Milan housing

Marco Nobili

Anno Accademico 2024/2025

Abstract

La presente analisi si pone l'obiettivo di fornire una previsione del prezzo di alcune abitazioni nella città di Milano (e dintorni) sulla base delle caratteristiche di tali immobili. È stata effettuata una consistente procedura di preprocessing e sono stati costruiti modelli quali regressione lineare (con backward e forward regression), ridge regression, lasso ed elastic net. Il miglior modello in termini previsivi, sulla base del MAE, è risultato essere un modello lineare applicato ad un'opportuna trasformazione della variabile risposta.

1 Introduzione

L'analisi è stata effettuata sul dataset disponibile al seguente link. Il dataset `training.csv` contiene 8000 osservazioni con la relativa variabile risposta (i.e. il prezzo di vendita), mentre il dataset `test.csv` è composto da 4000 record senza tale informazione. La performance su quest'ultimo dataset viene valutata tramite il Mean Absolute Error (MAE), definito come $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$.

1.1 Descrizione del dataset: analisi descrittive e pre-processing

Di seguito vengono presentate le variabili contenute nel dataset. Per ognuna di esse, si riportano le operazioni di pre-processing effettuate ed eventuali statistiche descrittive (se interessanti per l'analisi).

- `square_meters`: Numerica, no `NA`

Sono presenti 7 osservazioni con una superficie inferiore a 14 m^2 (minimo per abitante per legge). Potremmo imputare questi valori sulla base di altre caratteristiche della casa (come faremo per i dati di test) ma, per evitare inutili distorsioni, decidiamo di escludere queste osservazioni dall'analisi.

- `bathrooms_number`: Categoriale, con `NA`

Sebbene rappresenti un concetto numerico, presenta 4 categorie: *1 (bagno)*, *2*, *3*, *3 o più*. Sono presenti 25 `NA`, che decidiamo di imputare assegnando un bagno ogni 75 m^2 di superficie.

- `floor`: Categoriale, no `NA`

Questa variabile indica il piano in cui si trova l'edificio. Viene codificata come categoriale in quanto (oltre che i valori da 1 a 9) sono presenti le classi *Piano Terra*, *Mezzanino* e *Seminterrato*.

- `total_floors_in_building`: Numerica, con `NA`

Imputiamo i valori mancanti come $max(1, \texttt{floor})$, in quanto supponiamo che un `NA` possa essere presente se la casa ha un solo piano (esempio: una villetta), ma se l'abitazione è ad un piano superiore al primo sicuramente non siamo in questo caso (e quindi assegniamo all'immobile il minimo numero di piani possibile). Per pulire i dati, sostituiamo anche tutti i valori di questa variabile che sono minori di `floor` con il valore presente in tale variabile.

- `lift`: Categoriale, con `NA`.

Questa variabile indica la presenza (*Sì*) o meno (*No*) dell'ascensore. Molti degli `NA` sono relativi ad abitazioni al piano terra o in edifici con un solo piano. Aggiungiamo quindi la categoria *Inutile* in cui inseriamo tutte le osservazioni (anche quelle non mancanti) che rientrano in tale casistica. I restanti `NA` vengono inseriti nella classe *No*.

- `rooms_number`: Categoriale, no `NA`

Come per `bathrooms_number`, sono presenti le classi *1 stanza*, ..., *5 stanze* e *5 o più stanze*.

- `car_parking`: Categoriale, no `NA`

Molte categorie hanno poche osservazioni. Riassumiamo quindi tutti i valori in tre classi: *Box*, *Posto Auto* e *No*. Se in un'abitazione sono presenti sia box che posto auto, essa viene inclusa nella prima categoria (in quanto più impattante sul prezzo totale).

- `availability`: Categoriale, con `NA`

Sono presenti numerose categorie sulla base della data in cui la casa sarà disponibile. Poichè non possiamo calcolare i mesi mancanti a tale data, costruiamo due categorie: *Disponibile Ora* e *Disponibile in Data Fissata*. I valori mancanti potrebbero essere relativi a nude proprietà, in cui non si sa quando l'immobile risulterà disponibile. Inseriamo quindi tutti gli `NA` in una terza categoria a parte.

- `condominium_fees`: Categoriale, con `NA`

Nonostante la natura numerica di questa variabile, vista la presenza di diversi `NA` e del valore *No Spese Condominiali*, trattiamo questa variabile come categorica con le seguenti classi, costruite a partire dal valore iniziale: *Da 1 a 100 €*, *Da 101 a 200 €*, *Da 201 a 400 €*, *Da 401 a 600 €*, *Più di 600 €*, *No Spese Condominiali*, *Valore mancante*.

- `year_of_construction`: Categoriale, con `NA`

Per un discorso del tutto analogo, trattiamo questa variabile come categoriale con le seguenti classi: *Valore mancante*, *Prima del 1900*, *Anni 1910*, *Anni 1920*, ..., *Anni 2010*, *Anni 2020*, *In costruzione*, dove nell'ultima classe inseriamo tutte le abitazioni con valore ≥ 2025 in questa variabile.

- `conditions`: Categoriale, con `NA`

Poichè non è possibile imputare la condizione di un'abitazione (*Nuova*, *Eccellente*, *Buona* o *Da Ristrutturare*) sulla base delle sue altre caratteristiche, lasciamo gli `NA` in una quinta classe separata.

- `zone`: Categoriale, con `NA`

Mi aspetto che questa sia la variabile più impattante (insieme a `square_meters`) sul prezzo finale. Per questo motivo e poichè il valore mancante è uno solo, tale osservazione viene esclusa dall'analisi.
Poichè il numero di zone è molto elevato (146), in una fase preliminare si era provato a rendere numerica questa variabile calcolando la distanza in km dal centroide di ogni zona al Duomo. I modelli costruiti in questo modo davano però risultati sotto tutti i punti di vista peggiori rispetto ai modelli in cui questa variabile veniva lasciata come categorica, quindi si è proceduto in quest'ultimo modo.

- `heating_centralized`: Categoriale, con `NA`

Come per `conditions`, lasciamo gli `NA` in una classe a parte. Le categorie sono quindi: *Centralizzato*, *Indipendente* e *NA*.

- `energy_efficiency_class`: Categoriale, con `NA`

Sono presenti le usuali categorie dalla *A* alla *G*. I valori in cui non è presente una lettera e i valori mancanti vengono inseriti in una classe (*NA*) a parte.

- `other_features`: Stringa, con `NA`

Estraiamo le caratteristiche più importanti, sia in base alla tipologia, sia in base alla presenza nel dataset. Creiamo quindi le seguenti variabili categoriali (e poi rimuoviamo `other_features`):

  - `Giardino`: *Privato*, *Condiviso*, *No*
  - `Strutture`: *Campo Tennis*, *Piscina*, *Idromassaggio*, *Niente*
  - `Servizi`: *Concierge Intera Giornata*, *Concierge Mezza Giornata*, *Reception*, *Niente*
  - `Arredamento`: *Arredato*, *Parzialmente arredato*, *Solo Cucina*, *Niente*
  - `Altre Stanze`: *Taverna*, *Terrazzo*, *Cantina*, *Solaio*, *Niente*
  - `Sicurezza`: *Allarme*, *Porta Blindata*, *Niente*
  - `Esposizione`: *Doppia*, *Interna*, *Esterna*, *Non Specificato*

1.2 Variabile risposta

La variabile risposta `selling_price`, la cui distribuzione è riportata in Figura 1, non contiene valori mancanti e i suoi valori sono sensati: vanno da un minimo di 25629€ ad un massimo di 2980000€, limiti ragionevoli per un mercato come quello di Milano.

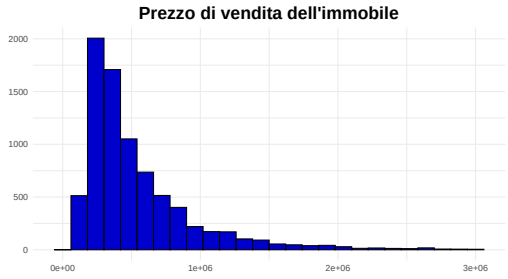

Figura 1: Istogramma di `selling_price`

Decidiamo di fare una trasformazione di questa variabile. Nei modelli che utilizzeremo in fase di analisi, l'effetto delle covariate sarà infatti di tipo additivo, ma risulta più facile credere che i m^2 di un'abitazione abbiano un effetto moltiplicativo sul prezzo finale. Per ovviare a questo problema, decidiamo quindi di modellare il prezzo al m^2 invece che il prezzo di vendita, ottenendo poi il prezzo totale semplicemente moltiplicando la stima ottenuta per `square_meters`.
Nonostante ciò, decidiamo di lasciare anche la variabile `square_meters` tra i regressori. Questa scelta è dettata dal fatto che, a parità di tutte le altre caratteristiche, ci aspettiamo che case di metratura maggiore abbiano un prezzo al m^2 leggermente inferiore rispetto a case più piccole. Poichè ci aspettiamo che questo effetto sia molto rilevante, inseriamo tra le covariate anche `square_meters`^{2}. In ogni caso, dovessero questi due previsori risultare irrilevanti, verranno identificati come tali nei modelli che costruiremo.

2 Analisi

2.1 Operazioni preliminari

Dividiamo i dati di `training.csv` in training (60%), validation (20%) e test (20%). Nella restante parte di questo report, con *test* faremo riferimento a questi dati, mentre ci riferiremo al test set finale su cui fare previsione come *nuove osservazioni*. Definiamo inoltre *fulltraining* l'insieme di training e validation.

2.2 Modelli Lineari

Per avere un benchmark, costruiamo innanzitutto un modello lineare basato solo su `zone` e `rooms_number`. Facciamo poi la stessa cosa ma considerando la variabile risposta `pricemq` in scala logaritmica (trasformazione che ci permette di ottenere previsioni sicuramente positive dopo averne preso l'*exp*). Con modelli di questo tipo allenati sul training set, otteniamo un MAE sul validation set di circa 99000.
Costruiamo poi dei modelli analoghi ma considerando tutte le variabili. L'aumento della capacità previsiva è evidente: si ottiene un MAE sul validation di 83109 per il modello in scala originale e di 80333 per quello in scala logaritmica. Vista la sempre migliore perfromance dei modelli con la risposta in scala logaritmica, da ora in poi tutti i modelli costruiti useranno `log(pricemq)` come variabile target. Tranne qualche eccezione, i coefficienti stimati da questo modello presentano il segno che ci si attende a livello logico:

Variabile	Coefficiente
square_meters	-0.003
zone = Duomo	1.041
floor = 9	0.065
energy_efficiency_class = G	-0.055

Tabella 1: Alcuni dei coefficienti stimati

[Nota:] Le classi baseline per le variabili categoriali qui presentate sono `zone = Affori`, `floor = 1` e `energy_efficiency_class = A`.

2.3 Forward e Backward Regression

Per ottenere risultati comparabili con i precedenti e per massimizzare la potenza previsiva, ad ogni step selezioniamo il miglior modello sulla base del MAE ottenuto sul validation set (invece che utilizzare indicatori come AIC o simili). La procedura (circa 60 secondi di calcolo) ha prodotto i seguenti risultati:

- Forward Regression: il modello con la capacità previsiva maggiore è risultato essere quello in cui si considerano tutte le covariate, ovvero il modello stimato sopra.
- Backward Regression: il modello che ha prodotto il minor MAE sul validation set (80151) è risultato essere quello che esclude `total_floors_in_building`, `car_parking`, `rooms_number` e `facilities`.

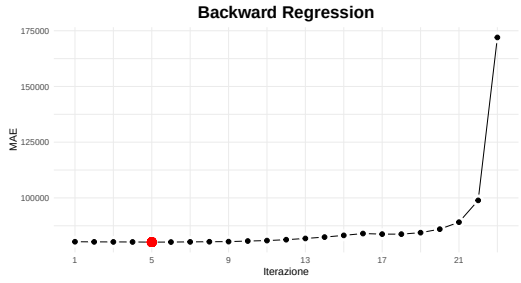

Figura 2: Percorso Backward Regression

2.4 Ridge, Lasso ed Elastic Net

Per cercare di ridurre la complessità del modello, sono state utilizzate queste tre tecniche. Sono stati dunque costruiti diversi modelli per diversi valori di λ sui dati di training ed è stata valutata la loro performance previsiva sul validation. In tutte e tre i casi, il miglior valore di λ è risultato essere il più piccolo presente nella griglia di valori proposta, ovvero un valore quasi nullo. Tutti questi modelli risultano quindi per costruzione estremamente simili ad OLS, in quanto la penalità inserita da ognuno di essi viene pesata ≈ 0.

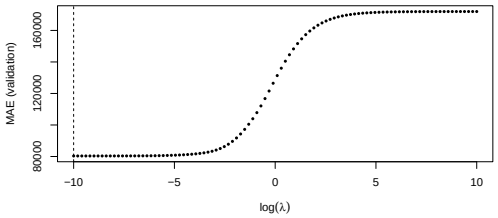

Figura 3: Performance per diversi valori di λ nella Ridge Regression

2.5 Test set

Il validation set è stato utilizzato per scegliere i valori di λ delle tre tecniche di shrinkage e i previsori da includere nella backward e forward regression (anche se da ora in poi non considereremo più quest'ultimo metodo in quanto coincidente con OLS).
Procediamo ora riallenando tutti questi modelli sul *fulltrain* set e valutiamo la loro capacità previsiva sul test set per scegliere quale utilizzare con le nuove osservazioni. I risultati ottenuti sono riportati in Tabella 2.

Scegliamo dunque di utilizzare il modello lineare in scala logaritmica per prevedere le nuove osservazioni. A questo punto risulta necessario fare un'importante osservazione. Nella procedura seguita in questa analisi, il dataset di validation è stato utilizzato per la selezione dei parametri dei diversi modelli, mentre il dataset di test per la scelta del modello finale. Quindi il valore MAE=79545 qui ottenuto non può essere considerato come l'errore che ci attendiamo di commettere sulle nuove osservazioni, in quanto è "influenzato" dal fatto che abbiamo usato questi valori per scegliere il miglior modello. Tuttavia, si è comunque deciso di procedere in questo modo in quanto risulta la procedura più conveniente per usare il maggior numero di dati in tutte le fasi dell'analisi ed ottenere quindi il miglior risultato possibile sulle nuove osservazioni.

Metodo	MAE
Modello lineare	80249
Modello lineare, scala log	79545
Backward regression	79967
Ridge	79591
Lasso	79545
Elastic net	79604

Tabella 2: MAE dei diversi modelli sul test set

3 Nuove osservazioni

Importiamo i nuovi dati ed eseguiamo le stesse operazioni di preprocessing effettuate sul dataset iniziale, con le seguenti differenze:

- Poichè non possiamo escludere alcuna osservazione, imputiamo i valori mancanti di `square_meters` sulla base dei m^2 medi delle osservazioni presenti nei dati originali con lo stesso numero di stanze e bagni.

- In questi dati sono presenti zone che non erano presenti nei dati originali. Sostituiamo questi valori con una zona ad essa vicina.

Alleniamo ora il modello scelto su tutti i dati presenti nel dataset originale per avere la migliore accuratezza possibile e facciamo previsione sulle nuove osservazioni. Le previsioni proposte hanno ottenuto un MAE pari a 73169.6 sul 38% di tale osservazioni, come riportato nella classifica pubblica di Kaggle.