Tecniche di Data Mining per la previsione dei prezzi delle case nella città metropolitana di Milano

Tommaso Pozzi matricola 864654

19 Giugno 2025

1 Introduzione

In questa analisi l'obiettivo principale è sfruttare tecniche di data mining con lo scopo di prevedere il prezzo delle case nella città di Milano. Il data set su cui vengono addestrati i modelli è disponibile su Kaggle ed è composto da 8000 osservazioni e 16 variabili. Il report è costituito principalmente da 4 sezioni: una prima, materiali, dove vengono presentati i dati e tutte le operazioni di pre-processing effettuate, una seconda dove vengono descritti i modelli e le scelte effettuate durante l'analisi ed un'ultima contenente i risultati e le conclusioni finali.

2 Materiali

La variabile target è il prezzo di vendita di un immobile nella città di Milano caratterizzato nel data set dall'etichetta *Selling Price*. Le covariate includono: *square meters* (superficie in mq), *bathrooms number* e *rooms number* (numero di bagni e stanze), *floor* e *total floors in building* (piano e numero totale di piani), *lift* (presenza dell'ascensore), *conditions* (quattro categorie: *excellent/refurbished*, *good/liveable*, *new/under construction*, *to be refurbished*), *car parking* (numero e tipologia di posto auto), *condominium fees* (spese condominiali), *zone* (zona di Milano), *heating centralized* (presenza di riscaldamento centralizzato), *energy efficiency class* (classe energetica), *year of construction* (anno di costruzione) e *other features* (caratteristiche aggiuntive dell'immobile). Attraverso una prematura ispezione dei dati, si è osservato come siano presenti ben 3955 dati mancanti suddivisi tra le diverse variabili, esclusa la variabile risposta. Prima di procedere con l'imputazione si sono estratte le caratteristiche aggiuntive dalla variabile *other features*. Per ognuno di questi attributi si è creata una nuova variabile dummy con valore 1 indicante la loro presenza. Dopo questo processo, si è passati all'imputazione dei valori mancanti. In primis, si è osservata la variabile *bathrooms number* contente 25 valori mancanti. Per l'imputazione di questa variabile si è considerata la sua associazione con la variabile *room number* osservando una forte corrispondenza tra i rispettivi valori. Si è deciso di assegnare valore 1, quando il numero delle stanze totali era inferiore a 3, 2 quando era inferiore a 5, mentre in presenza di 5 o più stanze un valore pari a 3 bagni. Tra le caratteristiche aggiuntive si è osservato che alcuni condomini possedevano concierge e reception. Queste caratteristiche sono state unite in un'unica variabile dicotomica. Per quanto concerne, invece, la variabile *floor*, si è osservata la presenza di valori come *mezzanine*, *ground floor* e *semi basement*. Con lo scopo di rendere la variabile numerica si sono assegnati rispettivamente i valori 0.5, 0 e -1. Successivamente questa variabile è stata anche categorizzata in 5 livelli: *basement, low floor, medium floor, high floor* e *penthouse* attraverso la creazione di una nuova variabile. L'esposizione, altra caratteristica presente in *other feature* è stata raggruppata in un' unica variabile. Infatti, nelle altre caratteristiche erano indicate le direzioni a cui era esposta la casa. Si è deciso di assegnare valore pari a 1 quando l'esposizione era presente in una sola direzione, 2, 3 e 4 negli altri casi. Quando, invece, non era presente questa informazione si è creata una nuova classe apposita denominata *Info not available*. Inoltre, in molte osservazioni era presente l'informazione generica *double exposure*. A queste osservazioni è stata assegnata la classe coerente 2. L'informazione relativa alla composizione delle finestre, ottenuta attraverso la dicotomizzazione della variabile *other feature*, si basa principalmente su due aspetti: lo spessore del vetro (singolo, doppio o triplo) e il materiale del telaio (legno, PVC o metallo). A partire da queste informazioni, sono state create due variabili distinte: una per lo spessore del vetro e una per il materiale del telaio. Anche in questo contesto, nei casi in cui tali caratteristiche non fossero disponibili all'interno della variabile originale, è stato assegnato il valore *info not available*. Le informazioni relative alla caratteristica dell'arredamento, *furnished*, sono state inglobate in un unica variabile a 4 livelli indicando se la casa fosse arredata, arredata parzialmente, solo la cucina oppure spoglia. La variabile *car parking* era composta principalmente da caratteri che indicavano il numero di posti auto disponibili e se questi fossero in un parcheggio condiviso o in un garage. Dato che il numero di parcheggi era specifico alla singola osservazione e troppo variabile si è deciso di escludere questa informazione e mantenere soltanto l'informazione riguardante la tipologia di parcheggio, ottenendo una variabile a quattro livelli: no parking, shared parking, garage e sia condiviso che garage. La variabile *lift* presentava 41 valori mancanti. Si è deciso di imputare la presenza dell'ascensore in tutte le osservazioni il cui numero totale di piani dell'immobile era superiore a 3, assenza quando inferiore. La variabile *condominium fees* è stata resa numerica sostituendo 0 quando il valore associato era *No condomiunium fees*. Anche all'interno della variabile *total floor* erano presenti 74 valori mancanti. L'imputazione di questa variabile è risultata più complessa perchè non esiste una regola euristica per la loro definizione. Inoltre, la maggior parte delle case considerate nell'analisi erano palazzi con un numero di piani totali davvero elevati e, l'imputazione attraverso modelli, ad esempio alberi di classificazione, produceva previsioni composte solo da piani elevati. Si è optato, quindi, per imputare attraverso una moda condizionata. In particolare, ci si è condizionati alla variabile *floor*. Data l'incertezza presente in questo metodo si è deciso di creare una variabile dicotomica che indicasse

dove fosse avvenuta l'imputazione. In questo modo si è anche in grado di individuare la presenza di caratteristiche latenti dovute alla presenza dei valori mancanti, che spesso vengono tralasciate. Inoltre, la dummy potrebbe essere in grado di individuare bias introdotti dall'imputazione e valutare se le informazioni imputate influenzano i risultati del modello predittivo. I valori mancanti delle variabili *conditions* e *energy efficiency class* sono stati sostituiti con un nuovo livello indicato come *Info not available*. Per quanto riguarda il riscaldamento, si è considerato di imputare come indipendente quando il numero totale di piani dell'edificio era pari a 1, quando vi era un camino, quando vi era una piscina o un campo da tennis e quando le tasse condominiali erano pari a 0, mentre centralizzato altrimenti. Questa scelta è dettata dalla logica che case indipendenti o con attrazioni da casa di lusso, tipicamente predispongono per una struttura a riscaldamento autonomo. Inoltre, difficilmente il costo del riscaldamento è gratuito, quindi risulta inverosimile avere case con riscaldamento centralizzato e zero spese condominiali. Queste ultime, sono state invece imputate attraverso un modello di regressione lineare con covariate *concierge, tennis court, pool, electric gate, video entryphone, heating centralized, energy efficiency class, lift e condomio*. *Condominio* è una variabile creata con l'intento di verificare se l'immobile fosse un condominio o meno. Essa è una dummy che assume valori pari a 1 quando le spese condominiali sono nulle e la casa non è costituita da più di 2 piani. Anche in questo contesto si è creata una variabile dicotomica rappresentante l'imputazione. Infine, l'ultima variabile da imputare è quella relativa all'anno di costruzione. Questa variabile era problematica perchè presentava ben 789 valori mancanti. Inoltre, utilizzare un modello per l'imputazione di una data è complesso. Si è deciso di optare per un'imputazione attraverso la mediana, pari a 1960. Data l'incertezza di questo metodo imputativo, si è creato, anche in questo caso, una variabile dicotomica rappresentante l'avvenimento dell'imputazione. In conclusione, si è osservato come alcune osservazioni presentavano valori anomali, in particolare per le variabili *square meters* e *condominium fees*. Per quanto riguarda la prima, erano presenti case con valori di superficie inferiori ai 20 metri quadrati. Per la normativa italiana (Decreto Salva Casa), la metratura minima per un monolocale deve essere pari a 20 metri quadrati. Si sono rimosse queste osservazioni dal data set. Per la seconda variabile, invece, si sono osservati valori di *condominium fees* superiori a 12000 euro. Si sono considerati, osservando i valori delle altre variabili, errori di battitura e si è pertanto diviso per 100. Fortunatamente questi errori erano presenti in quantità ristrette e queste modifiche non hanno alterato in modo decisivo l'analisi finale. La variabile *zone* presenta una totale di 146 zone, tuttavia, molte di queste erano presenti in frequenze piuttosto ridotte. Si è deciso di unire le zone con una frequenza assoluta inferiore a 10 con quella più vicina geograficamente. In conclusione al metodo di pre processing applicato a questo data set, è stato eseguito un principio di *feature engeneering*. Si sono ottenute le coordinate relative a ciascuna zona e queste sono state utili per il calcolo della distanza in metri dal duomo di Milano. Inoltre, sono state create due variabili: *luxury* e *optional*. La prima, conteggia il numero di caratteristiche di lusso che presentava una determinata osservazione. Si sono considerate come caratteristiche di lusso la piscina, il campo da tennis, la concierge e l'idromassaggio. Si sono considerate come optional, invece, la porta di sicurezza, il cancello elettrico, il camino, la fibra ottica, il citofono video e il sistema tv centralizzato con satellite.

3 Metodi

Attraverso un'approfondita analisi esplorativa si sono osservate, nel caso continuo, le variabili maggiormente correlate con la variabile risposta per l'implementazione dei diversi modelli. Si è scelto di non utilizzare tecniche come la best subset regression a causa dell'elevato numero di variabili presenti all'interno del data set. Si sono testati approcci di shrinkage e selezione delle variabili attraverso regressioni di tipo lasso ed elastic net ($\alpha = 0.5$), tuttavia il parametro di penalizzazione λ ottenuto attraverso cross-validation ha portato, in entrambi i casi, a valori talmente piccoli da far collassare le stime ai minimi quadrati. Un ulteriore approccio si è provato attraverso la regressione *group lasso*, non ottenendo tuttavia significative riduzioni della dimensionalità (anche in questo caso il parametro di complessità λ è stato stimato attraverso cross-validation). Si è optato per un approccio manuale alla selezione delle variabili svolto attraverso un'approfondita analisi esplorativa. In particolare, si è osservato come variabili come *square meters* (correlazione con la risposta 0.7567) fossero di fondamentale importanza per il modello predittivo. Le altre variabili selezionate, dove nel caso di variabili dicotomiche venivano osservati i valori mediani del prezzo per ogni modalità verificando se vi era una significativa differenza, sono: *condominium fees, floor, luxury, alarm system, optional, pool, conditions, lift, rooms number, terrace, energy efficiency class, cellar, Condominio, year of construction, is na year, is na condominium fees, zone, distanza dal duomo, furnished, floor category, private garden, bathrooms number*. Oltre a ciò, si sono considerate le interazioni tra le variabili *bathrooms number* e *square meters*, questo per il semplice fatto che il numero di bagni è altamente correlato con l'ampiezza della casa. A parità di superficie, una casa che presenta un numero più elevato di bagni può risultare più comoda e di conseguenza acquisire un valore maggiore. Una seconda interazione considerata è quella tra *floor* e la variabile *luxury*. Osservazioni a piani elevati con un numero elevato di caratteristiche di lusso si comportano in maniera differente. Nei contesti non di lusso, i piani alti potrebbero essere meno desiderabili se l'edificio non ha ascensore o se aumenta la fatica nell'accesso. Al contrario, nei contesti lussuosi, i piani alti possono essere visti come un maggior privilegio data la scarsa rumorosità e l'elevato livello di privacy. Infine, l'ultima interazione considerata riguarda la variabile creata durante il pre-processing indicante la distanza dal Duomo e la variabile luxury. Tale interazione è stata introdotta per modellare in modo più flessibile il rapporto tra centralità e prezzo in funzione delle caratteristiche qualitative dell'immobile. Questo principalmente per contrastare l'effetto, importante, dato dalla variabile *zone* sul prezzo delle

case. In generale, la distanza dal Duomo è una variabile fortemente correlata al prezzo, in quanto sintetizza la vicinanza al centro storico e alle zone di maggiore prestigio come Duomo e Brera. Tuttavia, l' effetto della distanza può non essere uniforme per immobili di lusso il cui valore di mercato rimane elevato anche in zone più periferiche. Inoltre, l' introduzione di questa interazione contribuisce a ridurre la dipendenza esclusiva dalla variabile *zone*, che potrebbe assorbire effetti strutturali rilevanti ma difficili da interpretare. Sempre per quanto riguarda la variabile *zone*, in Figura 1 è rappresentata la distribuzione del prezzo, in scala logaritmica, al variare delle zone, ordinata per prezzo mediano crescente.

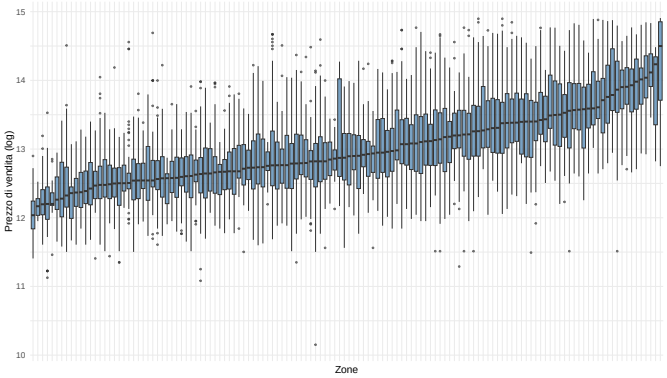

Figura 1: Distribuzione del prezzo degli immobili in scala logaritmica rispetto alle zone

Come ben visibile il numero delle zone è elevato, 132. Tuttavia, ciò che è fortemente interpretabile riguarda il forte impatto delle zone sul prezzo. Dato il numero elevato di livelli di questa variabile, e il ristretto numero di frequenze per alcuni livelli, in questo lavoro si sono provati principalmente due approcci. Un primo, mantenendo la variabile *zone* come un effetto fisso, mentre un secondo come un effetto casuale. Importante evidenziare che, quando *zone* viene utilizzata come intercetta aleatoria si è anche interessati ad una sua stima da includere nel modello. Per fare ciò è stato necessario utilizzare una codifica a somma al posto della classica codifica a riferimento. Quando si lavora con effetti misti con lo scopo di effettuare previsione, come in questo caso, non vi è necessaria una categoria di riferimento, tuttavia è fondamentale che la somma degli effetti sia pari a zero. In questo modo, imponendo una distribuzione casuale (gaussiana) sulle intercette è come se si stesse eseguendo un'operazione di shrinkage dei coefficienti andando a penalizzare zone con un numero inferiore di osservazioni e una variabilità elevata della risposta al loro interno. Questa tipologia di shrinkage può essere vista come una penalizzazione ridge, dove tuttavia, il parametro λ è adattivo e non fisso, e viene calcolato empiricamente considerando la varianza all'interno di ciascuna zona e tra le diverse zone. L'obiettivo dell'introduzione degli effetti fissi è quindi quello di penalizzare l'importanza delle zone sulle previsioni finali. I modelli testati sono principalmente 6: modello lineare con *zone* considerato come effetto fisso, modello lineare con *zone* considerato come intercetta aleatoria, un modello lineare dove le variabili continue sono modellate attraverso spline cubiche naturali, due modelli additivi dove queste sono state modellate attraverso thin plate regression splines, corrispondenti a spline penalizzate dove i parametri λ di penalizzazione sono stati stimati attraverso il metodo REML. La variabile risposta, essendo strettamente positiva, è stata trasformata in scala logaritmica così come la variabile *square meters*. Per quanto riguarda *condominium fees* si è considerato come variabile esplicativa il rapporto tra questa e *square meters*, in scala logaritmica.

4 Risultati

Per valutare l'accuratezza del modello si è diviso il training set in un training ristretto (composto dal 70% delle osservazioni) e in un validation set (composto dal restante 30%). Le metriche utilizzate per il confronto tra modelli sono state calcolate sulle previsioni effettuate sul set di validazione. In Tabella 1 sono rappresentati i risultati dei 6 modelli confrontati in termini di MAE.

Modello	MAE
Lineare	81334.31
Lineare con B-Spline (degree = 3)	79791.95
Lineare con zone random	81430.15
Lineare con zone random e B-spline(degree = 3)	79871.46
GAM con zone fisso	78183.12
GAM con zone random	78207.45

Tabella 1: Risultati in termini di MAE dei modelli stimati sul validation set

Come ben visibile, l'utilizzo di un'intercetta aleatoria non ha particolari effetti predittivi in termini di MAE. Tuttavia si è deciso comunque di optare, per effettuare previsione sul test set, per il modello GAM con zone

come intercetta aleatoria. Questo, per i motivi descritti precedentemente riguardanti lo shrinkage dei coefficienti. Effettuando questa operazione si introduce bias controllato nella stime, tuttavia il modello dovrebbe essere in grado di generalizzare in modo migliore. Inoltre, il numero di gradi di libertà effettivi associato alla variabile *zone*, è risultato pari a 125, a differenza dei 131 del modello con effetti fissi. Questo risultato porta ad un modello più parsimonioso rendendolo più efficiente e meno propenso all'overfitting. Inoltre, in un contesto a dati sbilanciati come questo, gli effetti misti sono in grado di gestire al meglio la loro variabilità e i coefficienti di zone con un numero di frequenze ridotto vengono attirati verso la media generale. I risultati del modello mostrano, inoltre, come l'ipotesi della relazione non lineare tra *square meters* e il prezzo sia corretta (edf $= 5.986$). Infatti, il logaritmo del prezzo cresce in modo lineare fino a 244 metri quadrati circa, per poi crescere in maniera più lenta. Anche le spline attribuite alle variabili *condominium fees / square meters* (edf $= 6.422$)) e *year of construction* (edf $= 7.492$) sono significative nello spiegare la relazione non lineare con la variabile risposta. Interessante osservare soprattutto il comportamento di quest'ultima. Infatti è parecchio oscillatorio, ma descrescente. A case più antiche viene attribuito un prezzo medio più elevato. Nonostante ciò vi sono periodi storici dove gli immobili riprendono valore fermando il trend decrescente. Un esempio è il periodo tra il 1500 ed il 1750. Evidente nello studio dell'andamento della spline è inoltre come, oltre al fatto che nel campione siano presenti più case costruite recentemente, il prezzo medio negli ultimi anni sembra avere una forte risalita. Per quanto riguarda, invece, le spline stimate per l'interazione tra *floor* e *luxury* si è osservata una forte significatività della relazione per un numero di caratteristiche di lusso basso $(0 - 1 - 2)$, mentre nessuna significatività per valori elevati. Lo stesso accade nel caso delle spline stimate per la variabile riguardante la distanza dal duomo. Questo risultato può essere in parte spiegato dalla effettiva scarsa frequenza di case che possiedono un elevato numero di caratteristiche di lusso e per questo motivo le stime attribuite non risultano significative. Risultati interessanti riguardano principalmente le covariate dicotomiche. Infatti, per la variabile *alarm system* viene stimato un coefficiente di 0.03321, ciò significa che, a parità di tutte le altre variabili, il prezzo medio di una casa a Milano aumenta del 3.37% se questa ha installato un sistema di antifurto. Interessante osservare inoltre, come, la presenza della piscina in un immobile a Milano aumenti il suo valore di ben 10.84% ($\beta = 0.10295$). Il semplice passaggio, invece, da una casa in condizioni eccellenti ad una in buona condizioni comporta una riduzione del 10.12% ($\beta = --0.10669$) del prezzo medio, riduzioni inferiori si riscontrano invece quando questa informazione non è disponibile ($4.69\%, \beta = -0.04807$). Quando la casa è nuova vi è un incremento sul prezzo medio del 4.97% ($\beta = 0.04856$). Le due variabili dicotomiche relative all'imputazione delle variabili *year of construction* e *condominium fees* sono risultate entrambe significative. Una casa arredata, anche solo parzialmente, comporta un aumento del prezzo di vendita rispetto ad una casa spoglia. In particolare, una casa con arredata soltanto la cucina ha un aumento percentuale sul prezzo di vendita del 6.35% ($\beta = 0.06165$), superiore a quando la casa è arredata completamente ($2.49\%, \beta = 0.02464$)) o parzialmente ($2.6\%, \beta = 0.025673$). Tutte le interpretazioni dei coefficienti stimati sono valide soltanto a parità delle altre covariate e riguardano una variazione sul prezzo medio degli immobili.

Di rilevanza sono i coefficienti stimati riguardo alla variabile *zone*. L'interpretazione in questo contesto è leggermente diversa dovuta al cambio di definizione delle diverse variabili dicotomiche inserite all'interno del modello. Infatti, il coefficiente β stimato per le variabili *zone* non rappresenta la variazione, dopo l'opportuna trasformazione $((\exp(\beta) - 1) \times 100\%)$, rispetto ad un valore di baseline, bensì rappresentano la deviazione rispetto alla media delle zone, ossia il prezzo medio di una casa di Milano in scala logaritmica quando tutte le altre covariate sono 0 o corrispondono al loro livello base poichè è stata utilizzata una codifica a somma. Pertanto il valore di ciascun coefficiente β_j attribuito a *zone* quantifica la variazione percentuale attesa del prezzo di un immobile situato nella zona j rispetto al prezzo medio stimato su tutte le zone, a parità delle altre covariate. L'intercetta stimata è pari a 11.77322 e corrisponde ad un prezzo medio delle case di Milano di circa 129731 euro. Si osserva come l'acquisto di un immobile nel quartiere Wagner / Corso Vercelli comporti un incremento del prezzo medio del 209.26% ($\beta = 1.12902$), a parità delle altre covariate. Questa zona si contraddistingue come quella che apporta il maggior incremento nella stima dei valori della case. Un ulteriore zona che presenta questa caratteristica è Turati, comportando un incremento del 73.58% ($\beta = 0.55149$). Zone come Roserio o Ponte nuovo comportano invece un significativo decremento del prezzo medio, rispettivamente del 62.87% ($\beta = -0.99086$) e 48.55% ($\beta = -0.66467$). Questi valori sono coerenti con la geografia della città essendo, Wagner e Turati zone centrali, mentre Roserio e Ponte Nuovo quartieri periferici. Tra gli effetti stimati con maggior impatto sul prezzo medio vi sono anche le zone di Duomo, City life, Brera e Lanza, Pagano, Vincenzo Monti, Palestro, Moscova e zone universitarie come Bocconi, Bicocca, Città studi e San Vittore (Università Cattolica).Questo conferma ulteriormente la rilevanza della posizione nella determinazione del prezzo degli immobili. L'analisi dei residui del modello evidenzia come esso tenda a sottostimare il prezzo di alcuni immobili particolarmente costosi che, pur non essendo collocati in zone centrali o di pregio, non presentano un elevato numero di caratteristiche di lusso. Queste situazioni potrebbero indicare case atipiche, che sfuggono ai pattern appresi dal modello. In conclusione, i modelli stimati si sono dimostrati efficaci nella previsione del prezzo degli immobili a Milano, mantenendo un buon compromesso tra accuratezza predittiva e interpretabilità, anche con l'impiego di tecniche più flessibili come i modelli additivi. L'utilizzo dell'intercetta aleatoria ha consentito di mantenere interpretabili i coefficienti associati alla variabile zone, sfruttando la variabilità tra e all'interno dei gruppi per effettuare shrinkage sui coefficienti. L'interpretazione dei coefficienti risulta quindi funzionale per identificare le caratteristiche degli immobili che influenzano maggiormente il prezzo, mentre l'impiego delle spline permette di catturare relazioni non lineari tra variabili quantitative e risposta, migliorando ulteriormente la flessibilità del modello.