HOUSE PRICE PREDICTION

Alessandro Zanotta

13/06/2022

Abstract

Il dataset trattato mette a disposizione dati relativi alle abitazioni vendute in una contea non ben specificata dello stato di Washington.
Obiettivi dell'elaborato sono:

- Trovare attraverso l'utilizzo di modelli statistici e del machine learning il miglior modello in grado di prevedere il prezzo di vendita di una determinata abitazione date **determinate** caratteristiche
- Migliorare la previsione sul prezzo fornita dal modello lineare completo sul test senza nessun approfondimento in merito alle covariate a disposizione

La metrica utilizzata per misurare le performance è il mae (**mean absolute error**),il modello completo ha ottenuto sul test un mae pari a 0.091.
I risultati ottenuti sono stati soddisfacenti in quanto, è stato ottenuto un mae pari a 0.05055439 attraverso il metodo xgboost, e quindi ad abbassare di molto il risultato del modello completo.

1 Introduzione

Il training set disponeva di 17293 osservazioni e 19 features su cui allenare i modelli per prevedere la variabile *price* su un test set di 4320 osservazioni e 18 variabili in quanto la variabile target non era a disposizione nel test set.
Obiettivo prinicipale dell'analisi è trovare il miglior modello in grado di minimizzare il *MAE*, motivo per cui si è optato per una divisione attraverso un campionamento stratifitcato della variabile target *price* del training a disposizione in ulteriori training e test.
Un altro obiettivo dell'analisi è di migliorare il *MAE* del modello lineare completo. In tale modello sono state utilizzate tutte e 18 le variabili a disposizione per prevedere la variabile *price*, senza nessun tipo di analisi esplorativa e feature engigneering, riportando un *MAE* pari a 0.09 che sarà poi utilizzato come benchmark per la valutazione successiva dei modelli proposti.

2 ANALISI ESPLORATIVA

Le variabili a disposizione per le analisi sono le seguenti: date sold,bathrooms,sqft living,sqft lot,floors,waterfront,view,condition,sqft above,sqft basement,year built (yr built),year renovated,zip code,latitude,longitude,nn sqft living,nn sqft lot. Nessuna di queste presenta valori mancanti

La variabile target price, nel training è presente in scala logaritmica base 10, tale trasformazione porta ad una distribuzione gaussiana del prezzo.

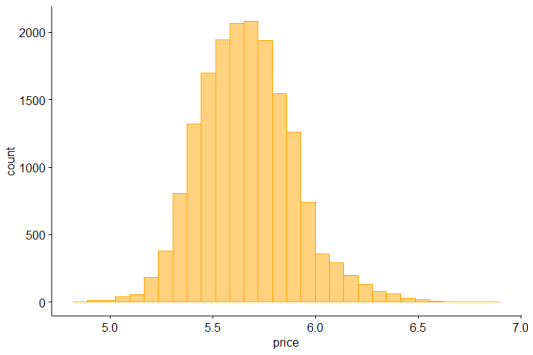

Figure 1: Distribuzione Log(Price)

Tra le variabili all'interno del dataset sono presenti latitudine e longitudine, attraverso cui è possibile fare un grafico spaziale delle posizioni delle abitazioni ed una valutazione dei diversi prezzi in base alla locazione. Come si nota dal grafico è

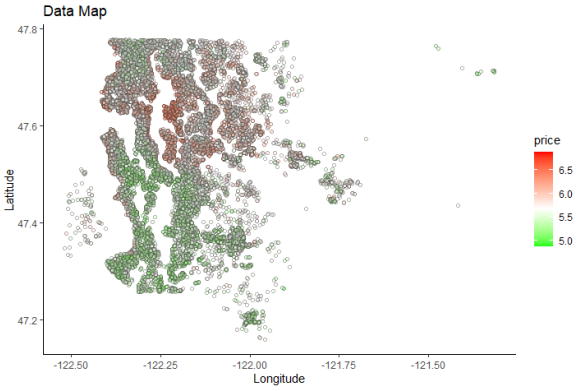

Figure 2: Price in base a latitudine e longitudine

possibile evincere la presenza di corsi d'acqua tra le abitazioni, con una concentrazioni di prezzi elevati per le abitazioni che si trovano verso nord su un corso d'acqua. Nel dataset, a conferma di tale osservazione è infatti presente la variabile waterfront, la quale indica se l'abitazione affaccia su un corso d'acqua o meno. Nell'analisi esplorativa dei dati emerge un'importante osservazione: la variabile sqft living è combinazione lineare delle variabili sqft basement ed sqft above. Si è quindi optato per una trasformazione logaratmica di base 10 per entrambe le variabili, in modo tale da perdere il legame lineare tra le due, ed abbassare notevolmente la correlazione tra le tre variabili in questione. La trasformazione avvicina la distribuzione della variabile sqft living alla distribuzione della variabile target price come mostrato in figura 3. Un ragionamento analogo è stato seguito anche per le variabili sqft lot, nn sqft living,nn sqft lot.
La variabile bedrooms presentava i valori riportati nella tabella qui di fianco a destra:

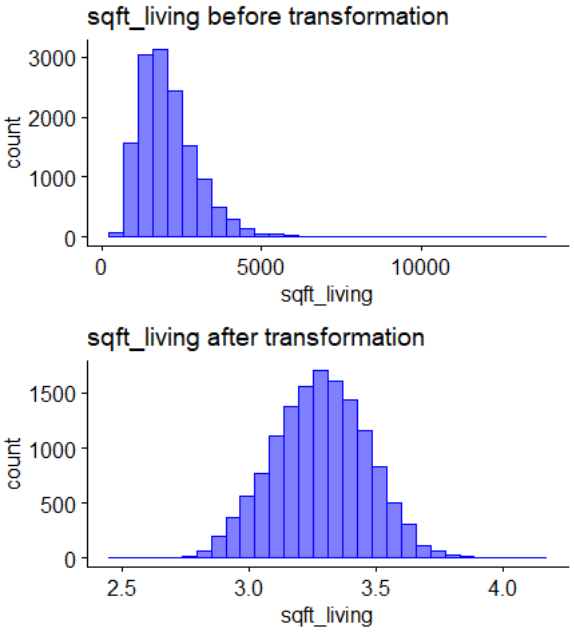

Figure 3: Distribuzione Log(Sqft living)

numero camere	numero case
0	11
1	131
2	1779
3	6285
4	4390
5	1028
6	166
7	28
8	8
9	4
10	1
11	1
33	1

Sono presenti case con rispettivamente 10,11,33 camere da letto, che per un'abitazione risultano valori piuttosto anomali. Si è deciso quindi di investigare sulla presenza di tali valori

Il grafico qui riportato4 mostra come la dimensiona dello spazio abitabile per le case che presentano tali valori anomali sia molto simile a case con un minore numero di camere da letto. Più in particolare l'abitazione con 33 camere da letto presenta un valore pari alla mediana (metrica preferebile in presenza di ouliers) del prezzo delle abitazioni con

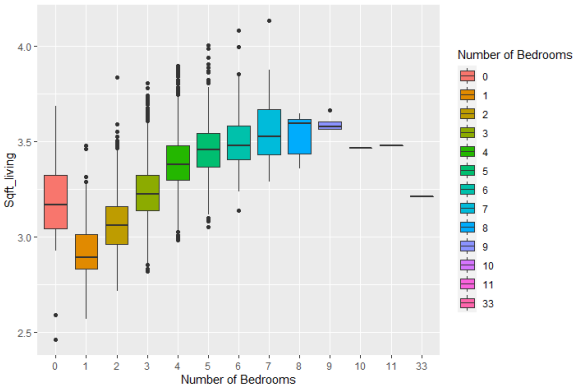

Figure 4: Metri quadri abitabili in base al numero di camere da letto

tre camere da letto, motivo per cui si è optato per l'imputazione di 3 camere da letto all'abitazione che ne riportava 33, considerando tale numero conseguenza di un errore di imputazione. Ragionamento simile è stato seguito per le case che presentavano un numero di camere da letto maggiore a 8, scegliendo come numero da imputare l'8.
Altri valori inusuali presenti nel dataset riguardano la variabile *bathroooms*, sono infatti presenti alcuni numeri che riportano una parte decimale. Effettuando una ricerca si è scoperto che in America il numero *0.5* riferito ai bagni indica un bagno con solo toilet e lavandino e sprovvisto di doccia [2] .Motivo per cui si è deciso di lasciare i valori relativi ai bagni come erano presenti nel dataset.
Situazione simile si ha per la variabile floors, che indica il numero di piani presenti all'interno della casa, 0.5 si è ipotizzato che potesse essere per la presenza di un soppalco, motivo per cui è stata lasciata inalterata. Sqft lot rappresenta il terreno, su cui è stata costruita l'abitazione, tale valore è stato mantenuto inalterato. Per le altre variabili si riporta che la variabile waterfront è molto importante, in quanto c'è un cambio di prezzo significativo al variare delle modalità, 0 (abitazione non su corso d'acqua) 1 (abitazione su corso d'acqua). Ragionamenti analoghi valgono per le variabili qualitative condition che ha 4 modalità e view che descrive la qualità della vista.

3 Feature Engigneering

Nella fase di feauture engigneering sono state create diverse variabili per i diversi modelli utilizzati e sono state effettuate diverse operazioni, tenendo conto in alcuni casi particolari che la variabile numerica più correlata con la variabile risposta è sqft living.

3.1 Modello lineare

Per il modello lineare sono state effettuate le seguenti operazioni:

- E' stata rimossa la variabile sqft above, in quanto anche dopo la trasformazione con il logaritmo base 10 presenta un'elevata correlazione con la variabile sqft living, è infatti pari a 0.8659[1].

- E' stata creata la variabile tipologia, che in base al numero di bagni assume tre diverse modalità: **open space** per le osservazioni che presentano tra 0 e 1 bagno. **normale** per le osservazioni che presentano tra 1 e 4.25 bagni, **grande** per le osservazioni che hanno più di 4.25, sono state considerate abitazioni grandi, tale ragionamento si è basato sul boxplot qui riportato.

- Per la variabile floors si è optato per dividere in categorie di case da 1 piano fino a 2.5 piani ed assegnare per la categoria **others** per alle abitazioni con più di 2.5 piani. Tale operazione è giustificata dal cambiamento significativo del prezzo a seconda dell'appartenenza a queste tre macro categorie

- è stata creata la variabile *storica*, che assume modalità si o no, in base all'anno di costruzione, come benchmark è stato preso in considerazione il 1980. Un'altra variabile creata è stata *renovated* che assume modalità si oppure no. Altra variabile creata è stata **sqft garden**, è una variabile numerica che presuppone l'esistenza di un giardino in quanto formata dalla sottrazione tra sqft lot e sqft living, se tale differenza è maggiore di 0 allora la casa ha un giardino di dimensione pari alla differenza.

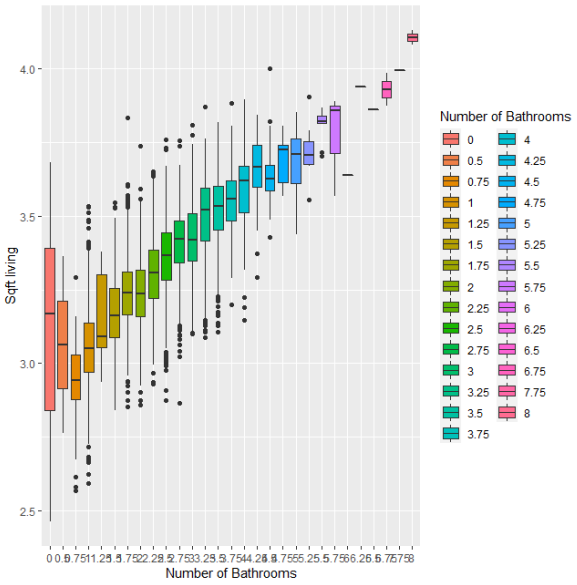

Figure 5: Metri quadri abitabili in base al numero di bagni

3.2 Modelli con alberi

- E' stata presa la decisione di rimuovere la variabile sqft above in quanto altamente correlata con sqft living e sqft basement, come suggerito dalla letteratura, tale operazione migliora le performance dei modelli con gli alberi [1]
- E' stata creata la variabile **year old**, formata dalla differenza tra l'anno in cui è stata venduta e l'anno in cui è stata rinnovata oppure costruita, viene preso in considerazione il maggiore tra i due.
- La variabile date sold è stata mantenuta e considerata come variabile numerica, ad esempio la data '11/05/2014' diventa 11052014
- Per la variabile zip code, sono state assegnate a categoria "others" gli zip code che si presentavano con frequenza relativa minore di una soglia pari a 0.05.
- Anche in questo caso come per i modelli lineari per le variabili sqft living, sqft basement,sqft lot, nn sqft living, nn sqft lot, è stata applicata la trasformazione logaritmica di base 10,nonostante la lettura suggerisca che non sia necessario [1].
- Sono state create le variabili **percabove** e **percbase** formate dalle percentuali di sqft above e sqft basement rispetto al totale dell'abitazione (sqft living).
- Anche in questo caso è stata creata la variabile sqft garden, che rappresenta la dimensione di un eventuale giardino.

4 Scelta dei modelli

Per la valutazione dei modelli e la scelta degli iperparametri è stato seguito il principio della repeated cross validation(eccezion fatta per i modelli lineari senza splines), da cui è possibile ricavare la stima del generalized error, che descrive la capacità di generalizzazione del modello. Tale risultato è stato poi confrontato con l'empirical error, ovvero con l'errore che il modello riporta sui dati di training. Se i due numeri sono simili, allora significa che il modello è robusto e ci si può attendere sui nuovi dati un risultato simile.

4.1 Modello lineare

Come primo modello è stato utilizzato il modello lineare senza interazioni e senza splines, con il feature engigneering esposto in precedenza, i risultati del primo modello senza interazioni e senza splines sono qui riportati:

Generalized Error	Empirical Error
0.0591	0.05841458

Table 1: Modello lineare base

Due valori molto simili tra di loro che suggeriscono robustezza del modello, tale assunzione sarà poi verificata più avanti con la stima sul test a disposizione.

4.2 Modello lineare con interazioni

Il modello lineare con le interazioni ha riportato i seguenti risultati:

Vengono aggiunte quindi le interazioni al nostro modello, in quanto si è verificato un sostanziale miglioramento di 0.01, rispetto al modello senza, e data la natura della variabile target, questo è

Generalized Error	Empirical Error
0.0582	0.05487891

Table 2: Modello lineare con interazioni

un risultato significativo. In questo caso i risultati del Generalized error e dell'Empirical error suggeriscono una scarsa robustezza del modello, in quanto differiscono di 0.04. Tale assunzione sarà poi verificata più avanti con la stima sul test a disposizione.

4.3 Modello lineare con interazioni e splines

Come ultima aggiunta al modello sono state fatte le splines, sulle variabili latitudine e longitudine, come gradi di libertà sono stati scelti rispettivamente 18 e 22. Ecco i risultati ottenuti:

Generalized Error	Empirical Error
0.055	0.05246732

Table 3: Modello lineare con interazioni e splines

Si ha ancora un ulteriore miglioramento delle performance. I due valori si presentano distanti ma comunque più vicini rispetto al modello con le interazioni, tale osservazione può essere sinonimo di sovraddamento ai dati, ipotesi che sarà verificata nel momento della stima sul test a disposizione.

4.4 Random Forest

Il primo modello non lineare utilizzato è stato il Random Forest con iperparametri pari a 9 per mtry, ovvero numero di variabili scelte casualmente ad ogni split, 5 per min n, ovvero il numero minimo di osservazioni che devono essere presenti all'interno di ogni nodo prima di un'ulteriore suddivisione. I risultati ottenuti sono qui riportati:

Generalized Error	Empirical Error
0.0574	0.05385573

Table 4: Random Forest

Come ci si attendeva il Random Forest si adatta particolarmente ai dati, mostrando una differenza significativa tra l'empirical ed il generalized error.

4.5 XGBoost

L'ultimo modello valutato è stato l'xgboost, che ha riportato i seguenti risultati:
Come ci si attendeva c'è una differenza sostanziale

Generalized Error	Empirical Error
0.0516	0.03311142

Table 5: XGBoost

tra il generalized error e l'empirical error, in quanto l'xgboost è particolarmente portato ad adattarsi ai dati su cui viene stimato.L'xgboost è il modello che presenta i risultati migliori in termini di generalized error. Gli iperparametri scelti in base al principio della repeated cross validation sono stati i seguenti :

- tree depth = 15
- trees = 500
- learn rate = 0.0309754
- min n = 47
- lossreduction = 2.2e-06

5 Scelta Modello finale

Vengono qui riportate le performance dei modelli ottenute sul test a disposizione:

Modello	Test Error
Lineare con splines e interazioni	0.05385573
Random Forest	0.0536
Xgboost	0.05055439

Table 6: Risultati sul test

In base a questi risultati si sceglie di utilizzare il modello xgboost per la previsione finale. Gli obiettivi posti all'inizio dell'elaborato sono stati raggiunti, si è riusciti infatti a migliorare di parecchio la performance del modello completo e si è stati in grado di trovare un modello che garantisce un MAE relativamente basso e una certa robustezza nelle analisi.

References

[1] *A Recommended Preprocessing — Tidy Modeling with R*. https://www.tmwr.org/pre-proc-table.html. (Accessed on 06/17/2022).

[2] *What Is a Half-Bath, Quarter Bath and Three-Quarter Bath? — The Realty Firm*. https://therealtyfirms.com/half-bath-quarter-bath-three-quarter-bath/. (Accessed on 06/17/2022).