PREVISIONE: PREZZO DELLE CASE A MILANO
Data Mining 2024/2025

Aurora Musitelli
Matricola: 856741

1 Introduzione

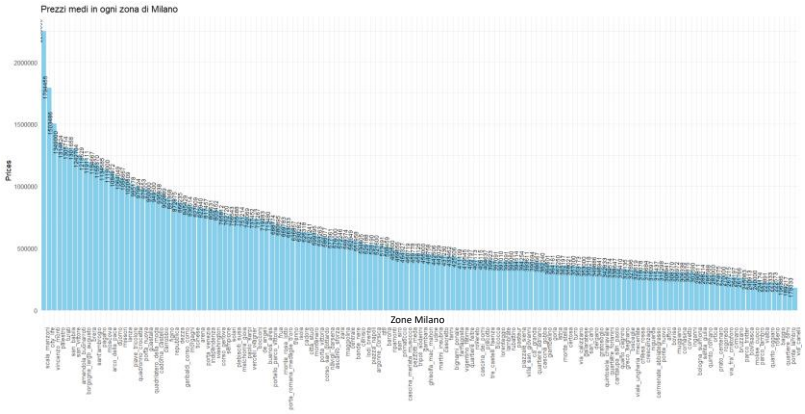
L'obiettivo dell'analisi consiste nel prevedere il prezzo di vendita delle case localizzate nella città di Milano applicando modelli statistici per osservare il modello che minimizza il valore MAE (mean absolute error). Il lavoro si è concentrato sull'analisi esplorativa e nel feature engineering, si riportano quindi in questa relazione le osservazioni, le idee e i risultati ottenuti durante lo studio del dataset in questione. Il dataset in esame complessivamente comprende 12800 osservazioni e 16 variabili, di cui 10 variabili qualitative e 6 variabili quantitative, tra cui la variabile target selling_price il prezzo di vendita dell'immobile. Come prima cosa è stata svolta un'attenta fase di pre-processing per identificare la natura delle variabili in esame e la presenza di eventuali anomalie all'interno dei dati. Dopo aver unito i dataset di training e di test al fine di effettuare un pre-processing comune, si è svolta un'analisi sulle caratteristiche delle variabili qualitative e sulle variabili quantitative.

2 Analisi Esplorativa

Analisi delle variabili qualitative

Le variabili categoriali sono state codificate per renderle idonee all'utilizzo nei modelli statistici

Lift	presenza ascensore nell'immobile, codificata come binaria (1 = presente, 0 = assente)
Floor	piano in cui si trova l'immobile. Le categorie come "mezzanine", "semi-basement" e "ground floor" sono state convertite in valori numerici
Floors_build	numero totale di piani dell'edificio, trattata come variabile numerica, "1 floor" codificato come valore numerico 1
Car_parking	informazioni sulla presenza di parcheggi, codificato come "box" e "parcheggio condiviso", successivamente codificate come variabili dummy
Availability	se la casa è già disponibile o, in caso contrario, quando sarà disponibile codificato come valore binario (1 = disponibile, 0 = non disponibile)
Conditions	condizioni categoriche della casa "excellent / refurbished", "good condition / liveable", "new / under construction", "to be refurbished" codificate come valori numerici in scala da 1 a 5 e successivamente codificate come variabili dummy escluso il livello baseline
Zone	zona di Milano in cui è situato l'immobile. Sono state mantenute solo le zone con almeno 20 osservazioni, mentre le restanti sono state raggruppate nella categoria "altre" per ridurre il rischio di over fitting causato da categorie scarsamente rappresentate, preservando al contempo l'informazione contenuta nelle zone più frequenti
Other_features	elenco di caratteristiche aggiuntive dell'immobile. Sono state selezionate le 10 più ricorrenti e trasformate in variabili dummy
Heating_centralized	presenza di riscaldamento centralizzato codificata come (1 = centralizzato, 0 = indipendente)
Energy_class	classe di efficienza energetica della casa codificato come valore numerico su una scala da 1 a 7 e successivamente codificate come variabili dummy escluso il livello baseline



Guardando da più vicino la distribuzione dei prezzi medi degli immobili nelle zone di Milano, si osserva una forte variabilità che riflette le differenze socio economiche tra i vari quartieri di Milano. Questa variabilità influisce sulla variabile target selling_price, motivo per cui si è deciso di codificare la variabile "zone" mediante variabili dummy. Includere le zone come variabili dummy consente di catturare l'effetto specifico di ciascuna area sul prezzo di vendita, migliorando l'accuratezza predittiva e riducendo il rischio di distorsioni. Inoltre, la scelta dell'aggregazione delle zone con poche osservazioni nella categoria "altre" contribuisce a minimizzare over fitting poiché sono presenti 148 zone totali.

Analisi delle variabili quantitative

Di seguito si descrive il trattamento delle principali variabili identificate come quantitative

Square_meters	superficie dell'immobile in metri quadrati. È stata mantenuta come variabile continua, poiché rappresenta una misura fondamentale direttamente proporzionale al prezzo di vendita
Bathrooms_number	numero totale di bagni nella casa, codificato il valore 3+ come 4 bagni
Rooms_number	numero totale di stanze della casa, codificato il valore 5+ come 6 camere
Condominium_fees	importo mensile delle spese condominiali. È stata applicata una trasformazione logaritmica per ridurne l'asimmetria e l'effetto degli outlier
Year construction → Age	è stata eliminata la variabile year of construction e creata la variabile age (identificativa dell'età dell'abitazione) più informativa e facilmente interpretabile. Anche qui è stata applicata una trasformazione logaritmica per migliorare la distribuzione

Outlier detection e valori mancanti

Durante la fase di pre-processing, è stata condotta un'analisi per gestire in modo appropriato i valori mancanti e gli outlier, al fine di garantire la qualità del dataset e migliorare le prestazioni dei modelli predittivi. Per quanto riguarda il controllo dei valori mancanti è stato svolto su tutte le variabili (esclusa la variabile target selling_price). I valori NA per alcune variabili sono stati gestiti come non presenza quindi imputati come valore 0, questa scelta è giustificata dal fatto che la mancanza di informazione su questi attributi può essere interpretata come assenza di informazione nel contesto del mercato immobiliare. Per le variabili square_meters e condominium_fees i valori mancanti sono stati imputati con la mediana. Le variabili costanti (senza varianza) sono state rimosse.

Invece per i valori outlier sono stati individuate le seguenti anomalie:

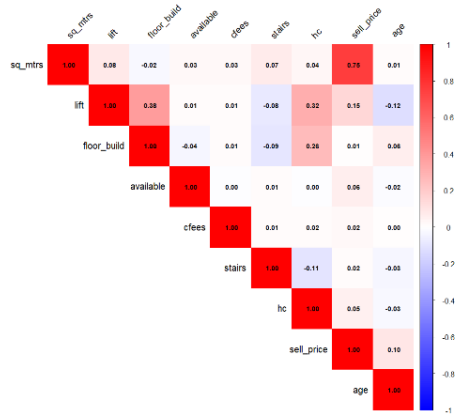
→ Sono state considerate anomale le abitazioni con superficie inferiore a 28 m², in quanto tale dimensione è statisticamente rara e, nella realtà, corrisponde spesso a immobili atipici

→ Condominium_fees sono stati rilevati valori estremamente elevati rispetto alla media, potenzialmente dovuti a casi eccezionali (es. immobili di lusso).

→ Age (età dell'immobile), alcuni immobili risultavano eccessivamente vecchi, generando una forte dispersione nella distribuzione.

Per gestire questi casi, è stata adottata una trasformazione logaritmica su condominium_fees e age, al fine di: stabilizzare la varianza (mitigando problemi di eteroschedasticità), linearizzare la relazione con la variabile target selling_price, facilitando la modellizzazione e ridurre l'impatto degli outlier, comprimerne l'influenza e migliorare la robustezza dei modelli. Queste trasformazioni permettono di preservare l'informazione contenuta nei dati estremi senza doverli necessariamente escludere, migliorando al contempo l'efficienza predittiva del modello.

Correlogramma variabili quantitative

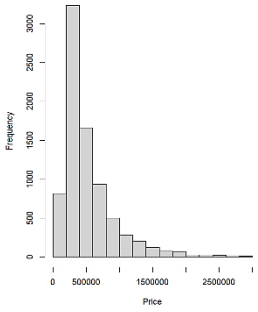


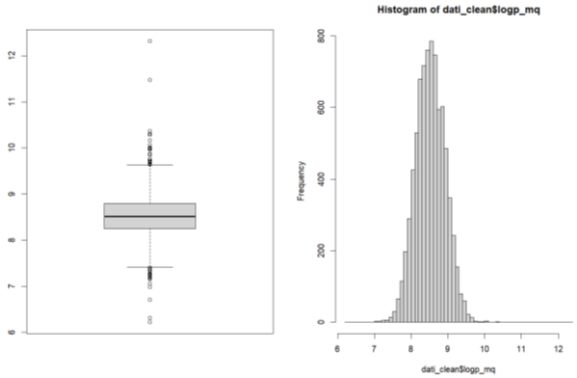
Si è analizzata la relazione tra la variabile target selling_price e le variabili esplicative tramite l'analisi del correlogramma. Dall'analisi è emerso che: la variabile square_meters presenta una forte correlazione positiva con selling_price, confermando l'intuizione che la superficie abitabile è uno dei principali determinanti del prezzo di un immobile. Le altre variabili quantitative mostrano correlazioni deboli o moderate con il prezzo, ma possono comunque apportare valore in un'ottica di previsione del prezzo delle case. Non sono emersi segnali critici di collinearità tra le variabili esplicative. È stata inoltre posta attenzione nella creazione delle variabili dummy, escludendo il livello di riferimento (baseline) per ciascuna categoria, evitando così problemi di multicollinearità.

Analisi della variabile target

La variabile target selling_price rappresenta il prezzo di vendita finale della casa in euro. La distribuzione della variabile dipendente è asimmetrica a destra (asimmetria positiva) con:

- valore minimo: €25629
- media: €540568
- valore massimo: €2980000





Questa asimmetria è tipica nei dati immobiliari, dove immobili di lusso possono influenzare fortemente la media e la varianza. Quindi, per migliorare la qualità della modellazione, si è deciso di creare la variabile $\log\left(\frac{selling\_price}{\text{square\_meters}}\right)$ utilizzandola come variabile target. Questa trasformazione è stata motivata da diverse considerazioni pratiche: stabilizza la varianza poiché riduce l'eteroschedasticità, ovvero la variabilità non costante del prezzo rispetto alla dimensione dell'immobile. Consente di modellare in maniera più lineare una relazione che, in scala originale, è tipicamente moltiplicativa (ad esempio: raddoppio dei metri quadrati, raddoppio del prezzo). Riduce l'impatto degli outlier: i valori estremi elevati vengono compressi, migliorando la robustezza del modello e per ultima interpretabilità economica poiché il rapporto prezzo/metri^2 è una metrica nota nel settore immobiliare e facilita la lettura dei risultati in chiave pratica. Questa scelta ha permesso di normalizzare la distribuzione della variabile target e di renderla più adatta all'applicazione di modelli statistici, migliorando al contempo la capacità predittiva e la stabilità del modello finale.

3 Modellistica

Già dall'analisi del correlogramma si evince che l'applicazione di modelli come le componenti principali e la ridge regression potrebbero non essere ottimali, in quanto questi risultano particolarmente performanti in presenza di variabili altamente correlate tra loro, situazione che nel nostro dataset non si è rilevata, questo risultato è stato confermato anche in fase di stima. Si è quindi deciso di modellare la variabile risposta come $\log\left(\frac{selling\_price}{\text{square\_meters}}\right)$: in primo luogo, è stato stimato un modello lineare completo (OLS) e successivamente, tra i modelli implementati, quelli che hanno prodotto le migliori performance predittive sul validation set in termini di MAE sono:

1.Lasso Regression modello di regressione penalizzata L1, che consente la selezione automatica delle variabili riducendo così over fitting. È stato selezionato il valore ottimale di penalizzazione tramite cross-validation (lambda.min), il modello risultante ha azzerato alcuni coefficienti, identificando solo le variabili rilevanti.

2.XGBoost modello basato su gradient boosting di alberi decisionali, adatto a catturare interazioni complesse tra le variabili. È stato effettuato un tuning approfondito dei parametri tramite cross-validation (5-fold) per determinare il numero ottimale di iterazioni (nrounds).

3.Forward Stepwise Regression è stata implementata una regressione stepwise forward con pesi inversamente proporzionali ai metri quadrati, per dare più peso agli immobili più piccoli. Creato il modello nullo e aggiunto progressivamente le variabili più significative secondo il criterio BIC (k = log(n)). La formula finale è risultata parsimoniosa, mantenendo solo le covariate statisticamente significative. Questo approccio ha consentito di migliorare le stime rispetto a un OLS semplice.

4.Iteratively Weighted Least Squares (IWLS) a partire da una selezione inziale delle variabili tramite stepwise su un modello OLS, è stato implementato un algoritmo iterativo di regressione pesata in cui si è stimata la varianza dei residui, applicato uno smoothing tramite LOESS su tali varianze rispetto ai fitted values. Si sono costruiti dei pesi inversi alla varianza stimata ed infine si è riadattato il modello pesato con tali pesi.

Il confronto tra i modelli è stato condotto utilizzando le metriche principali sul validation set:

- MAE sul prezzo totale stimato: exp(predizione $\log\left(\frac{selling\_price}{\text{square\_meters}}\right)$)*square_meters
- MAE su $\log\left(\frac{selling\_price}{\text{square\_meters}}\right)$

Modelli fittati	MAE validation €
IWLS	87651.48
Lasso	87689.99
XGBoost	87998.34
Forward Stepwise Regression	97641.48

4 Modello finale

Il modello migliore ottenuto al fine di prevedere il prezzo delle case a Milano risulta essere **Iteratively Weighted Least Squares (IWLS)**. Questo modello ha fornito la miglior combinazione di accuratezza e interpretabilità, garantendo previsioni più affidabili. Le previsioni finali, svolte sul test set, sono state ottenute modellando il logaritmo del prezzo al metro quadro e successivamente riportate alla scala originale tramite trasformazione esponenziale e moltiplicazione per la superficie dell'immobile. Questo metodo ha permesso di mantenere la coerenza con l'unità di misura originale del prezzo totale e ha contribuito a migliorare la qualità delle stime.