House sale prices prediction using machine learning algorithms

Davide Luperi

Esame di data mining - modulo Applied Predictive Modelling

Sommario

Le tecniche di machine learning in ambito previsivo comprendono modelli di tipo parametrico e non parametrico.

Nel corso degli anni sono state sviluppate tecniche di previsione sempre più precise con lo scopo di rendere minimo l'errore di previsione.

L'obiettivo di questo lavoro è stato quello di sviluppare, migliorare e confrontare modelli previsivi al fine di trovare il migliore (o i migliori) da usare per effettuare una previsione finale

---

1 Introduzione

Il lavoro effettuato è stato suddiviso nelle fasi classiche di un processo di analisi di dati: il focus della prima parte è quello dell'analisi esplorativa e della 'pulizia' dei dati, seguito da una fase di elaborazione dei modelli e tuning dei parametri ed infine confronto dei modelli sulla base dei risultati.

Il dataset (3) a disposizione era già suddiviso in train set (17293 osservazioni di cui era disponibile il valore della variabile risposta) e test set (4320 osservazioni di cui era ignota la variabile risposta)

Le osservazioni facevano riferimento a delle case, di cui erano a disposizione i valori di determinate caratteristiche quali ad esempio numero di camere da letto o metri quadri. La variabile risposta corrispondeva al prezzo delle stesse in trasformata logaritmica.

L'obiettivo è stato quello di prevedere il valore del prezzo delle case presenti nel test set sulla base delle informazioni del training set.

I modelli che sono stati valutati sul training set sono il modello lineare, il random forest, il kriging ordinario (per l'analisi spaziale) e l'xgboost. Ove possibile è stato effettuato un tuning dei parametri per ottimizzare al massimo le performance

La metrica statistica che è stata tenuta in considerazione per valutare i risultati dei modelli è stata il MAE (Mean Absolute Error)

2 Analisi esplorativa e pre elaborazione dei dati

Dopo aver verificato che nel dataset non ci fossero dati mancanti è stata studiata la distribuzione della variabile risposta nel training set. Come è possibile notare anche dalla curva sovraimposta all'istogramma, la distribuzione sembra approssimarsi molto bene ad una distribuzione normale

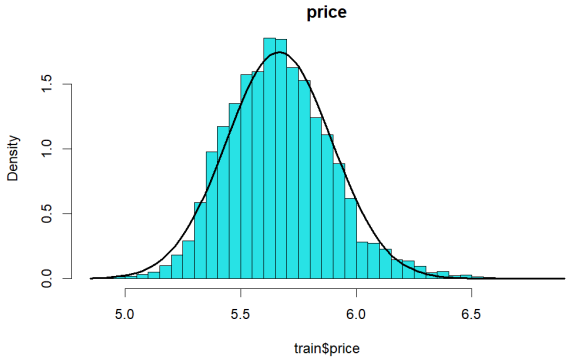

Figura 1: Distribuzione della variabile price

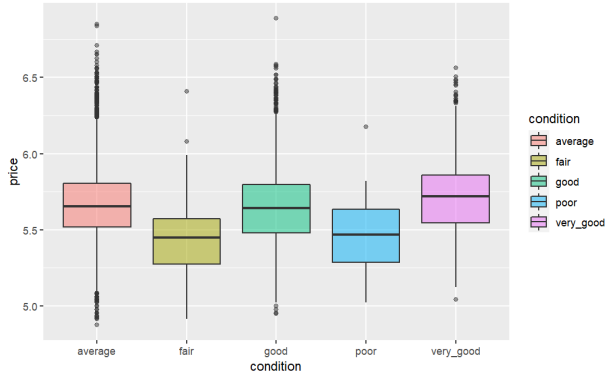

Figura 3: Distribuzione della variabile condition

2.1 Variabili fattoriali

Successivamente il lavoro si è concentrato sulla ricodifica delle variabili.

Nel dettaglio sono state convertite in variabili fattoriali le variabili *waterfront, condition* e *view* che erano in partenza di tipo carattere.

La variabile *date sold* contenente la data di vendita della casa, è stata trasformata tenendo in considerazione solo il mese di vendita. La nuova variabile creata è stata rinominata *month*

Si può analizzare la distribuzione delle variabili fattoriali in base alla variabile *price* utilizzando dei boxplot come mostrato di seguito.

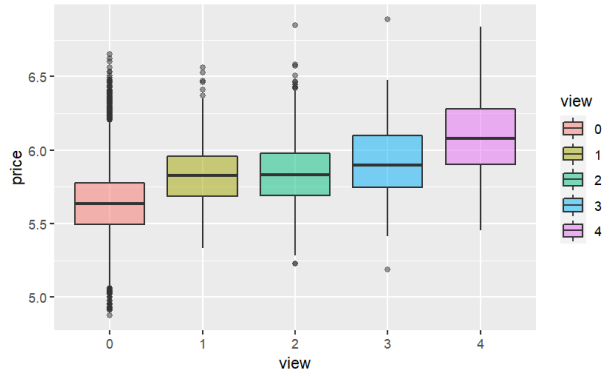

Figura 4: Distribuzione della variabile view

2.2 Variabili spaziali

Nel dataset sono presenti anche le variabili di geolocalizzazione delle abitazioni, ovvero *longitude* e *latitude*.

Grazie a queste informazioni è possibile sapere la collocazione delle case nel mondo, e se è possibile individuare aree in cui la dipendenza territoriale influisce notevolmente la variabile *price*

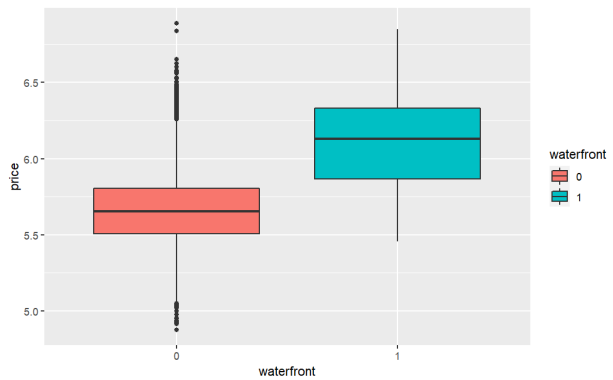

Figura 2: Distribuzione della variabile waterfront

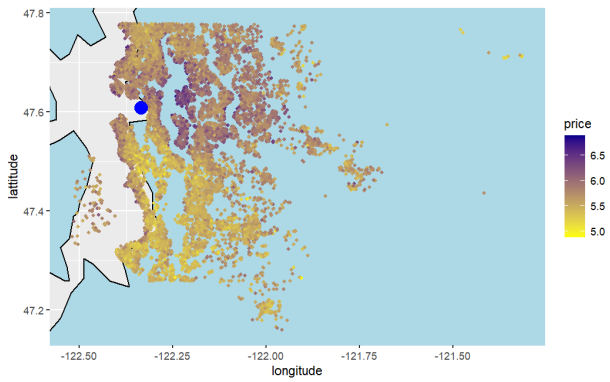

Figura 5: Distribuzione della variabile view

La mappa ci permette di geolocalizzare le osservazioni nella regione di Seattle, WA, Stati Uniti (2). Il centro di Seattle è rappresentato dal cerchio blu.

Si nota inoltre una certa correlazione tra posizione dell'abitazione e prezzo della stessa: sembra infatti che i quartieri nella zona centrale e più ad est del centro siano più cari, mentre risultano più economici quelli a sud del centro

2.3 Variabili continue

Mediante analisi di misure statistiche e boxplot, si è riscontrato un unico outlier nell'osservazione avente 33 come valore della variabile *bedrooms*. Questa osservazione è stata eliminata dal dataset.

Per avere una visuale sintetica della correlazione delle variabili indipendenti tra di esse e della correlazione che queste hanno con la variabile in oggetto viene presentato un grafico riassuntivo

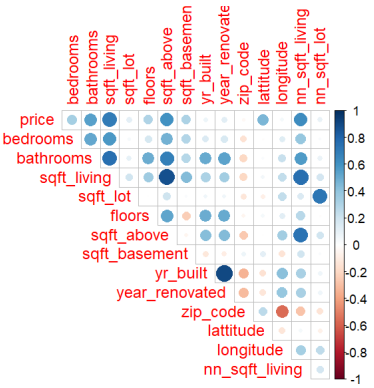

Figura 6: grafico delle correlazioni

Alcune coppie di variabili presentano tra di loro alta correlazione (2 coppie hanno correlazione maggiore di 0.8)

Si è deciso comunque di tenere tutte le variabili nel dataset

3 Modelli e risultati

In questa sezione vengono presentati i modelli allenati sul training set. Per valutare i modelli è stato utilizzata la tecnica delle k-fold con k pari a 4, e la misura del MAE presentato per ogni modello risulta essere la media del MAE calcolato sul validation set di ogni fold.

3.1 Modello lineare

Il primo modello che è stato provato è stato il modello di regressione lineare semplice, ovvero senza interazioni.

Il risultato ottenuto è stato:

MAE
0.09

Successivamente è stato valutato il modello lineare aggiungendo come regressori le variabili continue al quadrato e tutte le interazioni tra al più due variabili: è stato dunque appesantito notevolmente il modello con la speranza di ottenere un risultato migliore.

MAE
0.068

3.2 Kriging ordinario

Il secondo modello parametrico che è stato implementato è stato il kriging ordinario (1). Questo modello fa parte di una gamma di modelli parametrici utilizzati per l'analisi dei dati spaziali. Nello specifico, avendo a disposizione le coordinate delle osservazioni, oltre al valore della variabile target, i modelli di kriging permettono di effettuare previsioni su nuove coordinate sulla base dei valori assunti dalla variabile target nello spazio.

Si tratta comunque di modelli che per essere al meglio compresi richiedono di una più profonda conoscenza teorica dell'argomento, che non verrà trattata in questo testo.

Un approccio comune utilizzato con i modelli di kriging è quello di mappare lo spazio di riferimento mediante una griglia, dividendo dunque lo spazio in quadrati di dimensione uguali. Sul centroide di questi quadrati viene poi effettuata la previsione del valore della variabile target definisce il valore della variabile su tutto il quadrato.

Si può avere in questo modo una visuale delle previsioni effettuate dal modello

Utilizzando i dati a noi a disposizione è possibile visualizzare il comportamento del modello di kriging con la seguente mappa

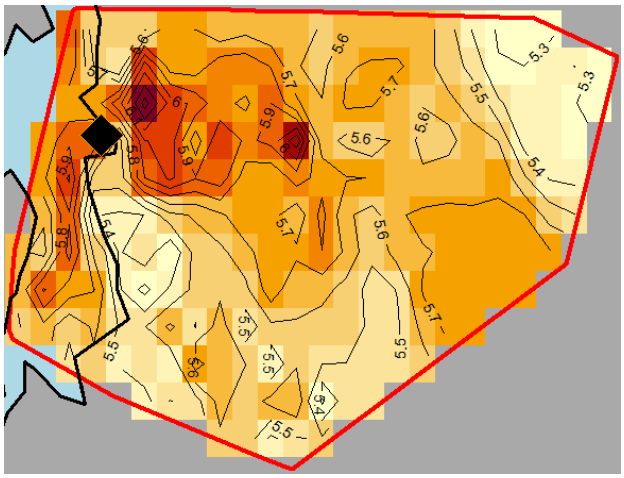

Figura 7: mappa delle previsioni fornite dal modello su una griglia di 861 aree

Per effettuare una previsione su dei punti specifici dello spazio è sufficiente sostituire i centroidi delle aree della griglia con le coordinate delle osservazioni

I risultati ottenuti con questo modello sono:

MAE
0.092

I valori iniziali utilizzati per fittare il variogramma sono:

- pesi di Cressie;
- parametri iniziali = 0.06, 0.2 ;
- modello di tipo esponenziale;
- nugget = 0.001.

3.3 Random forest

Il primo modello non parametrico implementato è stato il random forest. Questo algoritmo si basa su un principio di boosting delle osservazioni e delle variabili su cui implementare un numero elevato di alberi di regressione.

Si può osservare dalla seguente immagine come all'aumentare del numero di alberi di regressione creati, diminuisce l'errore.

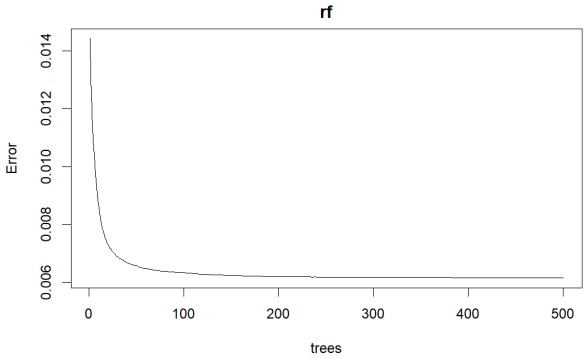

Figura 8: errore relativo al numero di alberi

Per questo algoritmo è necessario impostare degli impostare degli iperparametri a priori, tra cui il numero di alberi, il numero di variabili da considerare ad ogni nodo dell'albero ed il numero minimo di osservazioni ai nodi finali.

Su tutti gli iperparametri è stato eseguito un tuning per trovare la combinazione che minimizzasse l'errore

I valori finali trovati sono:

- nodesize=2;
- ntree=500;
- mtry=14;

MAE
0.057

3.4 XGboost

L'ultimo modello utilizzato per effettuare previsioni è stato l'XG-boost (4). L'algoritmo combina la logica predittiva degli alberi di regressione con l'ensemble (gradient boosting) di diversi modelli che operano sui dati originali e sui residui dei modelli.

Anche per questo algoritmo è necessario fornire come input il valore di alcuni iperparametri, la cui scelta viene effettuata mediante la tecnica di tuning:

- nrounds=100;
- eta = 0.1;
- maxdepth=15;
- min child weigth=1;
- max delta seps=4.

MAE
0.052

I risultati ottenuti con questo modello sono i risultati migliori ottenuti in assoluto, ed è quindi stato deciso di utilizzare il modello XGboost con i parametri presentati per effettuare la previsione sulle osservazioni presenti nel test set.

Riferimenti bibliografici

[1] Pesquer, Lluís, Ana Cortés, and Xavier Pons. "Parallel ordinary kriging interpolation incorporating automatic variogram fitting." Computers geosciences 37.4 (2011): 464-473.

[2] https://visitseattle.org/visitor-information/maps/

[3] https://aldosolari.github.io/DM/

[4] https://towardsdatascience.com/xgboost-theory-and-practice-fb8912930ad6