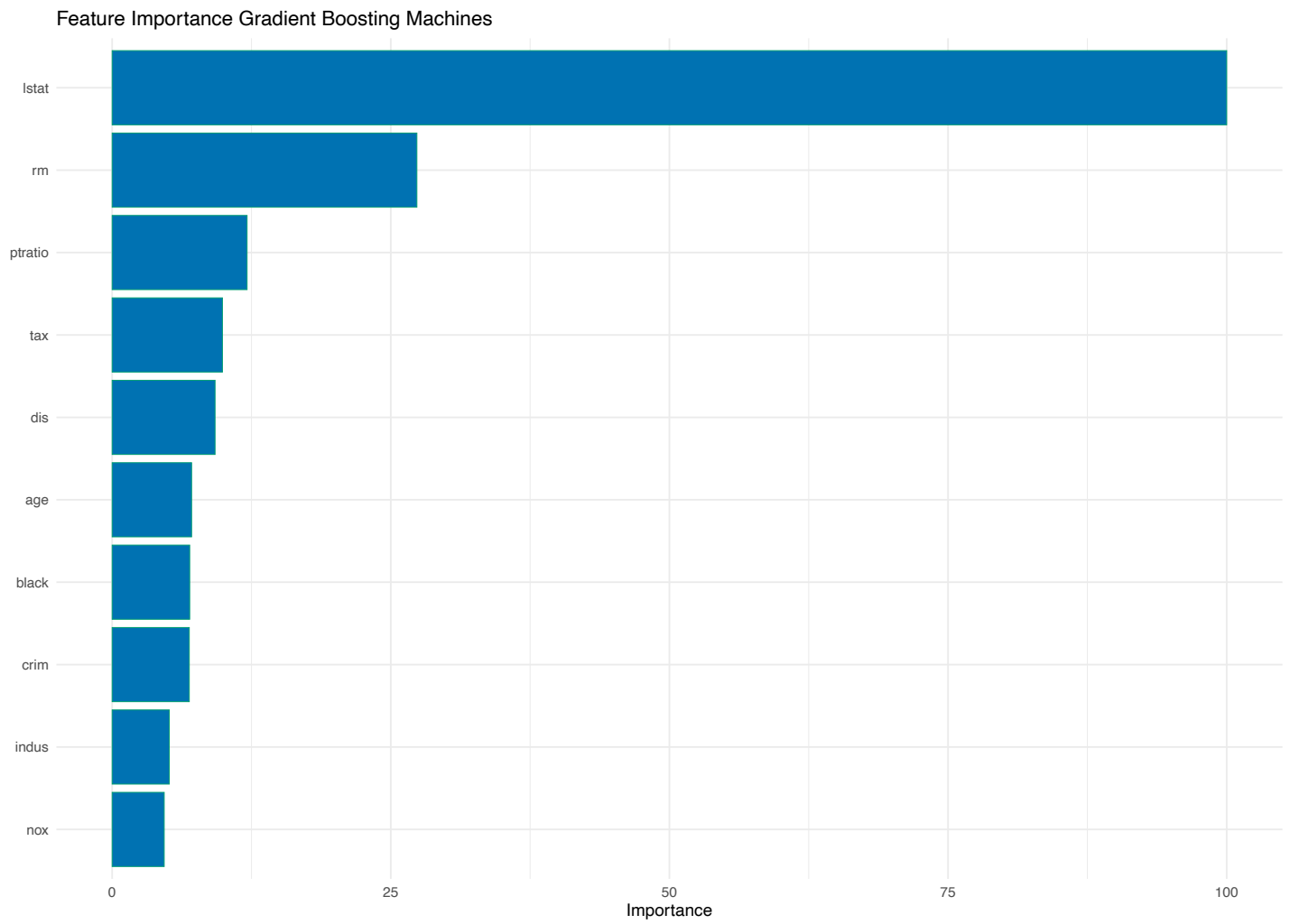
RF	
Parametri	Valori
	2
Mtry:	2

LASSO	
Parametri	Valori
	9
ALPHA:	0.1
LAMBDA:	0.00667

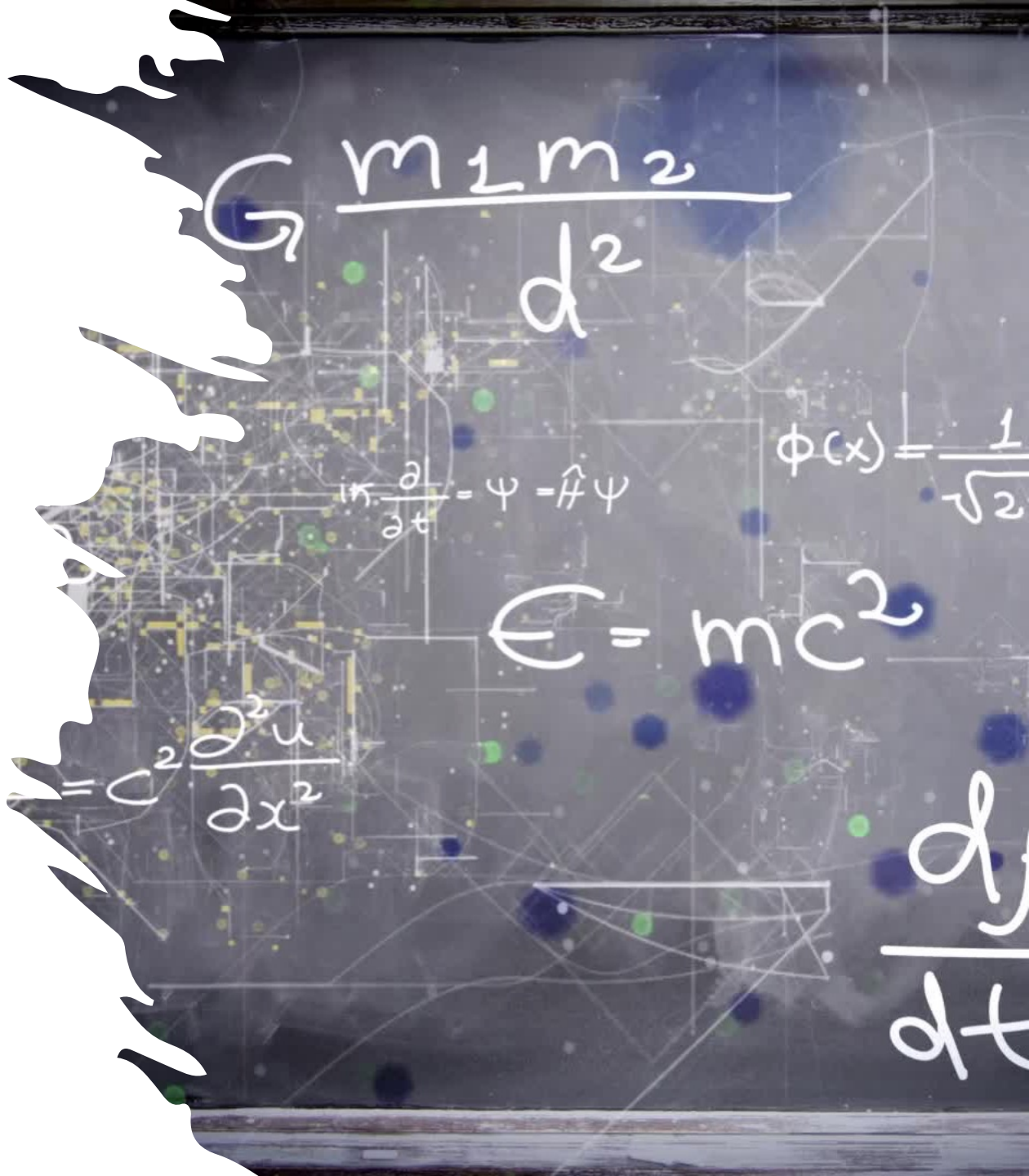
NN	
Parametri	Valori
	6
SIZE:	5
DECAY:	0.1

Feature Importance (FI): GBM

- *La FI è data sia dal numero di volte che la feature è utilizzata nei diversi modelli sia per quanta varianza migliora in un determinato split.*
- *In particolare indica il guadagno medio prodotto dalla caratteristica su tutti gli alberi in cui il guadagno è misurato dall'indice di Gini.*



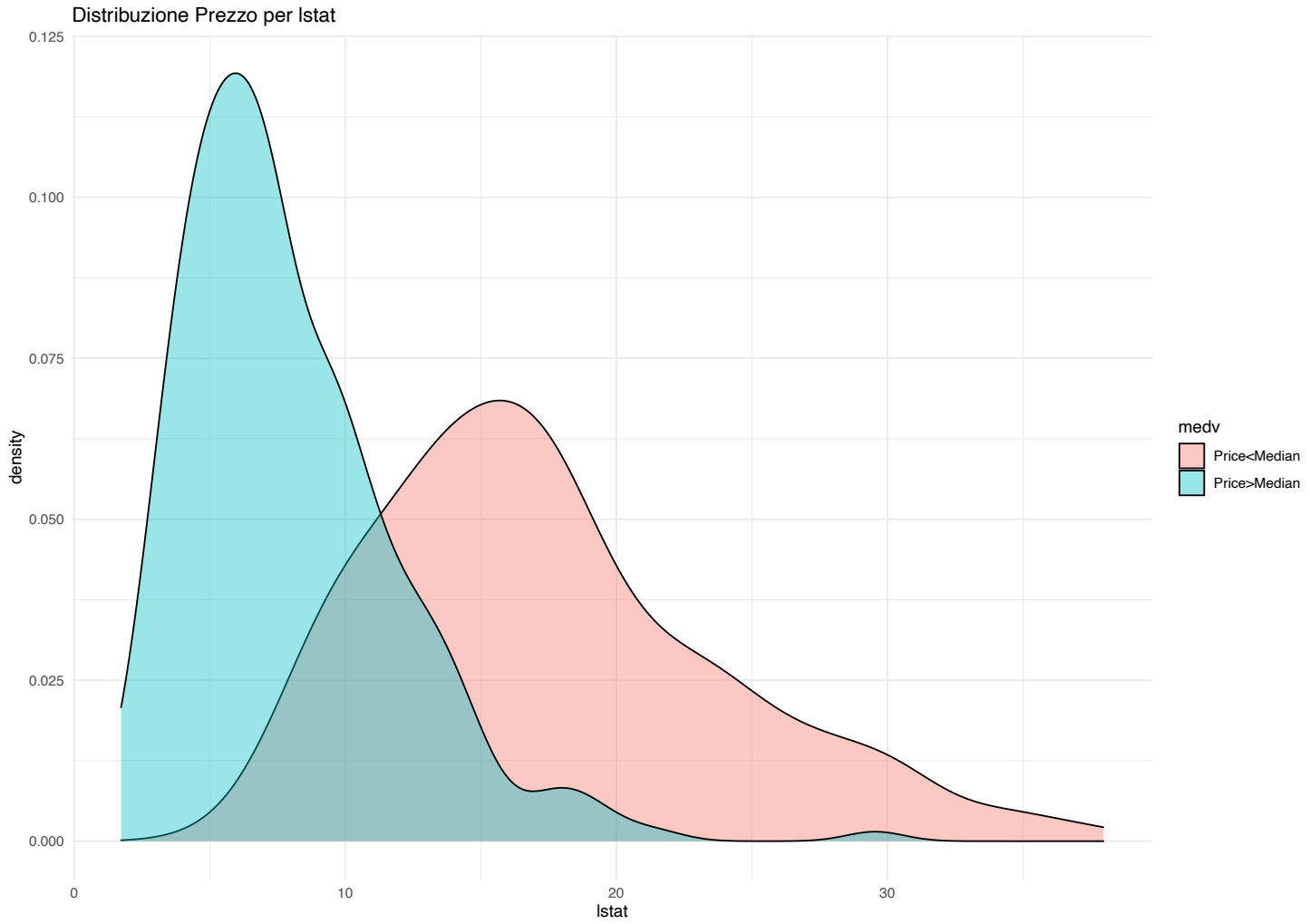
*Il fattore più importante risulta essere **LSTAT** (% Popolazione sotto la soglia di povertà). Tra le altre feature che hanno una maggiore importanza vi sono **RM** (numero medio di stanze per abitazione) e **PTRATIO**(rapporto alunni-insegnanti per città).*

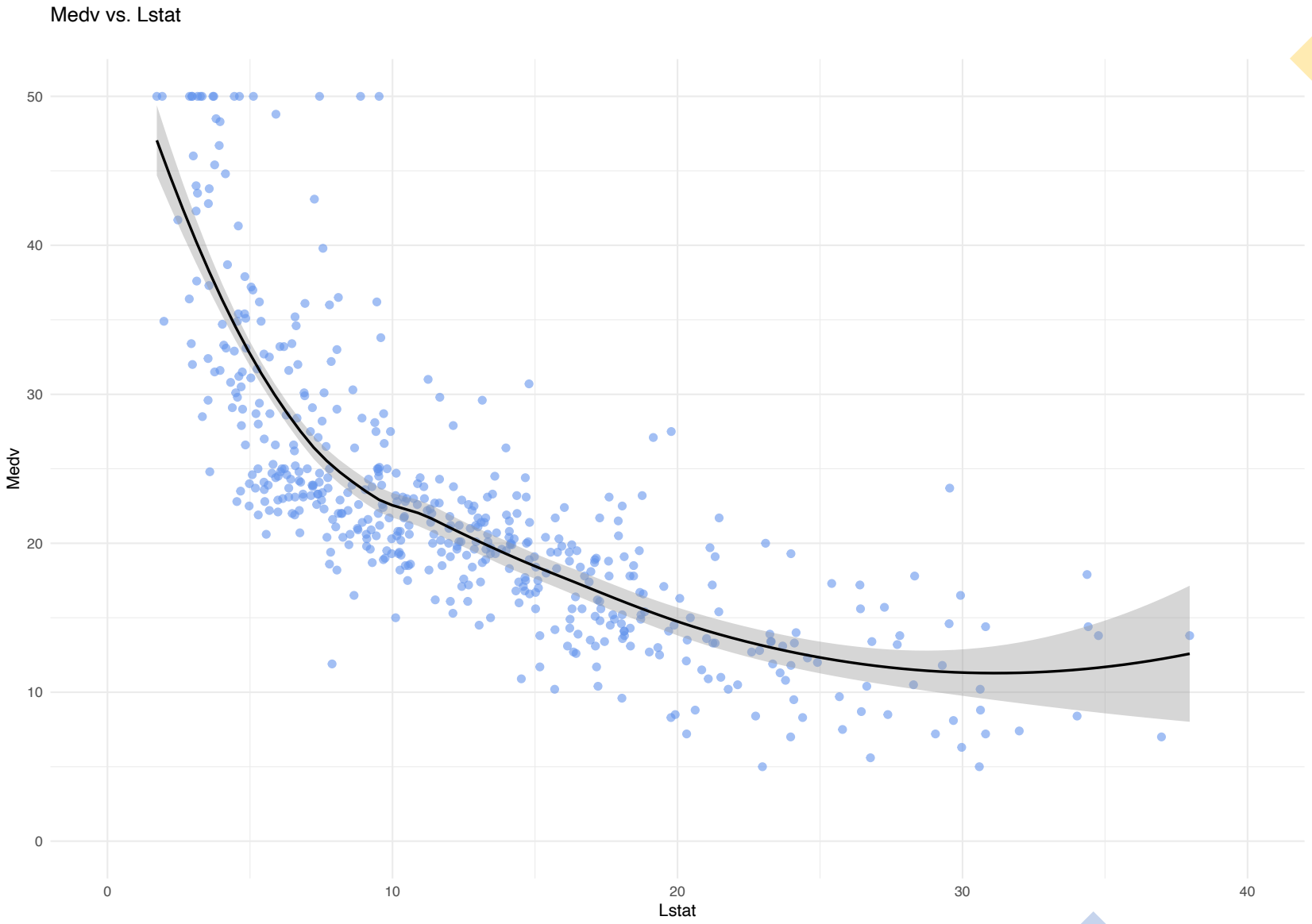


Feature LSTAT.

Osservando la funzione di distibuzione del valore medio degli appartamenti possono essere tratte alcune interessanti conclusioni.

- La % di popolazione sotto *la* ***soglia di povertà sembra avere un effetto negativo sul valore medio degli appartamenti***, infatti una bassa % di popolazione sotto la soglia di povertà contribuisce a determinare un prezzo degli appartamenti sopra il valore mediano. Viceversa nel caso opposto. ***Questo viene confermato anche dal grafico che segue***.




Medv vs. Lstat
Medv
Lstat
0
10
20
30
40
50
0
10
20
30
40


danke
謝謝
ngiyabonga
teşekkür ederim
спасибо
dank je
thank you
gracias
tapadh leat
hvala
mauruuru
mochchakkeram
bedankt
dziękuje
go raibh maith agat
sukriya
kop khun krap
obrigado
arigato
dakujem
grazie
merci
мерси