Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche

Previsione dei prezzi delle case utilizzando tecniche di machine learning: uno studio comparativo.

Relatore Prof. Livio Finos

Dipartimento di Scienze Statistiche

Correlatore: Prof. Daniel Vecchiato

Dipartimento di Territorio e Sistemi Agro-Forestali

Laureando Francesco Dubini

Matricola 2061939

Anno Accademico 2023/2024

Ai miei nonni.

Indice

Introduzione

La previsione del valore di mercato degli immobili rappresenta una sfida complessa che riveste grande importanza sia per operatori e investitori del settore, sia per privati cittadini. Stabilire in maniera accurata il prezzo di case e appartamenti è infatti cruciale in contesti come la compravendita, la stima del patrimonio, le decisioni di ristrutturazione o investimento.

Negli ultimi decenni, il diffondersi di grandi banche dati contenenti informazioni su transazioni passate ha aperto la strada allo sviluppo di modelli statistici ed econometrici per la valutazione automatica . Tuttavia, l'eterogeneità intriseca del mercato immobiliare e la molteplicità di fattori capaci di influenzare i prezzi rappresentano una sfida per i metodi tradizionali.

Più di recente, alcune ricerche hanno dimostrato come l'impiego di tecniche di machine learning possa portare a significativi miglioramenti nella precisione delle previsioni (Park e Bae, 2015). Algoritmi come random forest, boosted trees e reti neurali consentono infatti di catturare relazioni complesse all'interno di grandi masse di dati.

Nonostante i progressi compiuti, rimangono tuttavia margini di sviluppo, ad esempio nel confronto sistematico tra diverse metodologie e nell'integrazione di approcci "classici" e "black box". La prima parte di questo lavoro si occupa di rispondere a tali domande.

La seconda parte propone una metodologia che permette di ottenere per ogni immobile una stima di prezzo d'offerta e di acquisto, consentendo di misurare lo scarto. Per conseguire questo obiettivo si è ristretta l'analisi alla zona di Padova e, facendo leva sui risultati ottenuti nella prima parte, si è raggiunto l'obiettivo. I risultati possono contribuire alla comprensione empirica dello sconto, spesso poco indagato pur essendo rilevante per investitori. Inoltre,

modellizzare tali dati offre l'opportunità di acquisire informazioni aggiuntive per la valutazione, ottenibili in modo più agevole rispetto agli atti notarili. L'obiettivo è dunque fornire una stima attendibile del divario annuncio-transazione, delineando l'impatto di fattori chiave come l'ubicazione, le caratteristiche immobiliari e altri elementi rilevanti individuati tramite l'analisi quantitativa. Tale metodo è stato presentato a Padova in occasione del XLIX Incontro di Studi del Ce.S.E.T. dal titolo "IL RUOLO DEGLI INDICATORI SOCIO-ECONOMICO-AMBIENTALI NELLE POLITICHE E NELLE SCELTE DEGLI INVESTIMENTI PUBBLICI E PRIVATI".

Capitolo 1

Descrizione e creazione del dataset

1.1 Il dataset

Il dataset grezzo è stato ottenuto dal professore Daniel Vecchiato dell'Università di Padova con la tecnica del webscraping da un noto sito di compravendita di immobili. Questo consiste in 74281 inserzioni relative ad immobili ad uso residenziale nella regione Veneto estratte con la tecnica del webscraping in cinque differenti occasioni fra il 2022 e il 2023. Molte osservazioni sono di conseguenza ripetute e necessitano la rimozione. Il dataset è inoltre caratterizzato da una forte presenza di dati mancanti, in alcuni casi per variabili trascurabili, in altri per variabili rilevanti come l'anno di costruzione. La fase di pre-analisi è perciò cruciale per per raggiungere lo scopo dello studio, ovvero prevedere con la minore incertezza possibile l'offerta di vendita dell'immobile.

1.2 Utilizzo delle coordinate spaziali

La posizione geografica di un immobile è cruciale per valutare il valore dello stesso. Il dataset contiene per ciascun immobile latitudine, longitudine e comune di appartenenza. Per tenere conto di queste osservazioni si è ricorso al *Geohashing* (Moussalli, Srivatsa e Asaad, 2015). Il concetto di geohashing ha radici che risalgono al 1966, quando G.M. Morton introdusse idee simili all'interno del campo della geocodifica. Tuttavia, è stato nel 2008 che Gustavo Niemeyer ha formalizzato e diffuso il termine "geohashing" attraverso

un sistema pratico e accessibile. Niemeyer ha sviluppato un algoritmo che converte le coordinate geografiche in una stringa alfanumerica di breve lunghezza, rendendo possibile la rappresentazione di posizioni precise in modo conciso. Il geohashing si basa sull'utilizzo di una struttura gerarchica di suddivisione dello spazio in sottoinsiemi di forma griglia, utilizzando una tecnica nota come curva di ordine Z. Questa struttura consente di codificare una vasta gamma di posizioni geografiche con precisione crescente, dividendo lo spazio in sottoinsiemi sempre più piccoli. Le coordinate geografiche vengono quindi convertite in una stringa di caratteri utilizzando un algoritmo specifico, che fornisce una rappresentazione univoca della posizione. Per i dati oggetto di studio sono stati ottenuti 90 sottoinsiemi e per ciascun di essi è stato calcolato il valore medio di un immobile.

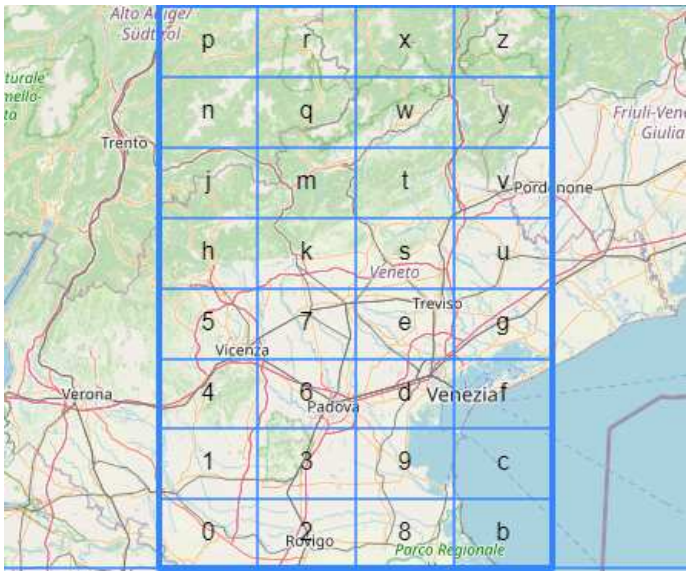


Figura 1.1: Esempio Geohashing

1.3 Imputazione della variabile anno di costruzione

La variabile anno di costruzione presenta numerosi dati mancanti, circa il 27% del totale. Per gestirli si è proceduto come segue:

1. Si Rimuove la variabile risposta **prezzo** dal dataset.

2. Come dataset di stima si tengono tutte le righe in cui la variabile **anno di di costruzione** non ha valore mancante. Come dataset di verifica utilizzo il restante delle osservazioni, ossia tutte quelle che hanno valore mancante.

3. Stimo in convalida incrociata il modello sul dataset di stima.

4. Prevedo l'anno di costruzione per quelle unità statistiche che hanno il dato mancante utilizzando il modello stimato in precedenza.

Per questa operazione di imputazione sono stati provati vari modelli ed è stato selezionato quello con la metrica MAPE più bassa.

Inoltre, viene aggiunta una variabile *dummy* che identifica quegli immobili il cui anno di costruzione è stato imputato. Si osserva dalla Figura 1.2 come la differenza fra immobili il cui anno di costruzione è stato predetto e non, sia molto lieve. Tuttavia le case per il quale è stato osservato presentano un valore al metro quadro leggermente inferiore.

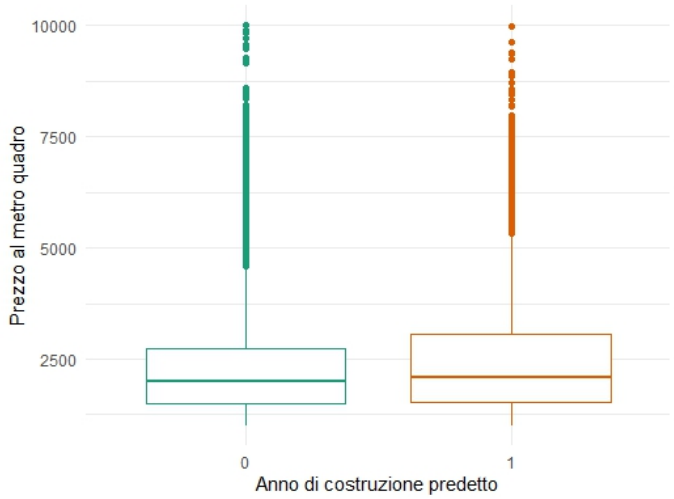


Figura 1.2: boxplot del prezzo al metro quadro condionato al fattore che indica se l'anno di costruzione è stato predetto

1.4 Dataset finale

Una volta completate le operazioni di pre-analisi, il dataset contiene 26969 osservazioni e 62 variabili predittrici. La numerosità campionaria si è ridotta di molto a causa della forte presenza di osservazioni ripetute e di inserzioni inserite in maniera errata. Per ciascun immobile si ha a disposizione:

- Posizione Geografica
- Caratteristiche fisiche dell'immobile
- Classificazione energetica
- Anno di costruzione
- Prezzo medio di un immobile nel quartiere

1.5 Analisi descrittiva

La variabile risposta utilizzata per l'analisi è il rapporto fa prezzo e superficie, ossia il **prezzo al metro quadro**, questo per ridurre l'eteroschedasticità che caratterizza questa tipologia di dati, difatti prezzi maggiori hanno una variabilità più alta, inoltre usare il prezzo al metro quadro normalizza i dati per la dimensione della proprietà, permettendo un confronto più equo tra proprietà di dimensioni diverse. Il grafico in figura 1.3 mostra una distribuzione con una coda lunga verso destra, indicando che ci sono alcune proprietà con prezzi estremamente alti rispetto alla maggior parte delle altre proprietà. La curva ha un picco netto, suggerendo che un intervallo di prezzo specifico è molto comune, ma poi la frequenza diminuisce rapidamente man mano che il prezzo aumenta. La distribuzione del prezzo al metro quadro in figura 1.4 sembra più concentrata intorno a una media, con meno variazione estrema.

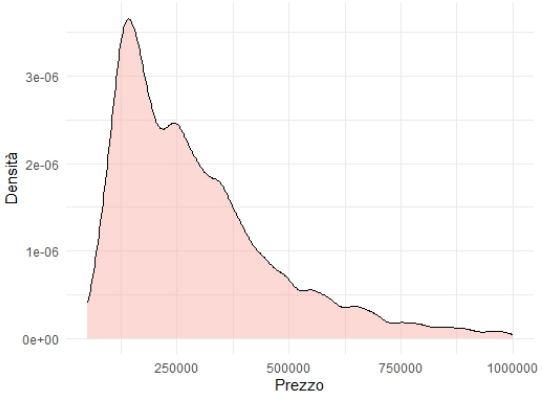


Figura 1.3: Densità del prezzo

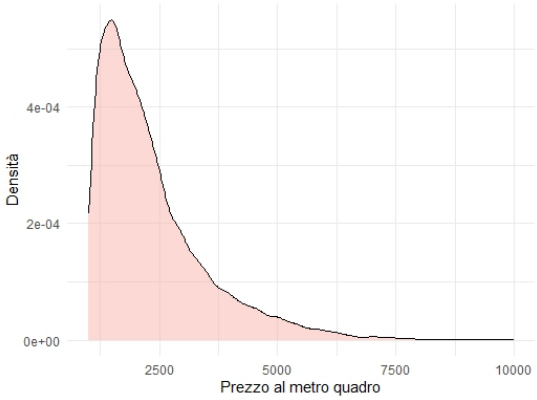


Figura 1.4: Densità prezzo al m^2

In figura 1.5 viene proposto un boxplot del prezzo al metro quadro condizionato allo stato dell'immobile. Ci sono differenze evidenti nel prezzo al metro quadro in base allo stato della proprietà. Le categorie con la mediana più alta risultano essere **Ottimo/Ristrutturato** e **Nuovo/Da ristrutturare**, suggerendo che queste proprietà hanno il prezzo al metro quadro più elevato. La categoria **Da ristruturare** presenta la mediana più bassa rispetto alle altre, con una distribuzione relativamente stretta e pochi valori estremi. Questo suggerisce che i prezzi al metro quadro per questa categoria sono relativamente bassi e consistenti.

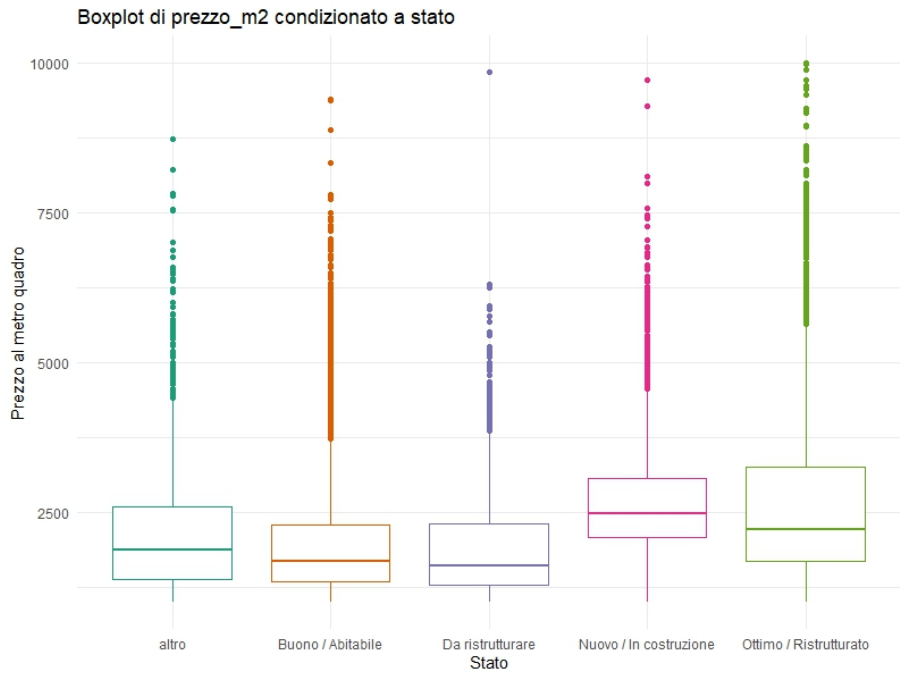


Figura 1.5: Boxplot del prezzo al metro quadro condizionato allo stato dell'immobile

La figura 1.6 mostra la densità della variabile risposta condizionata a quattro categorie della variabie **anno di costruzione**. Le proprietà costruite tra il 1300 e il 1900 potrebbero avere caratteristiche uniche, come valore storico o architettonico, che possono giustificare prezzi al metro quadro più alti. Tuttavia, il picco stretto suggerisce anche che non ci sono molte proprietà in questa fascia di prezzo, il che potrebbe essere dovuto alla scarsità di immobili che rimangono da quell'epoca o alla particolarità di quelli che sono sopravvissuti fino a oggi. Gli immobili più recenti (1990-2022) sono quelli che raggiungono i prezzi al metro quadro più alti fra gli immobili non d'eopca, come si può vedere dal grafico.

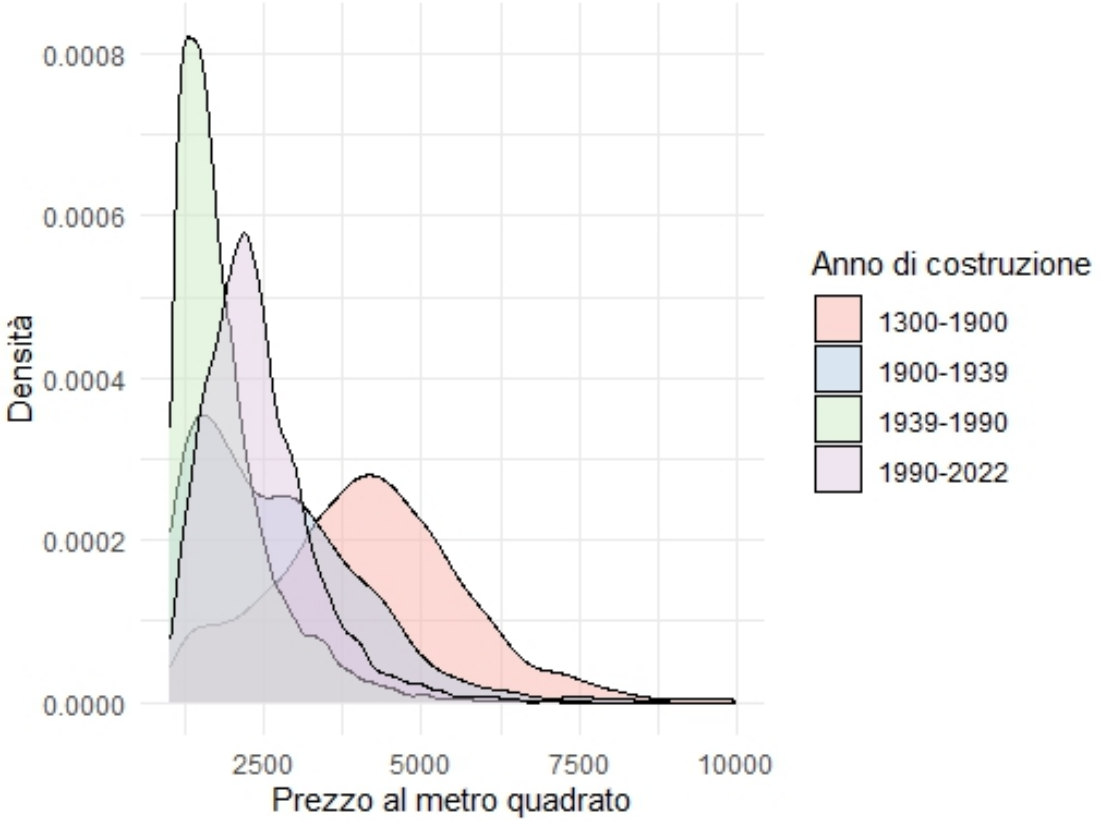


Figura 1.6: Densità prezzo al metro quadro condizionata all'anno di costruzione dell'immobile

La distribuzione dei punti in figura 1.7 mostra che, man mano che il prezzo medio per quartiere aumenta, anche il prezzo al metro quadro tende ad aumentare, ma con una variazione esponenziale. Il grafico in figura 1.8 mostra il prezzo al metro quadro rispetto alla superficie di un immobile. La tendenza generale indica che il prezzo al metro quadro tende a diminuire all'aumentare della superficie dell'immobile. La curva di densità mostra che la maggior parte degli immobili ha una superficie relativamente piccola e un prezzo per metro quadro più alto, mentre gli immobili più grandi sono meno numerosi e tendono a avere un prezzo per metro quadro inferiore.

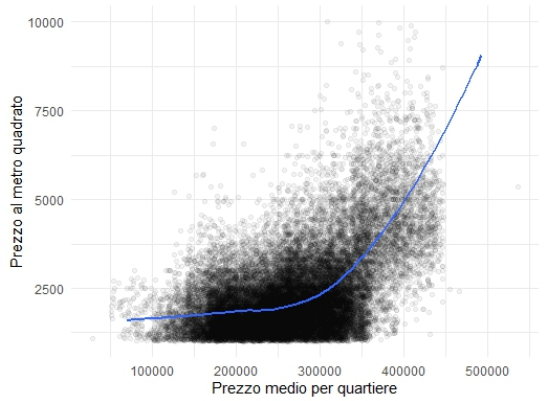


Figura 1.7: scatterplot del prezzo al metro quadro contro prezzo medio del quartiere

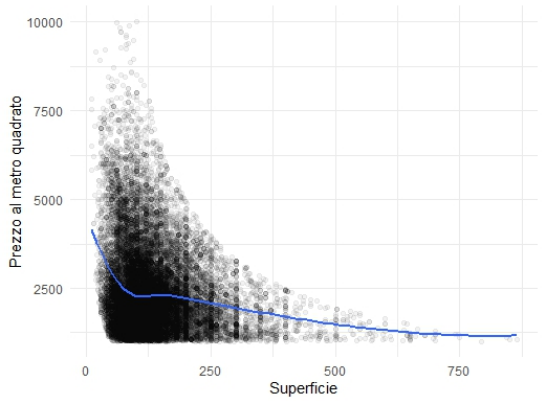


Figura 1.8: scatterplot del prezzo al metro quadro contro superficie

Il grafico in figura 1.9 mostra uno scatterplot tridimensionale rappresentante le variabili prezzo al metro quadro, superficie e valore medio di un immobile. Da questo si riesce a evincere che gli immobili caratterizzati da un valore al metro quadro più alto sono quelli situati in quartieri benestanti, ma con una superficie ridotta.

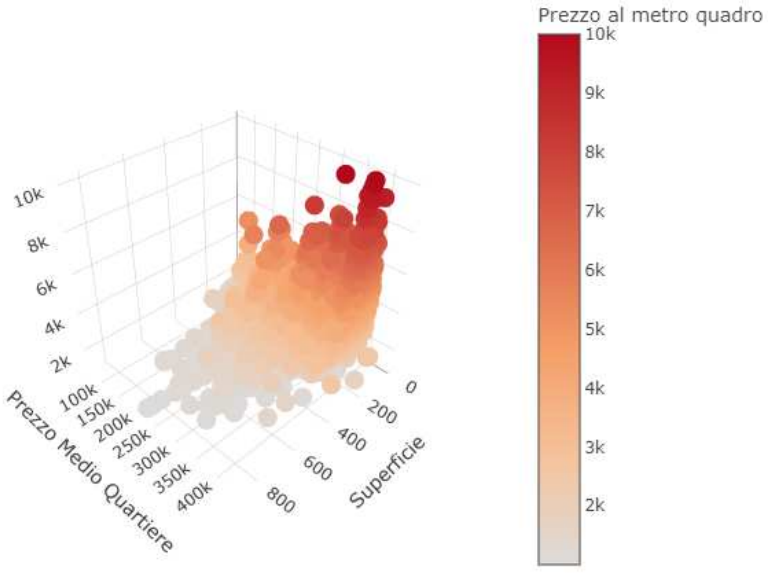


Figura 1.9: Prezzo al metro quadro contro superficie e prezzo medio di un immobile nel quartiere

Capitolo 2

Analisi

L'obiettivo principale di questo capitolo è quello di esplorare,valutare e confrontare diversi modelli interpretabili e combinazioni di approcci noti come "black box" nel contesto della stima del valore immobiliare. I modelli black box, come le reti neurali profonde e algoritmi di ensemble come Random Forest o Gradient Boosting, sono noti per la loro alta precisione e capacità di catturare relazioni complesse e non lineari nei dati. Sono particolarmente efficaci in compiti complessi e con grandi quantità di dati. Tuttavia, la loro complessità rende difficile comprenderne il funzionamento interno e giustificare le loro previsioni. I modelli interpretabili offrono una migliore comprensione di come le caratteristiche dei dati influenzano le previsioni del modello. Sono spesso più semplici e veloci da addestrare e richiedono meno dati. Questi modelli sono preferiti in situazioni in cui è necessaria la trasparenza e la comprensione delle decisioni del modello. Tuttavia, la loro semplicità può anche essere uno svantaggio, poiché possono non catturare relazioni complesse o non lineari così efficacemente come i modelli black box. Per quanto concerne i modelli interpretabili sono stati adatatti modelli lasso, ridge, elastic net, cart e MARS, mentre sono stati adattati i modelli Random Forest e XGBooost per quanto riguarda i modelli black box. La metrica di riferimento adottata per valutare e confrontare le prestazioni dei vari modelli è il Mean Absolute Percentage Error (MAPE), una misura che consente di catturare l'entità dell'errore in termini percentuali rispetto ai valori osservati.

$$\text{MAPE} = \left(\frac{1}{n} \sum_{t=1}^{n} \left| \frac{y_t - \hat{y}_t}{y_t} \right| \right) \times 100\% \tag{2.0.1}$$

La scelta del MAPE come metrica principale è motivata dalla sua capacità di fornire un'intuizione diretta sull'accuratezza del modello in termini relativi, rendendo più facile interpretare l'impatto degli errori di stima nel contesto pratico delle valutazioni immobiliari. Attraverso l'uso del MAPE, siamo in grado di quantificare l'errore medio in modo che rifletta la proporzione dell'errore rispetto al prezzo reale, permettendo così di comparare efficacemente le prestazioni dei diversi modelli indipendentemente dalla scala dei prezzi degli immobili. Di seguito sono presentati alcuni modelli, le cui implementazioni non sono illustrate in maniera estensiva poiché rappresentano approcci già affermati e largamente riconosciuti nell'ambito. Per un'analisi più approfondita di queste tecniche, si suggerisce di consultare le opere specialistiche pertinenti (Azzalini e Scarpa, 2012). Viene inoltre proposta una metodologia basata sullo Shapley Value, utile nell'interpretazioni dei modelli *ensemble*.

2.1 Modelli Interpretabili

2.1.1 Cart

La sezione è dedicata all'analisi del metodo CART (Classification and Regression Trees) e alla sua applicazione nella previsione dei prezzi immobiliari per metro quadro. Un valore di $\alpha = 0$ indica che non vi è alcuna penalità per la complessità dell'albero, il che può risultare in un albero completamente sviluppato, perfettamente adattato ai dati di addestramento ma potenzialmente sovradattato. Al contrario, un valore di $\alpha = 1$ è estremo e potrebbe condurre a un albero con una sola divisione o persino nessuna, riducendo in modo significativo la complessità del modello e rischiando di sottoadattare i dati. Il MAE riportato in Figura 2.1 diminuisce significativamente all'aumentare del parametro di complessità fino a raggiungere un valore intorno a 0.2, dopodiché l'errore si stabilizza e rimane circa lo stesso anche con l'aumento della complessità.Viene scelto quindi come valore $\alpha = 0.2$ per permettere di mantenere il modello il più semplice possibile riducendo al contempo il rischio

di overfitting.

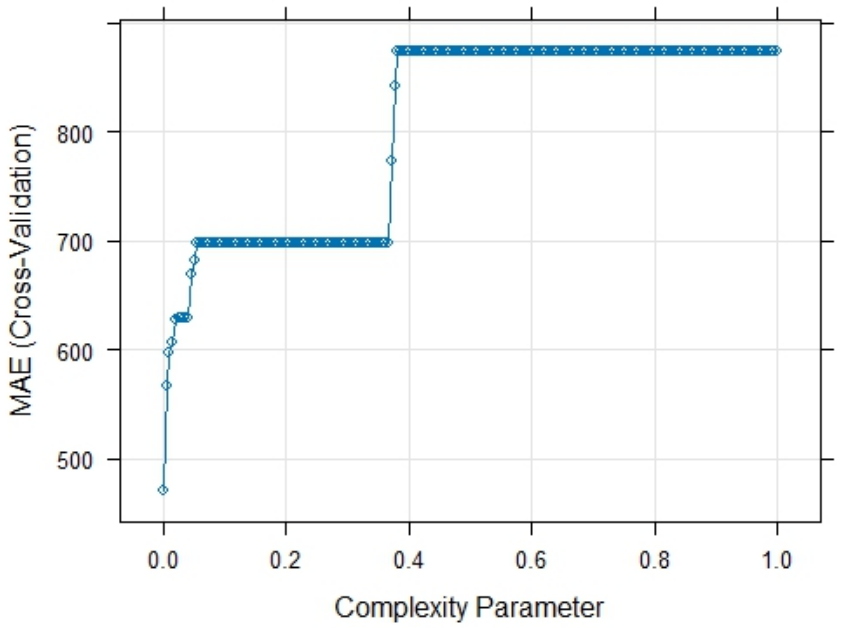


Figura 2.1: Scelta del parametro di complessità α

L'albero di decisione rappresentato nella Figura 2.2 illustra un modello di regressione sviluppato per stimare il prezzo al metro quadro di unità abitative basandosi su un insieme di variabili esplicative. Il nodo radice dell'albero effettua uno split sulla base del prezzo medio di un immobile nel quartiere, ciò a sottolineare come il contesto geografico-economico sia un importante fattore predittivo. La seconda divisione viene effettuata sulla base dello stato dell'immobile, questo sottolinea come siano importanti nella stima del valore di un immobile le condizioni dello stesso.

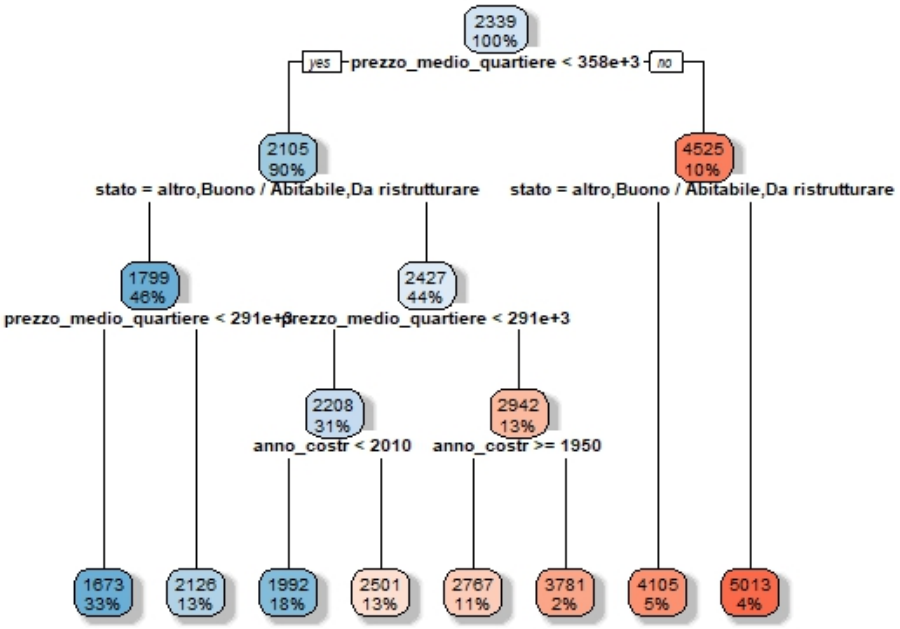


Figura 2.2: Albero di regressione

2.1.2 Regressione Regolarizzata

Regressione Lasso

Viene adattato un modello Lasso utilizzando 62 variabili e 5053 interazioni. Il grafico in Figura 2.3 mostra la metrica scelta in funzione del parametro di complessità λ. Sulla base del grafico si seleziona $\lambda = 4.41$. Il modello stima a zero 4548 parametri su 5115.

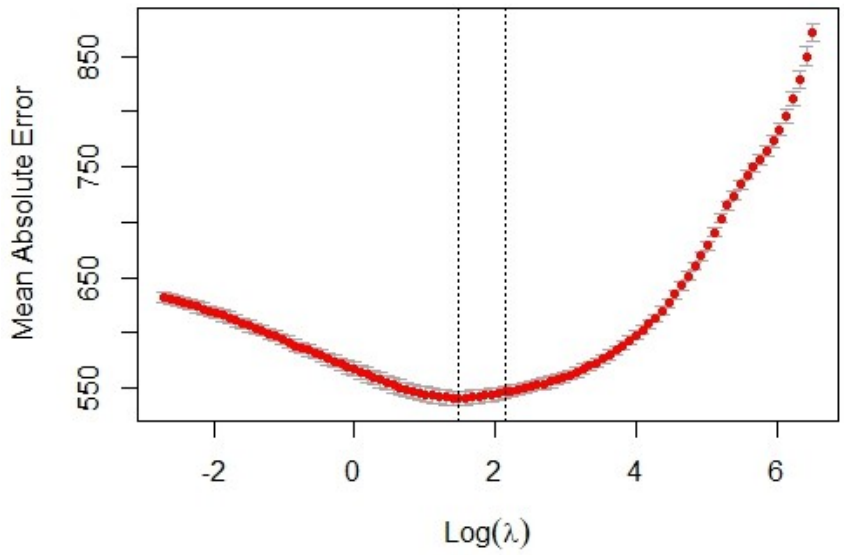


Figura 2.3: Scelta del parametro λ ottimale per il modello Lasso

Regressione Ridge

Viene adattato un modello lineare con penalizzazione L2. Viene selezionato il valore di λ ottimale sulla base del grafico in Figura 2.4.

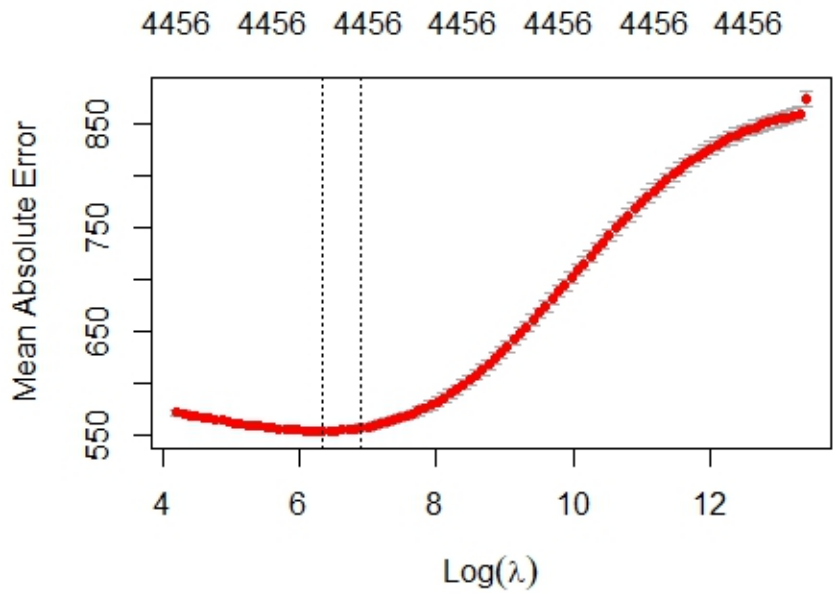


Figura 2.4: Scelta del parametro λ ottimale per il modello Ridge

Elastic Net

L'Elastic Net (De Mol, De Vito e Rosasco, 2009) è una tecnica di regressione che combina le penalità dei metodi Lasso (regressione dei minimi quadrati con penalità L1) e Ridge (regressione dei minimi quadrati con penalità L2), con l'obiettivo di sfruttare i vantaggi di entrambi i metodi di regolarizzazione. Questo è definito come un modello lineare in cui la funzione di costo è composta da una combinazione lineare della norma L1 e della norma L2 dei coefficienti del modello. Matematicamente, il problema di ottimizzazione può essere espresso come:

$$\min_{\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 + \lambda \left(\frac{1}{2}(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1 \right) \right\} \tag{2.1.1}$$

dove N è il numero di osservazioni, y_i è la variabile risposta, x_i è il vettore di predittori, β sono i coefficienti del modello, λ è il parametro di regolarizzazione che controlla l'intensità della penalità, e α è il parametro che bilancia l'importanza relativa delle penalità L1 e L2. Il grafico 2.6 mostra il criterio per la scelta del parametro di penalizzazione α, sulla base di esso viene selezionato $\alpha = 0.9$. l'Elastic Net con un valore α tra 0.25 e 1.00 ottiene i risultati migliori in termini di MAE, con una leggera preferenza per i valori appena sopra 0, dato che il MAE si stabilizza a un livello minimo. Questo suggerisce che una combinazione di penalità L1 e L2 è più efficace per il modello rispetto all'uso esclusivo di una penalità L2.

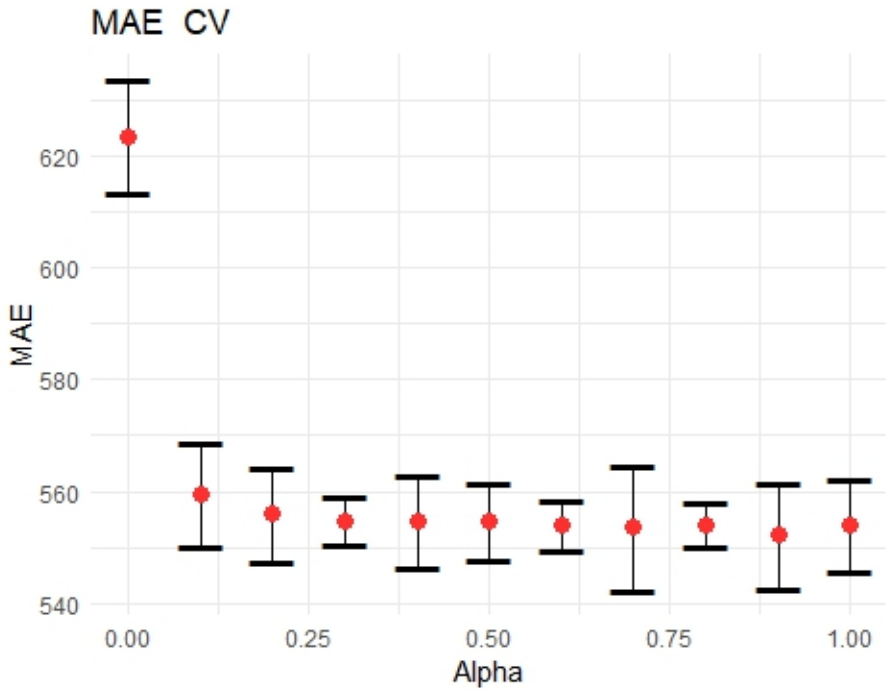


Figura 2.5: Scelta del parametro α di penalizzazione

Una volta fissato α, analogamente a quanto fatto in precedenza, si seleziona il valore del parametro λ che minimizza l'errore, come riportato in Figura 2.6.

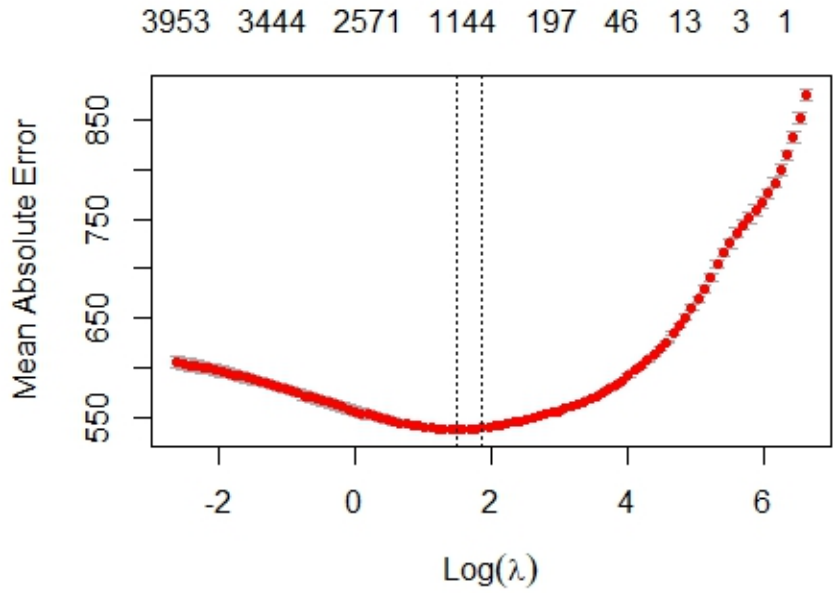


Figura 2.6: Scelta del parametro λ per l'Elastic net

2.1.3 Mars

Viene ora adattato un modello MARS (*Multivariate adaptive regression spline*). MARS estende il modello di regressione lineare tradizionale incorporando termini che catturano interazioni non lineari e relazioni a tratti attraverso l'uso di funzioni spline adattive. Queste funzioni, note come funzioni di base, sono combinate in un modello stepwise, permettendo al modello di adattarsi automaticamente alla forma e alle interazioni presenti nei dati. Il modello MARS è formulato come segue:

$$Y = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(\mathbf{X}) + \varepsilon \tag{2.1.2}$$

dove Y rappresenta la variabile risposta, β_0 è una costante, β_m sono i coefficienti delle funzioni di base $B_m(\mathbf{X})$, e ε è un termine d'errore casuale. Le funzioni di base $B_m(\mathbf{X})$ possono assumere diverse forme, tipicamente includono funzioni a cerniera di primo grado definite come:

$$B_m(\mathbf{X}) = \max(0, x_j - t_m) \quad \text{oppure} \quad B_m(\mathbf{X}) = \max(0, t_m - x_j) \tag{2.1.3}$$

dove x_j è la j-esima variabile predittiva e t_m è una soglia, che viene selezionata in modo adattativo dal modello. Per catturare le interazioni tra le variabili, MARS può costruire anche funzioni di base prodotto:

$$B_{mn}(\mathbf{X}) = B_m(\mathbf{X}) \times B_n(\mathbf{X}) \tag{2.1.4}$$

Il processo di costruzione del modello MARS procede attraverso due fasi principali: una fase di costruzione in avanti, dove le funzioni di base vengono aggiunte una per una, e una fase di potatura all'indietro, dove le funzioni di base meno significative vengono rimosse per prevenire l'overfitting e migliorare la capacità predittiva del modello. Nel grafico in Figura 2.7 viene proposto

un metodo per la scelta del grado di complessità del modello. Si opta per selezionare un grado del polinomio pari a 4.

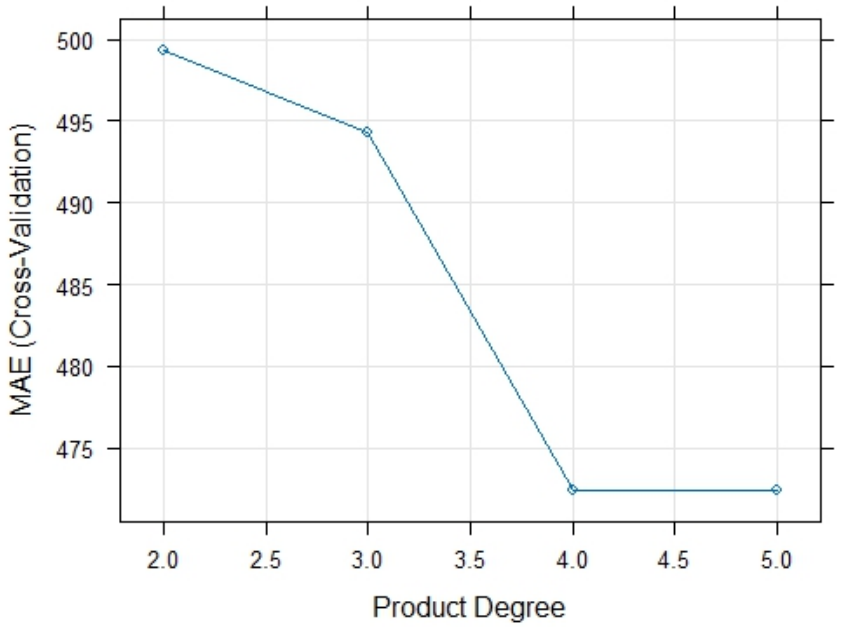


Figura 2.7: Scelta del parametro di complessità del MARS

Il grafico in Figura 2.9 mostra l'effetto stimato dal modello dell'interazione fra prezzo medio di un immobile nel quartiere e superficie. Da ciò si evince che immobili con un'ampia superficie situati in un quartiere costoso subiscono una svalutazione del loro valore al metro quadro. Le case con il prezzo al metro quadro più elevato sono quelle situate in quartieri pregiati, ma con superfici ridotte. Quello in Figura 2.8 riporta invece l'effetto dell'interazione fra anno di costruzione e superficie sulla risposta. Da questo si può osservare che gli immobili con il prezzo stimato più alto siano quelli d'epoca con una superficie ridotta.

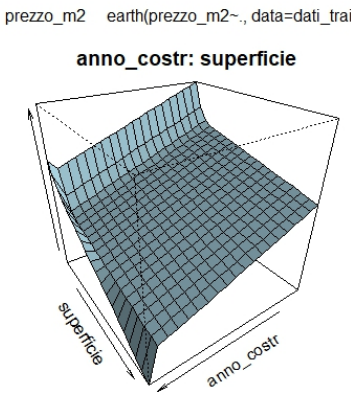


Figura 2.8: Effetto dell'interazione fra anno di costruzione e superficie

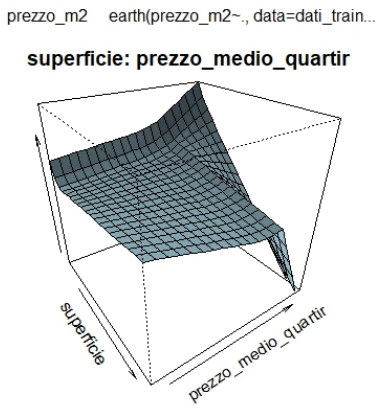


Figura 2.9: Effetto dell'interazione fra prezzo medio di un immobile e superficie

2.2 Modelli Black box

I modelli black box nel machine learning sono notoriamente potenti per la loro capacità di elaborare e modellare relazioni complesse e non lineari all'interno di grandi volumi di dati, risultando spesso in un'accuratezza predittiva superiore. Questa capacità li rende strumenti desiderabili per applicazioni che richiedono un'elevata performance. Tuttavia, la loro mancanza di trasparenza pone sfide significative: è difficile per gli utenti capire come vengono prese le decisioni all'interno di questi modelli, e la difficoltà di interpretazione può portare a problemi di fiducia e ad accettazione da parte degli utenti. Inoltre, senza una chiara visibilità del processo decisionale interno, i bias presenti nei dati possono passare inosservati e perpetuarsi nelle previsioni del modello.

Qui entra in gioco il valore di Shapley, adattato nel contesto dell'apprendimento automatico attraverso l'approccio SHAP. Questo metodo fornisce una spiegazione dettagliata dell'impatto di ogni singola feature sulla predizione di un modello. SHAP offre quindi una finestra nell'oscurità dei modelli black box, consentendo agli utenti di comprendere meglio come le diverse variabili influenzino i risultati.

2.2.1 Shapley Value

Introduzione Storica

Il Valore di Shapley è un concetto fondamentale nella teoria dei giochi cooperativi, formulato da Lloyd Shapley nel 1953 (Shapley et al., 1953). Questo concetto ha rivoluzionato il modo in cui i matematici e gli economisti considerano la distribuzione equa dei guadagni generati da una coalizione di agenti. Shapley ha introdotto un metodo basato su principi di equità e simmetria, che è stato poi premiato con il Nobel nel 2012 per il suo impatto profondo e duraturo. SHAP si basa sui valori di Shapley per fornire una spiegazione dei modelli di machine learning. SHAP connette la teoria dei giochi con le interpretazioni dei modelli, assegnando a ogni feature un valore che rappresenta il suo contributo al risultato predittivo. Questo fornisce uno strumento per interpretare modelli complessi in modo intuitivo e giustificabile. Il metodo SHAP considera ogni possibile combinazione di feature come una "coalizione" e calcola il contributo di ogni feature alla predizione, tenendo conto dell'effetto di interazione tra feature. Un valore SHAP elevato per una data feature indica un forte impatto su una predizione specifica.

Fondamenti Teorici del Valore di Shapley

Per implementare questa metodologia nel contesto dei modelli di Machine Learning (Lundberg e Lee, 2017), si definisce una funzione v che associa ad ogni sottoinsieme di caratteristiche S il prezzo che un immobile con solo quelle caratteristiche potrebbe realisticamente ottenere sul mercato.$v(N)$ rappresenta quindi il valore totale dell'immobile , dove N è l'insieme di tutte le variabili che si hanno per ciascuna casa.

Il Valore di Shapley di ciascuna caratteristica i viene calcolato utilizzando la formula:

$$\phi_i(N, v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left(v(S \cup \{i\}) - v(S)\right), \tag{2.2.1}$$

dove il contributo marginale di i $(v(S \cup \{i\}) - v(S))$ rappresenta l'incremento di valore che la caratteristica i apporta alla coalizione di caratteristiche S.

Questa metodologia consente di apprezzare il valore "vero" di ogni singola caratteristica, indipendentemente dall'ordine di valutazione, fornendo uno strumento di stima equo e basato su principi matematici solidi. Inoltre, può aiutare a identificare quali miglioramenti potrebbero aggiungere il maggiore valore all'immobile, guidando così le decisioni di ristrutturazione o di investimento.

Proprietà del Valore di Shapley

Il Valore di Shapley presenta diverse proprietà che lo rendono adeguato per la stima equa del valore delle caratteristiche immobiliari:

1. **Efficienza:** Il totale dei valori di Shapley per tutte le caratteristiche deve essere uguale al valore totale dell'immobile.

   $$\sum_{i \in N} \phi(i) = v(N) \tag{2.2.2}$$

   dove N è l'insieme di tutte le caratteristiche, $\phi(i)$ è il Valore di Shapley della caratteristica i, e $v(N)$ è il valore totale dell'immobile.

2. **Simmetria:** Se due caratteristiche contribuiscono allo stesso modo al valore dell'immobile, allora esse riceveranno lo stesso Valore di Shapley.

   $$\text{Se } v(S \cup \{i\}) = v(S \cup \{j\}) \text{ per ogni } S \subseteq N \setminus \{i, j\}, \text{ allora } \phi(i) = \phi(j). \tag{2.2.3}$$

3. **Additività:** Se le valutazioni di un immobile possono essere separate in parti indipendenti, il Valore di Shapley dell'immobile intero è la somma dei Valori di Shapley delle parti.

   $$\phi(i, v + w) = \phi(i, v) + \phi(i, w). \tag{2.2.4}$$

per ogni funzione di valutazione v e w.

4. **Assenza di Potere:** Se una caratteristica non contribuisce in alcun modo al valore dell'immobile, il suo Valore di Shapley sarà zero.

$$\text{Se } v(S \cup \{i\}) = v(S) \text{ per ogni } S \subseteq N \setminus \{i\}, \text{ allora } \phi(i) = 0. \quad (2.2.5)$$

5. **Linearità:** Il Valore di Shapley è lineare rispetto alle funzioni di valutazione.

$$\phi(i, \alpha v + \beta w) = \alpha\phi(i, v) + \beta\phi(i, w). \quad (2.2.6)$$

per ogni $\alpha, \beta \in \mathbb{R}$ e funzioni di valutazione v, w.

Applicazione del Valore di Shapley nella Stima del Valore Immobiliare

Nel contesto della valutazione immobiliare, il Valore di Shapley può essere utilizzato per attribuire un valore alle singole caratteristiche di un immobile in maniera equa e trasparente.

Algorithm 1 Calcolo del Valore di Shapley

Input: Un insieme di caratteristiche N e una funzione caratteristica v che associa un valore ad ogni sottoinsieme di N.

Output: Il Valore di Shapley $\phi(i)$ per ogni caratteristica $i \in N$.

for *ogni caratteristica* $i \in N$ **do**
| $\phi(i) \leftarrow 0$
end

for *ogni sottoinsieme* $S \subseteq N$ **do**
| Calcola $v(S)$
end

for *ogni caratteristica* $i \in N$ **do**
| **for** *ogni sottoinsieme* $S \subseteq N \setminus \{i\}$ **do**
| | $S_{con} \leftarrow S \cup \{i\}$
| | $contributo \leftarrow v(S_{con}) - v(S)$
| | $\phi(i) \leftarrow \phi(i) + \frac{|S|!\cdot(|N|-|S|-1)!}{|N|!} \cdot contributo\_marginale$
| **end**
end

return $\phi(i)$ *per ogni* $i \in N$

Ogni caratteristica della proprietà, come la localizzazione, la metratura, la presenza di un giardino o la vicinanza ai servizi, può essere considerata come un "giocatore" in un gioco cooperativo, dove il "guadagno" è il valore complessivo dell'immobile. L'algoritmo 1 mostra come calcolare per ciascuna feature il guadagno o la perdita generata per ciascuno degli immobili nel dataset.

2.2.2 Random Forest

Il grafico in Figura 2.10 mostra il MAE all'aumentare del parametro di complessità del modello. Sulla base di ciò si è deciso di utilizzaro un numero di variabili estratte casualmente pari a 10. Sebbene l'aumentare della complessità del modello non comporti un aumento della metrica MAE, quest'ultima non diminuisce in maniera significativa dopo il valore scelto e per evitare il sovradattamento si è optato per questa soluzione.

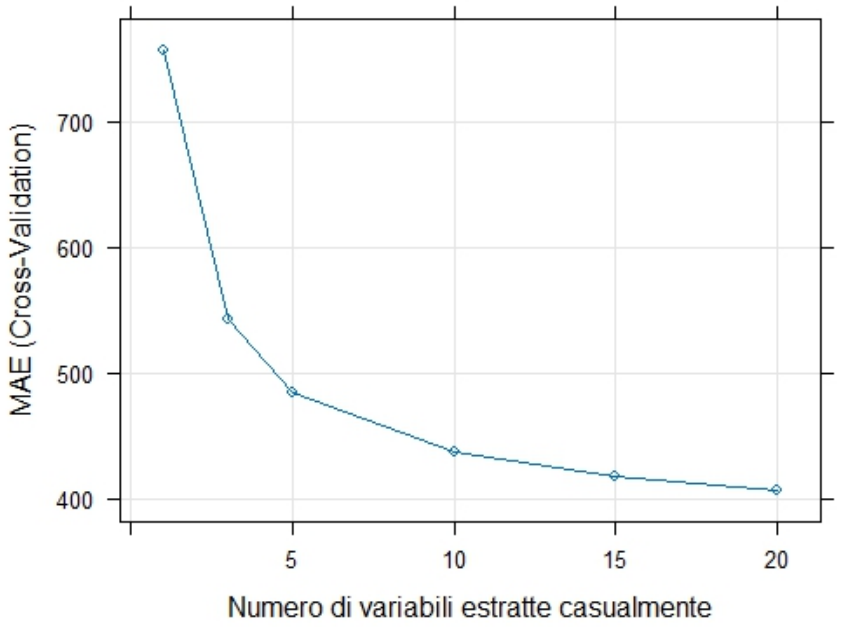


Figura 2.10: Scelta del parametro di complessità

Il grafico fornito in Figura 2.11 illustra l'applicazione dei valori di Shapley, derivati da un modello di foresta casuale, per determinare il contributo di ogni variabile indipendente nella previsione del prezzo al metro quadro di un'unità immobiliare specifica. Il valore medio di un immobile nel quartiere mostra il maggiore impatto positivo sul prezzo previsto, indicando che un valore più alto del prezzo medio nel quartiere corrisponde ad un aumento significativo del prezzo al metro quadro dell'immobile in questione. La condizione "Da ristrutturare" per lo stato dell'immobile ha un effetto negativo considerevole sulla

previsione del valore. Questo rispecchia la comprensione del mercato secondo cui gli immobili che richiedono ristrutturazioni sono generalmente valutati meno rispetto a quelli in condizioni migliori. Le coordinate geografiche, indicate da long e lat, mostrano effetti misti sul prezzo al metro quadro. Questo sottolinea l'importanza dell'ubicazione specifica all'interno di un quartiere, che può essere associata a fattori come la vicinanza a servizi, aree commerciali, o caratteristiche ambientali desiderabili.

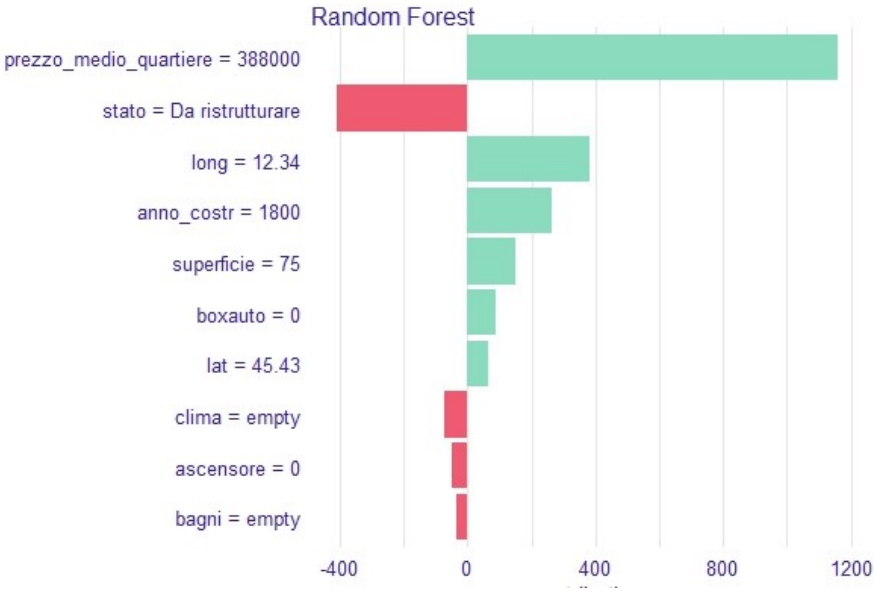


Figura 2.11: Shapley value di un immobile con Random Forest

Dall'analisi del grafico in Figura 2.12 si evince che la variabile *prezzo medio quartiere* è la più influente, seguita dall'anno di costruzione, che indica l'anno di costruzione dell'immobile. Le coordinate geografiche, long e lat, hanno un'influenza moderata, suggerendo che la posizione specifica all'interno del quartiere ha un impatto significativo sul prezzo.

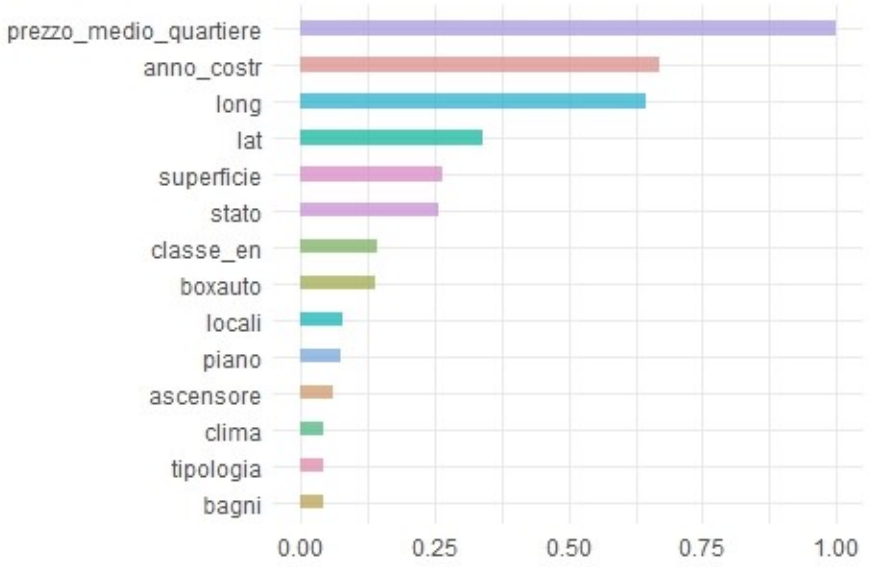


Figura 2.12: Importanza delle variabili ottenuta tramite permutazioni

2.2.3 Extreme Gradient Boosting

Il boosting è una tecnica di machine learning che crea un modello predittivo forte combinando più modelli deboli, tipicamente degli alberi di decisione, in una sequenza iterativa. Ogni modello successivo si concentra sugli errori commessi dal modello precedente, attribuendo un peso maggiore agli esempi che sono stati previsti in modo errato. L'obiettivo è di migliorare la previsione laddove il modello attuale è deficitario. Il boosting riduce il bias e può aumentare la precisione del modello predittivo finale. XGBoost (Chen e Guestrin, 2016) è un algoritmo che ottimizza la funzione obiettivo del boosting del gradiente in due modi principali: un approccio regolarizzato per controllare l'overfitting e un'implementazione computazionalmente efficiente che utilizza tecniche come il parallelismo, il pruning degli alberi e la gestione ottimizzata delle strutture dati. A differenza dell'implementazione standard del boosting del gradiente, XGBoost aggiunge un termine di regolarizzazione alla funzione obiettivo, che aiuta a migliorare la generalizzazione del modello su dati non visti riducendo la complessità del modello finale

Algoritmo XGBoost

L'algoritmo di Extreme Gradient Boosting (XGBoost) è un metodo di apprendimento supervisionato che migliora il modello in maniera additiva attraverso l'uso di alberi decisionali. Il processo include aspetti teorici e viene governato da formule specifiche:

1. **Inizializzazione**: Il modello viene inizializzato con una stima costante, solitamente il valore che minimizza la funzione di perdita L per il set di dati di training:
$$\hat{y}^{(0)} = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$$
dove $\hat{y}^{(0)}$ è la predizione iniziale, L è la funzione di perdita, y_i sono i valori target reali e γ è la costante iniziale.

2. **Costruzione Additiva di Alberi**:
   - A ogni iterazione t, un nuovo modello di albero $f_t(x)$ viene aggiunto, dove x rappresenta le caratteristiche dell'istanza.
   - L'albero è addestrato per adattarsi ai residui o agli errori del modello precedente utilizzando la somma dei gradienti:
$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + \nu f_t(x)$$
dove ν è il tasso di apprendimento (learning rate).

3. **Calcolo dei Gradienti e degli Hessiani**:
   - Per ogni istanza nel set di dati, vengono calcolati il gradiente g_i e l'hessiano h_i della funzione di perdita rispetto alla predizione attuale:
$$g_i = \partial_{\hat{y}^{(t-1)}} L(y_i, \hat{y}^{(t-1)})$$
$$h_i = \partial^2_{\hat{y}^{(t-1)}} L(y_i, \hat{y}^{(t-1)})$$
   - I gradienti g_i e gli hessiani h_i sono utilizzati per guidare la selezione delle suddivisioni durante la crescita dell'albero.

4. **Pruning degli Alberi**:

   - Viene calcolato il guadagno di informazione di ogni possibile suddivisione e confrontato con un parametro di complessità γ per decidere se eseguire il pruning.
   - Il pruning riduce la complessità del modello per prevenire l'overfitting.

5. **Ottimizzazione del Modello**:

   - I nuovi alberi vengono aggiunti minimizzando la funzione obiettivo $\mathcal{L}^{(t)}$:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} L(y_i, \hat{y}^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

   - $\Omega(f_t)$ è il termine di regolarizzazione che penalizza la complessità dell'albero f_t, definito come:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

     dove T è il numero totale di foglie dell'albero, w_j sono i pesi delle foglie.

6. **Aggiornamento del Modello**:

   - Dopo la costruzione di ogni albero, i parametri del modello vengono aggiornati per ottimizzare la funzione obiettivo.
   - Questo aggiornamento utilizza il gradiente della funzione di perdita e un tasso di apprendimento per ridurre l'errore.

7. **Fine del Training**:

   - Il modello finale è costituito dalla somma ponderata di tutti gli alberi costruiti durante l'addestramento.
   - Ogni albero contribuisce alla predizione finale con un peso determinato dal tasso di apprendimento ν. La predizione finale per un'istanza

è data da:

$$\hat{y} = \sum_{t=1}^{T} \nu f_t(x)$$

dove T è il numero totale di alberi nel modello finale.

Risultati XGBoost

Il grafico in Figura 2.13 mostra una serie di curve di validazione incrociata per un modello di XGBoost, con l'asse delle ordinate che rappresenta l'Errore Assoluto Medio (MAE) della validazione incrociata e l'asse delle ascisse che rappresenta il numero di iterazioni di boosting.

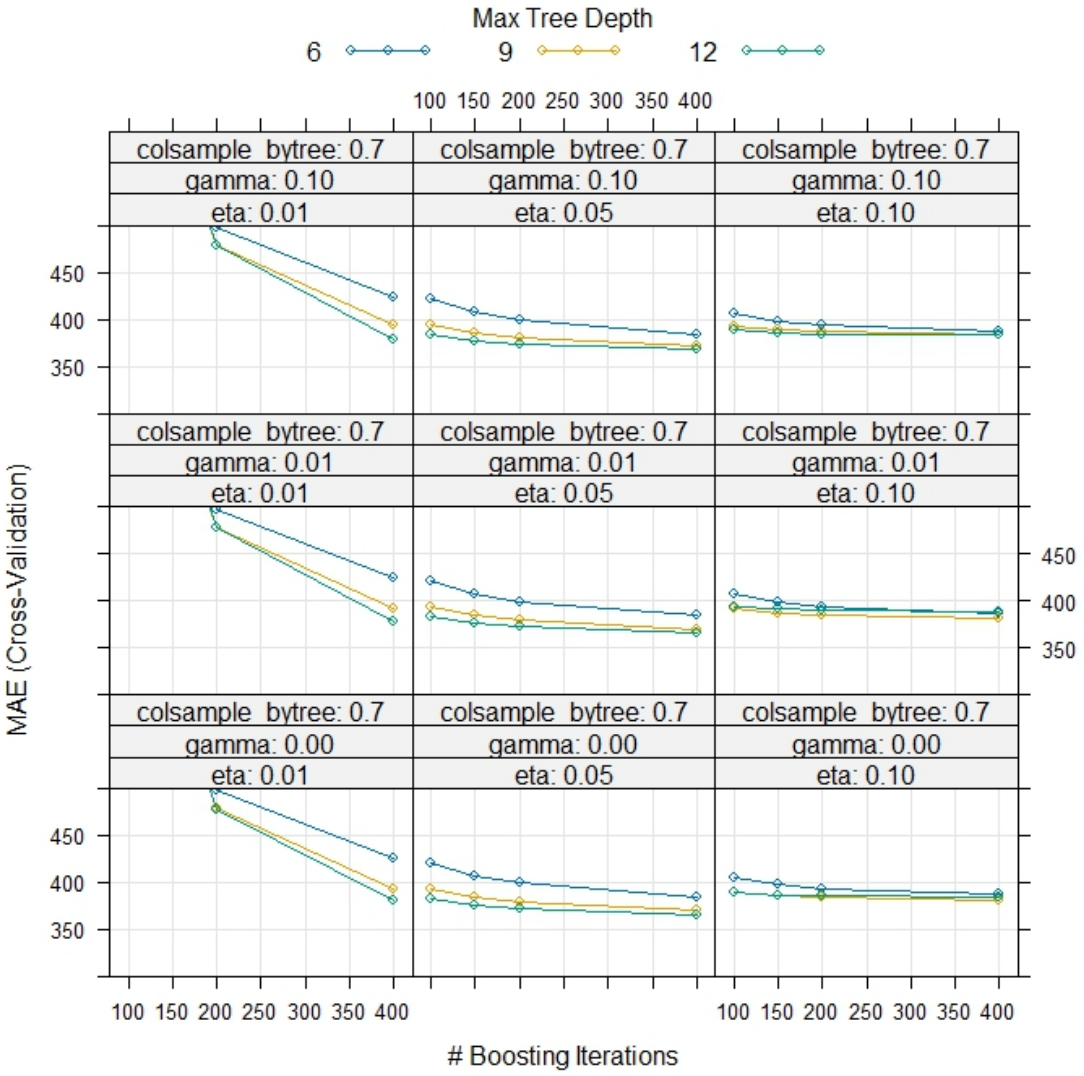


Figura 2.13: Tuning dei parametri per XGBoost

Vengono selezionati quindi come parametri ottimali:

- **proporzione di variabili estratte**: 0.7
- $\boldsymbol{\Omega}(f_t)$: 0.01
- $\boldsymbol{\nu}$: 0.05
- **Profondità degli alberi**: 6. Non sembra esserci una differenza significativa nel MAE per profondità diverse. Tuttavia, alberi troppo profondi potrebbero aumentare il rischio di overfitting. Pertanto, una profondità di 6 o 9 potrebbe essere preferibile.

Viene proposto nel grafico 2.14 in Figura l'importanza delle variabili ottenuta tramite permutazioni. Si può notare come le variabile che primeggiano siano il prezzo medio di un immobile nel quartiere, lo stato dell'immobile e la posizione.

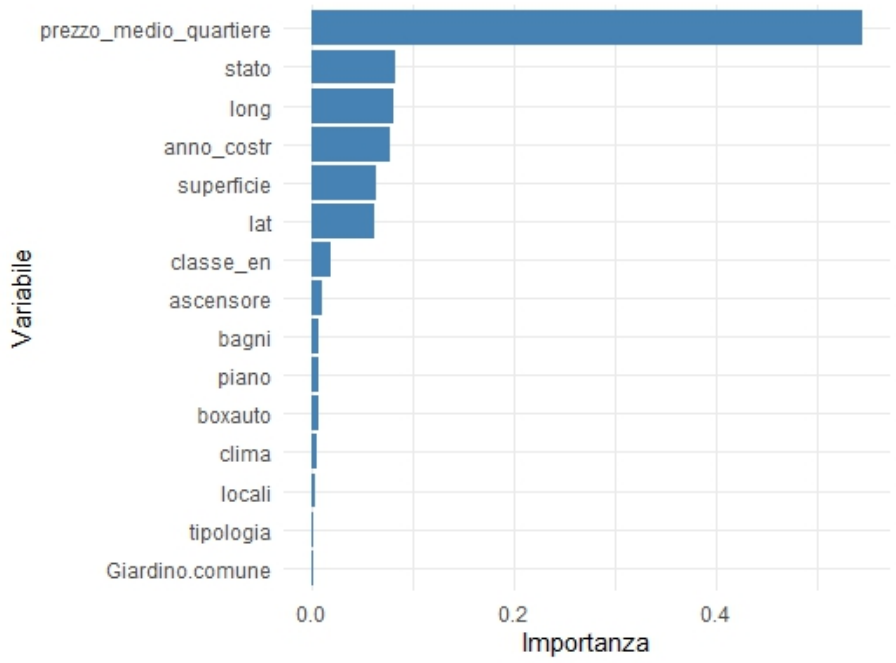


Figura 2.14: Importanza relativa delle variabili ottenuta tramite permutazioni

Il grafico in Figura 2.15 mostra gli Shapley Values per uno degli immobili nel dataset. Per quanto concerne la casa presa in esame il maggior au-

mento di valore è determinato dall'ottimo stato, da un quartiere costoso e da una buona classe energetica. La maggiore diminuzione del prezzo previsto è invece determinata dalla posizione all'interno del quartiere e dall'anno di costruzione.

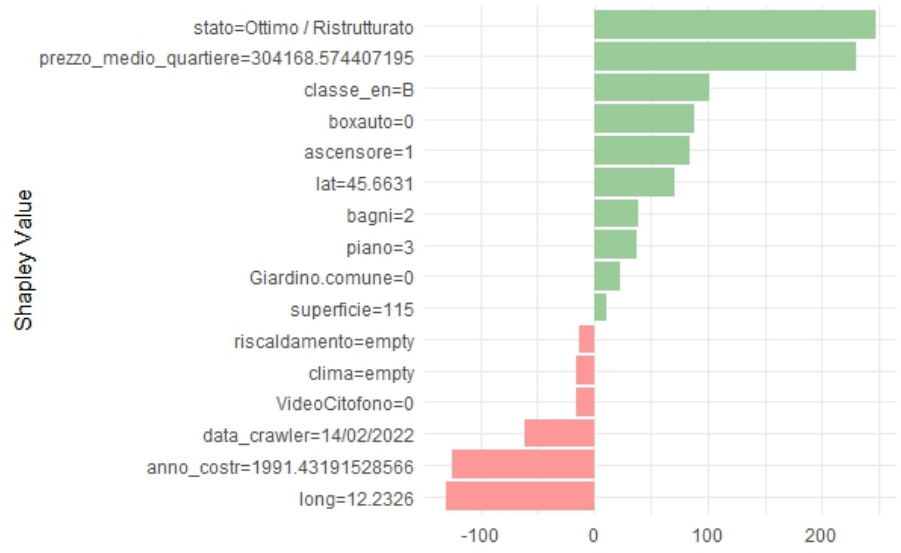


Figura 2.15: Shapley value di un immobile con XGBoost

Il grafico in figura 2.16 mostra l'importanza delle caratteristiche (feature importance) calcolata con lo Shapley Value, calcolata come il valore medio assoluto dei valori SHAP per ogni caratteristica. Inoltre consente di osservare se valori positivi o negativi corrispondono a valori alti o bassi di ciascuna variabili. Uno stato migliore dell'immobile, per esempio, comporta un aumento dello Shapley Value, mentre è evidente che più è alto il valore medio di un immobile nel quartiere, più aumenta il prezzo al medro quadro di una casa.

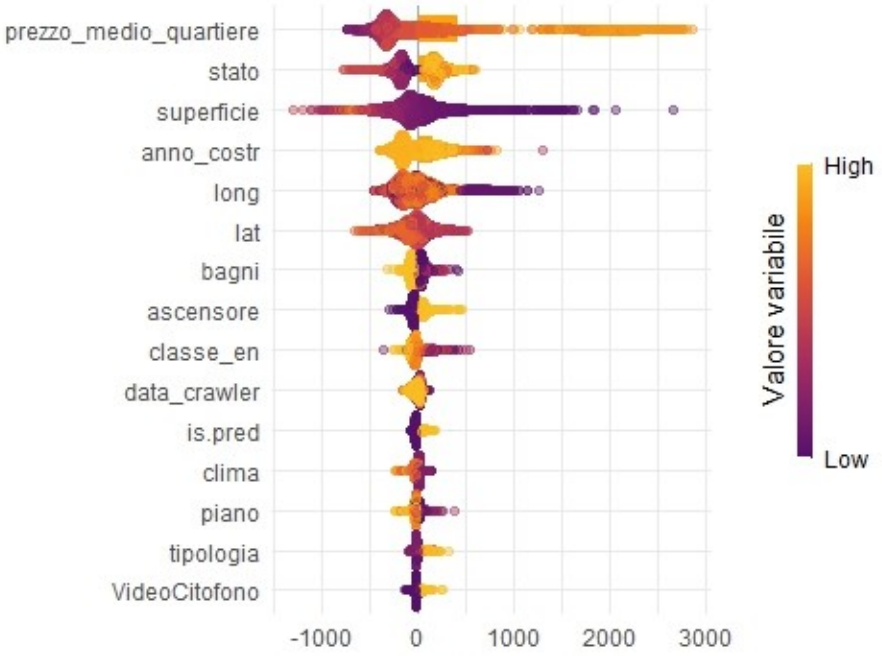


Figura 2.16: Importanza delle variabili utilizzando lo Shapley value

2.3 Confronto dei risultati

Vengono riportate le performance dei modelli nella tabella 2.1. I risultati evidenziano la superiorità dei più complessi, in particolare Random Forest e XGBoost, rispetto ai modelli lineari e ai semplici alberi decisionali per il dataset in esame. Il modello CART, nonostante la sua semplicità e la sua interpretabilità intuitiva, è risultato il meno performante. Il MARS, essendo un'estensione non lineare dei modelli lineari, migliora il MAPE attraverso l'utilizzo di spline adattive, che permettono di modellare relazioni non lineari e interazioni. Il confronto delle prestazioni di questi modelli evidenzia l'importanza di utilizzare tecniche di combinazioni di modelli e metodi di regolarizzazione per migliorare l'accuratezza predittiva e la generalizzazione del problema. Particolarmente importanti rsultano essere le interazioni ai fini della modellazione del fenomeno. XGBoost,sebbene sia il migliore, è tuttavia un modello black box la cui interpretabilità solitamente è molto scarsa. Si è analizzato tuttavia come l'impiego dello Shapley Value sopperisca, seppur parzialmente, a questa

Modello	MAPE (%)
CART	21.08
Lasso	25.9
Ridge	29.98
Elastic Net	24.6
MARS	20.1
Random Forest	17.2
XGBoost	16.3

Tabella 2.1: Performance di diversi modelli di previsione.

mancanza dei modelli ensemble. L'analisi dei valori SHAP tramite box plot in figura 2.17 rivela che lo stato dell'immobile svolge un ruolo distinto nelle previsioni del modello. Le condizioni migliori, come "Ottimo / Ristrutturato", tendono a influenzare positivamente le previsioni, mentre gli immobili "Da ristrutturare" o "Buono / Abitabile" sono associati a impatti negativi. La variazione all'interno delle categorie e gli outlier suggeriscono eccezioni individuali a questo schema generale.

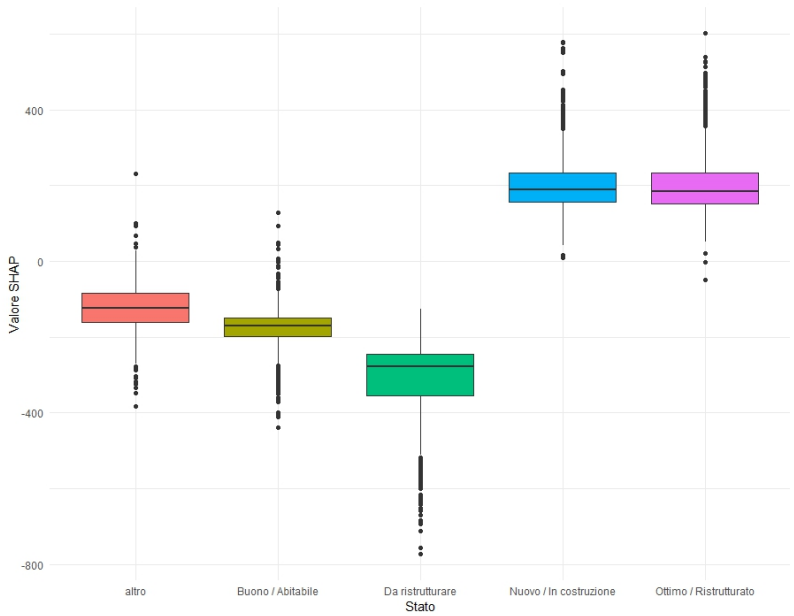


Figura 2.17: Boxplot degli Shapley Values condizionati allo stato.

Nella Figura 2.18, si presenta uno scatter plot che correla i Valori di Shapley con l'anno di costruzione degli immobili, con una differenziazione cromatica che rappresenta lo stato attuale dell'immobile. Dai dati si evince una tendenza interessante: gli immobili classificati come da ristrutturare sembrano presentare Valori di Shapley mediamente più elevati rispetto a quelli in condizioni ottime o ristrutturate, particolarmente per costruzioni più recenti.

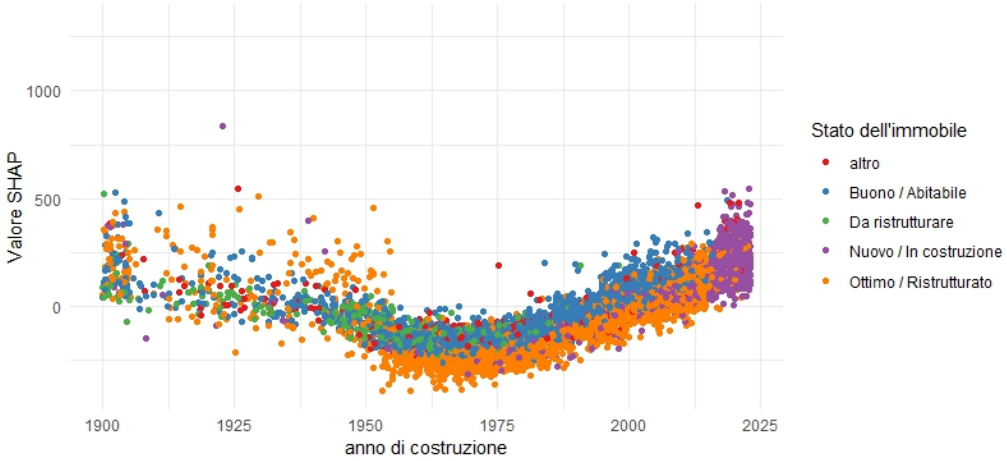


Figura 2.18: Scatterplot degli Shapley Values dell'anno di costruzione, colorati in base allo stato dell'immobile.

Capitolo 3

Conformal inference

3.1 Introduzione

L'inferenza intervallare è una branca della statistica in rapida ascesa applicata ai modelli di previsione. I metodi di machine learning classici producono come output una stima puntuale, senza tuttavia alcun intervallo di credibilità associato, ossia una misura dell'incertezza su ciascuna delle previsioni effettuate. Nel contesto dello studio condotto è di conseguenza di interesse prevedere per ciascun immobile un range di prezzo che sia il più verosimile possibile. Un approccio popolare è quello PAC (Probably Approximately Correct) ha come unica assunzione che i dati siano generati da una distribuzione sconosciuta i.i.d (Lei et al., 2018).

3.2 Split Conformal Inference

Sebbene sia un approccio piuttosto recente al machine learning, l'idea sembra essere nata negli anni '90 da una conversazione fra Vovk, Gammerman e Vapnik. L'idea alla base di questa tecnica è abbastanza semplice. Sia X l'insieme di stima e Y l'insieme di verifica. Per decidere se un valore y è incluso in un intervallo $\hat{C}(X_{n+1})$ si considera l'ipotesi nulla:

$$H_y : Y_{n+1} = y \tag{3.2.1}$$

E si costruisce il test basandoci sul campione aumentato:

$$(X_1, Y_1), .., (X_n, Y_n), (X_{n+1}, y) \tag{3.2.2}$$

L'intervallo di predizione sarà tale che i valori di y siano conformi a H_y. Per costruire un test si considera il caso semplificato dove $U_1, ..., U_n, U_{n+1}$ è generata da una distribuzione i.i.d scalare. Si ha quindi che:

$$T = \sum_{i=1}^{n+1} \mathbb{1}(U_i < U_{n+1}) \tag{3.2.3}$$

E' uniformemente distribuita su $1, ..., n+1$, si ha quindi che:

$$P(T \leq [(n+1)(1-\alpha)]) \geq 1 - \alpha \tag{3.2.4}$$

L'algoritmo della Conformal Inference classica è tuttavia molto oneroso computazionalmente, in questo paragrafo viene proposta una sua variazione chiamata *Split Conformal Inference* che consente di ridurre di molto lo stesso.

Algorithm 2 Split Conformal Prediction

Input: (X_i, Y_i), $i = 1, \ldots, n$, α, modello di previsione A

Output: $C^{\text{split}}(x)$ per ogni $x \in \mathbb{R}^p$

Function `SplitConformalPrediction(`(X_i, Y_i), α, A`):`

Divido casualmente $\{1, \ldots, n\}$ in due sottoinsiemi con la stessa lunghezza I_1, I_2

$\hat{y} \leftarrow A(\{(X_i, Y_i) : i \in I_1\})$

$R_i \leftarrow |Y_i - \hat{y}_i|$ per ogni $i \in I_2$

$d \leftarrow$ il k-esimo valore in $\{R_i : i \in I_2\}$, dove $k = \lceil d(n+1)(1-\alpha) \rceil$

$C^{\text{split}}(x) \leftarrow [\hat{y} - d, \hat{y} + d]$ per ogni $x \in \mathbb{R}^p$

return $C^{split}(x)$

L'espressione matematica

$$P[Y_{n+1} \in \hat{C}_{\text{split}}(X_{n+1})] \in \left[1 - \alpha, 1 - \alpha + \frac{1}{n+2}\right] \tag{3.2.5}$$

può essere interpretata come segue:

Si tratta della probabilità che la previsione corretta per l'osservazione successiva Y_{n+1} cada all'interno dell'intervallo di previsione $\hat{C}_{\text{split}}(X_{n+1})$,

che è un intervallo di confidenza calcolato utilizzando l'algoritmo di divisione conforme. Questa probabilità si trova nell'intervallo chiuso da $1-\alpha$ a $1-\alpha+\frac{1}{n+2}$.

In altre parole, ci si aspetta che la previsione corretta cada all'interno di questo intervallo con una probabilità compresa tra $1-\alpha$ e $1-\alpha+\frac{1}{n+2}$, dove α è il livello di significatività e n è la dimensione del campione. Questa proprietà è indicativa della precisione e della copertura dell'intervallo di previsione conforme.

Risultati

La Figura 3.1 rappresenta un grafico a dispersione generato attraverso un processo di conformal inference, con un livello di significatività fissato al 10%. Sull'asse delle ordinate (Y) è riportato il prezzo previsto dal modello, mentre sull'asse delle ascisse (X) è indicato il prezzo effettivamente osservato nel mercato. L'analisi del grafico evidenzia una correlazione positiva tra le variabili, indicativa di una buona aderenza del modello predittivo ai dati osservati. La maggior parte dei punti si concentra lungo la bisettrice, che rappresenta il luogo dei punti in cui il prezzo previsto coincide esattamente con il prezzo osservato. La banda grigia che si estende su entrambi i lati della bisettrice indica l'intervallo di conformal prediction al 10%. Questo intervallo rappresenta la regione entro cui ci si aspetta che cadano i prezzi previsti con una copertura probabilistica del 90%, tenendo conto del livello di incertezza del modello. Il fatto che la maggior parte dei punti si trovi all'interno di queste bande dimostra l'efficacia del modello nel catturare l'incertezza associata alle previsioni.

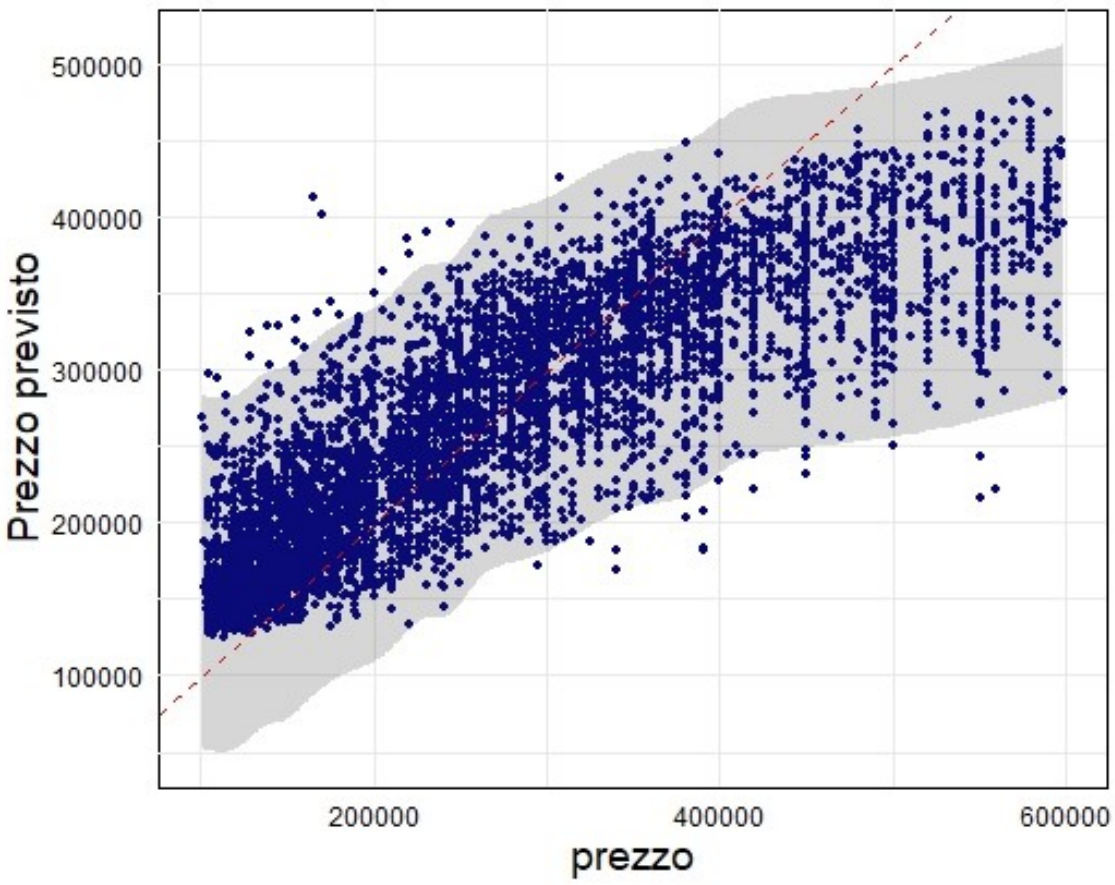


Figura 3.1: Inferenza conforme non pesata

3.3 Local Weighted Conformal Inference

Il metodo della *Split Conformal Inference*, tende a produrre intervalli di previsione $\hat{C}(x)$ la cui larghezza è approssimativamente costante su x.

L'inferenza conforme pesata (Tibshirani et al., 2019) è una tecnica derivata dall'inferenza conforme che mira a migliorare le prestazioni dell'inferenza considerando l'eteroschedasticità dei dati. In presenza di eteroschedasticità, ossia quando la varianza dei residui non è costante su tutto il dominio delle variabili indipendenti, l'inferenza conforme standard potrebbe produrre intervalli di confidenza non ottimali. Quando il rumore è eteroschedastico, possiamo modificare la definizione dei residui

nell'Algoritmo 2 scalando i residui adattati come

$$R_{y,i} = \frac{|Y_i - \hat{Y}|}{\hat{\rho}(X_i)}, \quad i \in I_2 \tag{3.3.1}$$

dove $\hat{\rho}(x)$ indica una stima della deviazione media assoluta condizionale (MAD) di $\{Y - \hat{Y}\}|X = x$ basata sui campioni in I_1. L'intervallo di previsione di output in un punto x deve anche essere modificato, diventando ora:

$$[\hat{Y} - \hat{\rho}(x)d, \hat{Y} + \hat{\rho}(x)d] \tag{3.3.2}$$

L'algoritmo sarà perciò:

Algorithm 3 Local weighted conformal inference

Input: (X_i, Y_i), $i = 1, \ldots, n$, α, modello di previsione A

Output: $C^{\text{split}}(x)$ per ogni $x \in \mathbb{R}^p$

Function `WeightedConformalPrediction`$((X_i, Y_i),\ \alpha,\ A)$**:**

Divido casualmente $\{1, \ldots, n\}$ in due sottoinsiemi con la stessa lunghezza I_1, I_2

$\hat{y} \leftarrow A(\{(X_i, Y_i) : i \in I_1\})$

$\hat{\rho} \leftarrow |y_i - \hat{y}_i|X = x$ per ogni $i \in I_2$

$R_{y,i} \leftarrow |\frac{Y_i - \hat{y}_i}{\hat{\rho}}|$ per ogni $i \in I_2$

$d \leftarrow$ il k-esimo valore in $\{R_i : i \in I_2\}$, dove $k = \lceil d(n+1)(1-\alpha) \rceil$

$C^{\text{split}}(x) \leftarrow [\hat{y} - d\hat{\rho}(x), \hat{y} + d\hat{\rho}(x)]$ per ogni $x \in \mathbb{R}^p$

return $C^{split}(x)$

Risultati

Il grafico in figura 3.3 rivela che il modello fornisce stime ragionevolmente accurate per un'ampia gamma di prezzi, con una dispersione che aumenta per prezzi più elevati. Questo potrebbe indicare che il modello gestisce meno efficacemente gli oggetti di superficie più grande o di valore più alto. Le previsioni per gli oggetti di valore inferiore sembrano essere più concentrate e quindi potenzialmente più precise. Le bande di previsione

catturano la maggior parte dei dati, ma la dispersione crescente con l'aumentare del prezzo suggerisce che il modello potrebbe essere migliorato per prevedere con maggiore precisione il valore degli oggetti di superficie più grande.

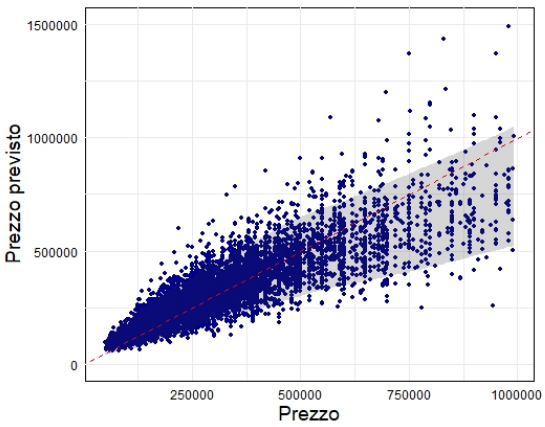


Figura 3.2

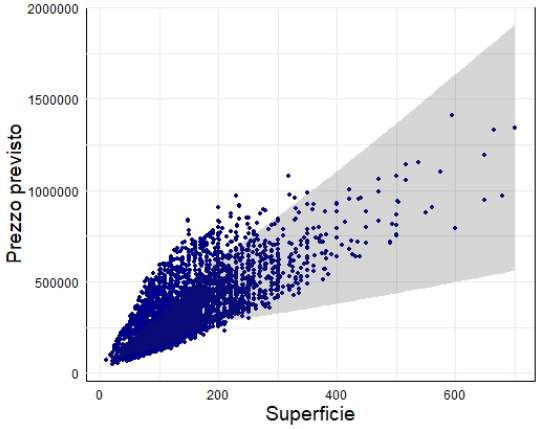


Figura 3.3

Figura 3.4: Inferenza conforme pesata

3.4 Conformalized Quantile Prediction

L'origine concettuale della conformal inference basata sui quantili può essere fatta risalire agli sviluppi teorici nei campi della statistica computazionale e dell'apprendimento automatico.

La CP classica è un metodo non parametrico che offre garanzie di copertura senza fare assunzioni specifiche sulla distribuzione dei dati. È versatile e relativamente facile da implementare ma potrebbe non essere ottimale in termini di lunghezza degli intervalli predittivi.

La QCP (Romano, Patterson e Candes, 2019) invece personalizza gli intervalli predittivi basandosi sui quantili, che possono fornire una misura più robusta e accurata della variabilità dei dati. Tuttavia, questa robustezza viene a un costo computazionale maggiore. I vantaggi della Quantile Conformal Prediction (QCP) includono:

- **Robustezza:** I quantili forniscono una misura più resistente agli outlier rispetto ai metodi basati sulla media. Ciò rende QCP particolarmente efficace nella gestione di distribuzioni di dati non simmetriche o con outlier significativi.
- **Adattabilità:** Gli intervalli predittivi generati dalla QCP si adattano meglio alla distribuzione locale dei dati. Questo significa che gli intervalli possono riflettere più accuratamente l'incertezza inerente a diverse regioni dello spazio dei dati.
- **Flessibilità:** La QCP mantiene la flessibilità della Conformal Prediction (CP) classica, potendo essere applicata a una gamma di modelli predittivi, rendendola una metodologia versatile per l'inferenza predittiva.

Algorithm 4 Quantile conformal Inference

Input: (X_i, Y_i), $i = 1, \ldots, n$, α_1,α_2, modello di previsione per quantili A

Output: $C^{\text{split}}(x)$ per ogni $x \in \mathbb{R}^p$

Function `QuantileConformalPrediction`$((X_i, Y_i)$, α_1,α_2, $A)$**:**

Divido casualmente $\{1, \ldots, n\}$ in due sottoinsiemi con la stessa lunghezza

I_1, I_2

$\hat{q_a}, \hat{q_\alpha} \leftarrow A(\{(X_i, Y_i) : i \in I_1\})$

$E_i \leftarrow max[\hat{q}_{a_1}(X_i) - Y_i, \hat{q}_{a_2}(X_i) - Y_i]$ per ogni $i \in I_2$:

$QE \leftarrow quantile(E, \alpha_2 - \alpha_1)$ per ogni $i \in I_2$;

$C^{\text{split}}(x) \leftarrow [\hat{q}_{\alpha_1} - QE(x), \hat{q}_{\alpha_2} + QE(x)]$ per ogni $x \in \mathbb{R}^p$

return $C^{split}(x)$

L'algoritmo 4 fornisce un modo per calcolare intervalli predittivi conformalizzati in modo che contengano la variabile di risposta reale con una probabilità almeno pari a $1 - (\alpha_2 - \alpha_1)$, assumendo che i dati siano scambiabili. La scelta degli insiemi I_1 e I_2 è cruciale per garantire che la procedura sia veramente non parametrica e distribuzione-free.

Risultati

I risultati ottenuti mediante l'applicazione del metodo della conformal inference quantilica sono illustrati in Figura 3.5. Sull'asse delle ordinate (prezzo previsto) e sull'asse delle ascisse (prezzo osservato), si osserva che la maggior parte dei punti si concentra lungo una linea diagonale che suggerisce una forte correlazione tra i prezzi osservati e quelli previsti dal modello. Questo indica che il metodo di conformal inference quantilica è generalmente efficace nel prevedere il prezzo degli immobili basandosi sulle caratteristiche inserite nel modello. La linea rossa nel grafico rappresenta la linea di identità, dove il prezzo previsto è esattamente uguale al prezzo osservato. Si può notare che molti dei punti cadono vicino a questa linea, il che implica che per molti immobili il prezzo previsto dal modello è molto vicino all'offerta di vendita reale.. Le bande di previsione grigie indicano l'intervallo di confidenza per le previsioni del prezzo. L'ampiezza di queste bande aumenta allontanandosi dalla linea di identità, riflettendo un aumento dell'incertezza nelle previsioni per i prezzi più elevati. Ciò potrebbe essere dovuto al fatto che immobili con prezzi più alti sono meno comuni e quindi il modello ha meno dati su cui addestrarsi, oppure potrebbe indicare la presenza di fattori influenti sul prezzo più difficili da catturare per gli immobili di fascia alta.

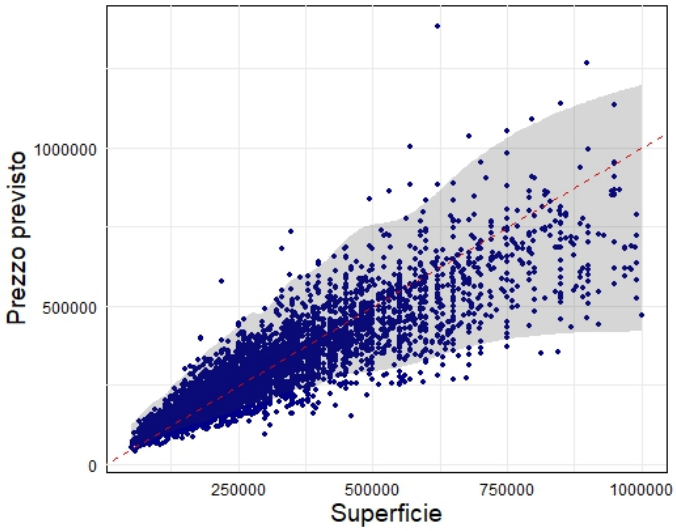


Figura 3.5: Inferenza conforme quantilica

Capitolo 4

Scostamento medio fra prezzo di acquisto e offerta di vendita

Il Capitolo 4 tratta una metodologia per stimare lo sconto applicato sulle offerte di vendite di alcuni immobili nell'area del comune di Padova. Il lavoro presentato è stato esposto in occasione del XLVIX Incontro di Studi del CeSET "IL RUOLO DEGLI INDICATORI SOCIO-ECONOMICO-AMBIENTALI NELLE POLITICHE E NELLE SCELTE DEGLI INVESTIMENTI PUBBLICI E PRIVATI".

4.1 Introduzione

Nella tradizione estimativa italiana una delle problematiche principali è costituita dall'accesso ai dati relativi ad atti di compravendita reali. Per quanto l'Agenzia delle Entrate abbia reso la procedura più snella con la creazione del portale web Sister, l'accesso agli atti risulta comunque oneroso per i professionisti ed i ricercatori e dispendioso in termini di tempo in quanto richiede l'analisi degli atti di compravendita in formato pdf. Una possibile fonte alternativa di dati al fine di acquisire dei comparables è costituita dagli annunci di vendita pubblicati dai portali specializzati online. Tali dati, per quanto facilmente accessibili, presentano delle criticità nel loro impiego a fini estimativi. Infatti, in un certo senso rappresentano le aspettative del venditore in termini economici, e pertanto si potrebbero discostare, per eccesso dall'effettivo prezzo di

vendita. Il loro impiego potrebbe dunque portare, da un punto di vista teorico, ad una sovrastima del valore di mercato del bene. Il loro impiego può essere giustificato qualora vengano impiegati per la creazione di scale di merito o coefficienti di differenziazione, ma non qualora vengano impiegati nell'ambito di procedure di stima dirette per l'elicitazione di valori di mercato. Risulta pertanto importante, data la facilità di accesso a tali dati, comprendere il loro scostamento dai valori di compravendita reali. A tal riguardo, il presente studio mira a stimare lo scostamento dei valori degli annunci immobiliari rispetto ai valori di mercato reali. A tal riguardo sono stati impiegati 2 database distinti. Il primo riguarda N=5660 comparables ottenuti da offerte di vendita relative all'anno 2022 e al primo trimestre 2023 relativi ad appartamenti del comune di Padova, mentre il secondo è costituito dall'analisi di N=281 atti di compravendita estratti dal database Sister dell'Agenzia delle Entrate, sempre relativi ad appartamenti del comune di Padova. Da un punto di vista metodologico, sono stati applicati algoritmi di machine learning per identificare i principali fattori che influenzano il suddetto gap. Per quanto riguarda i parametri presi in esame, particolare rilievo hanno assunto la posizione geografica, le caratteristiche delle proprietà, e altri fattori rilevanti.La metodologia proposta consiste in una modellazione in due step dove in un primo momento si stimano in cross-validation vari modelli sui dati relativi agli annunci, per ciascuno viene selezionato l'iperparametro che minimizza il MAPE (mean absolute percentage error) ed infine viene considerato solamente il modello migliore. Successivamente utilizzo il modello ottenuto per predirre, utilizzando come predittori quelli del dataset sugli acquisti. Tale metodologia ha consentito di ottenere per ciascun immobile un prezzo di vendita e un prezzo dell'annuncio, così da poter misurare il gap.

4.2 Dataset Compravendite

I dataset utilizzati per questa analisi sono:

- **Dataset offerte di vendita**: Offerta di vendita per ciascun immobile
- **Dataset compravendita**: Valore di acquisto effettivo dell'immobile

I dati per il valore di acquisto dell'immobile sono caratterizzati da 9 features e 281 osservazioni relativi all'area del comune di Padova. Di conseguenza, per quanto concerne il database delle offerte di vendita, vengono considerati solamente gli immobili afferenti all'area padovana e le variabili presenti nel database di compravendita.

Analisi esplorativa

La figura 4.1 mostra la heatmap delle offerte di vendita all'interno del comune di padova.

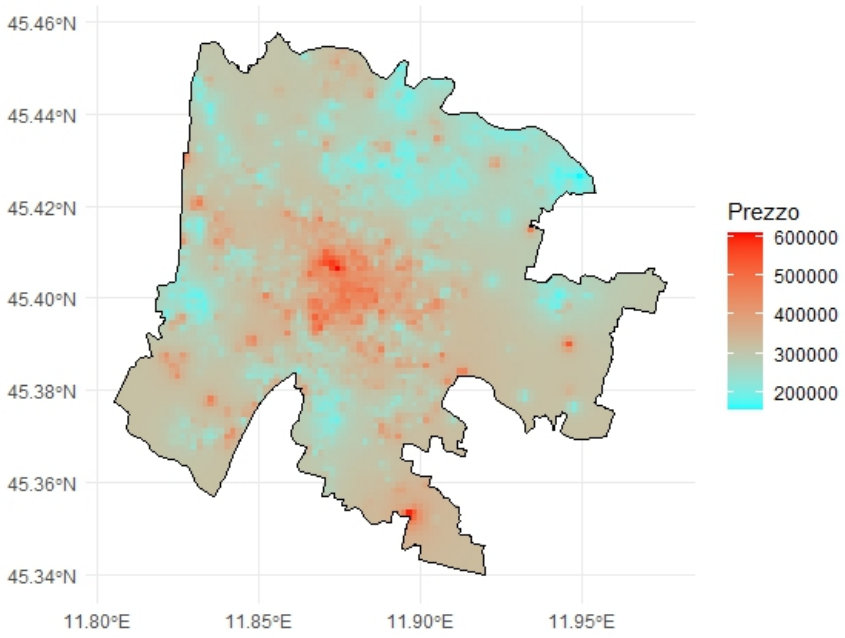


Figura 4.1: Mappa offerte di vendita nel comune di Padova

La mappa mostra come le zone con un'offerta di vendita più alta siano il centro della città, la zona industriale e il quartiere Guizza (zona sud).

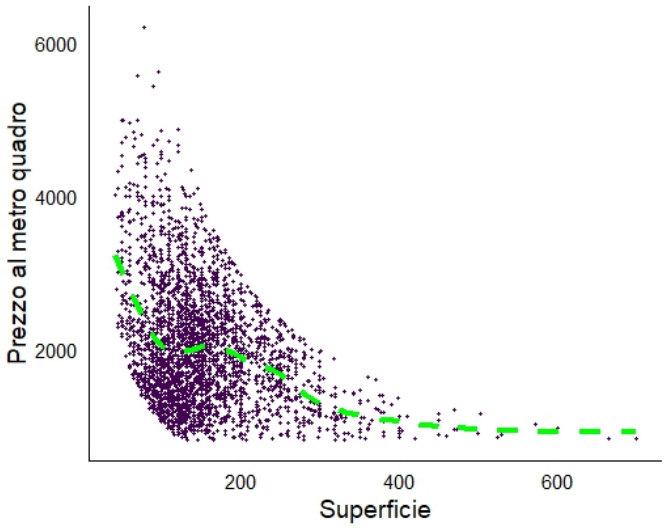


Figura 4.2: Scatterplot del prezzo al metro quadro contro superficie nel comune di Padova

La figura 4.2 suggerisce che il prezzo al metro quadro diminuisca esponenzialmente all'aumentare della superficie del locale.

4.3 Metodologia

La metodologia proposta consiste in una procedura a più step così definita:

(a) Stima di vari modelli in convalida incrociata sul dataset di compravendita.

(b) Selezione del modello migliore.

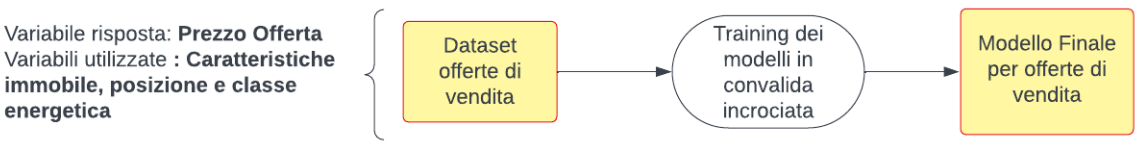


Figura 4.3: Primo step dell'analisi

(c) Previsione dell'offerta di vendita degli immobili facenti parte del dataset di compravendita.

(d) Stima dello sconto previsto applicato a ciascun immobile.

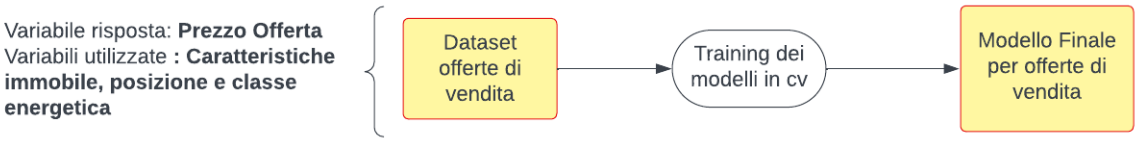


Figura 4.4: Secondo step dell'analisi

4.4 Risultati

Il modello predittivo ottimale identificato è l'Extreme Gradient Boosting (XGBoost), sulla base del più basso Mean Absolute Percentage Error (MAPE). L'analisi di importanza delle variabili in figura 4.5 evidenzia ubicazione, metratura e anno di costruzione come principali fattori di prezzo.

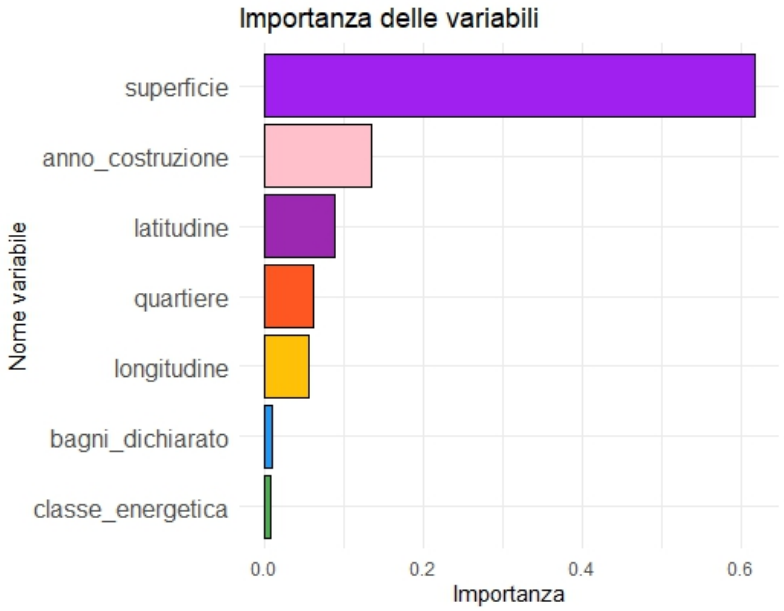


Figura 4.5: Importanza delle variabili

Viene esaminato il divario tra prezzo stimato e reale di vendita. Gli im-

mobili al di sotto dei 300.000€ ricevono in media uno sconto del 16,7%, indicando come gli immobili a prezzo più basso ottengano maggiori riduzioni. Sono riconosciuti i limiti nel prevedere con accuratezza gli sconti per le proprietà più costose, dovuti alla scarsità di dati.

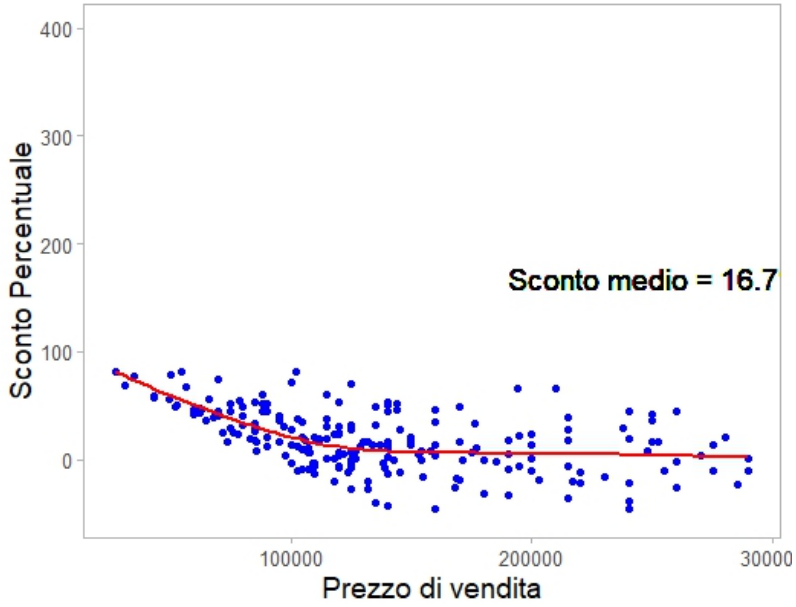


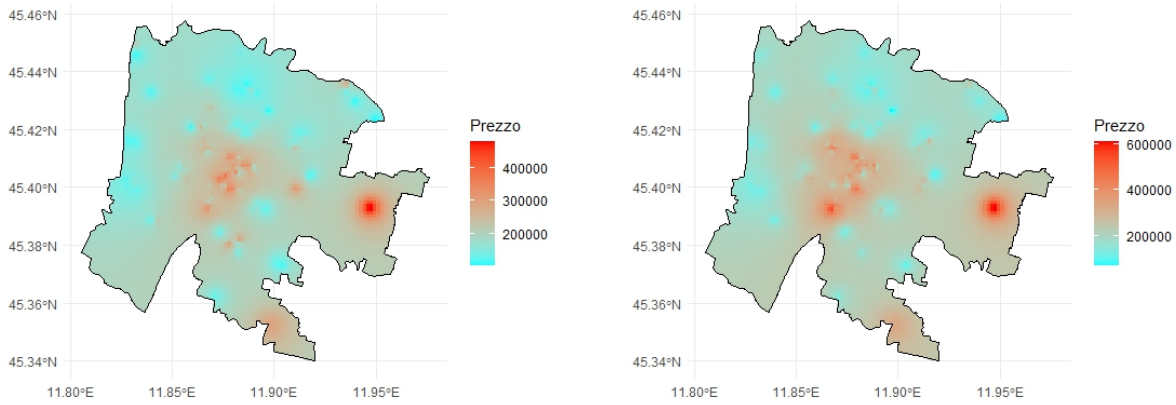
Figura 4.6: Sconto percentuale all'aumentare del prezzo

Le mappe presentatate in 4.7 rappresentano un'analisi termografica del territorio del Comune di Padova, in cui si visualizza la distribuzione spaziale del "Prezzo" attraverso un gradiente cromatico che varia dal blu al rosso. La legenda indica che il blu corrisponde ad un valore inferiore, situato intorno ai 200.000, mentre il rosso rappresenta un valore superiore che raggiunge i 600.000.

L'area geografica è chiaramente delimitata dalle coordinate latitudinali e longitudinali, che segnano i confini del territorio comunale. La presenza di concentrazioni calde, ossia zone in cui il colore vira verso il rosso, denota delle aree dove il "Prezzo" raggiunge i livelli massimi nell'ambito della scala considerata. Queste zone di maggiore intensità potrebbero corrispondere a quartieri o settori del comune caratterizzati da un più elevato valore immobiliare, commerciale o di altro tipo economico.

La distribuzione eterogenea dei valori suggerisce l'esistenza di una variabilità socio-economica all'interno del comune, con particolari aree che si distinguono per un maggiore prezzo di mercato rispetto al contesto circostante. Si può ipotizzare che tali zone corrispondano a centri di maggiore attività economica o a quartieri residenziali di alto standing.

In sintesi, la mappe mostrano come il modello adattato tenga in considerazione correttamente della posizione geografica dell'immobile e di come gli sconti più sostanziosi siano riservati agli immobili presenti nel centro della città.



Offerta di vendita osservata

Prezzo di acquisto stimato

Figura 4.7: Offerta di vendita stimata e prezzo di acquisto.

Il grafico a box plot presentato in figura 4.8 visualizza la distribuzione degli sconti percentuali applicati alle transazioni immobiliari in vari quartieri del comune di Padova. Ciascun box plot rappresenta un quartiere differente, e l'analisi comparativa di questi permette di discernere le differenze nei comportamenti di prezzo tra le diverse località.

L'asse delle ordinate elenca i quartieri, mentre l'asse delle ascisse quantifica lo sconto percentuale. Lo sconto è definito come la riduzione percentuale applicata al prezzo di listino dell'immobile durante la negoziazione di vendita. Dall'analisi del grafico , si può osservare che la distribuzione degli sconti varia notevolmente tra i quartieri. Alcuni quartieri mostrano

una distribuzione di sconti con mediana vicina allo zero, il che implica che la maggior parte delle transazioni in tali aree si è conclusa con un prezzo di vendita vicino al prezzo di listino. Altri quartieri presentano una mediana negativa, suggerendo che gli sconti sono comuni in queste aree.

Inoltre, l'ampiezza dei box plot e la lunghezza dei 'whiskers' riflettono la varianza negli sconti concessi, con alcune aree che mostrano una maggiore eterogeneità nelle negoziazioni rispetto ad altre. La presenza di valori anomali in alcuni quartieri indica la possibilità di sconti significativamente più alti o più bassi rispetto alla norma per quelle aree.

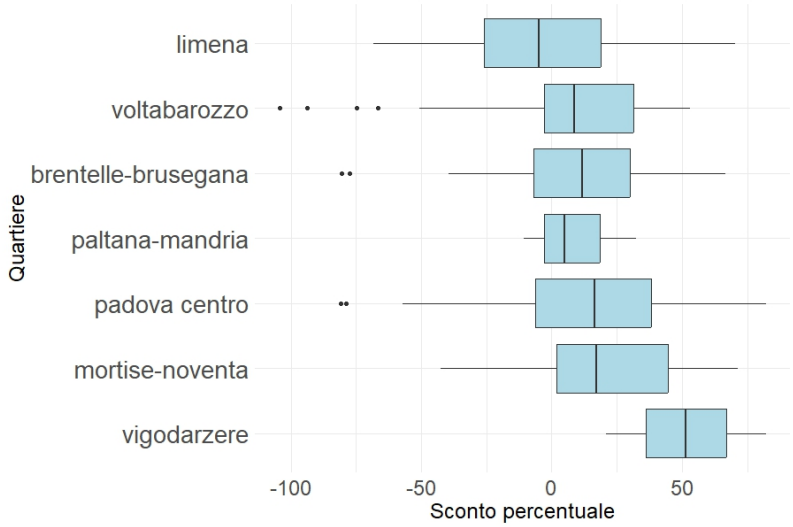


Figura 4.8: Sconto percentuale per quartiere

Capitolo 5

Conclusioni

Il presente lavoro di tesi si proponeva di conseguire due obiettivi: identificare il modello di machine learning più performante nella previsione dei prezzi degli immobili residenziali nella regione Veneto, e stimare in modo approfondito lo scostamento medio tra valori annunciati e prezzi effettivi di acquisto per le proprietà ubicate nel comune di Padova.

Per quanto riguarda il primo quesito, la sperimentazione condotta su oltre 70.000 inserzioni immobiliari ha consentito di valutare e confrontare le prestazioni di numerosi algoritmi di regressione, sia interpretabili che basati su tecniche di "deep learning". I risultati hanno evidenziato come approcci di ensemble learning come XGBoost e Random Forest riescano ad ottenere le migliori performance predittive, con errori inferiori anche del 35% rispetto a modelli lineari o alberi decisionali. L'analisi delle variabili di importanza e dei valori di Shapley ha inoltre fornito una visione più approfondita delle relazioni sottese ai prezzi, agevolando la comprensione del problema. Successivamente sono state proposte tre metodologie per la Conformal Inference, ossia Split Conformal Inference, Local Weighted Conformal Inference e Conformalized Quantile Prediction. Il metodo conformale suddiviso si è dimostrato in grado di generare stime probabilistiche coerenti con i principi teorici sottostanti, fornendo una misura dell'incertezza affidabile. L'introduzione della weighted conformal inference ha migliorato ulteriormente le prestazioni in presenza di eteroschedasticità. Inoltre, la conformal regression quantilica ha evidenziato il proprio potenziale nel produrre intervalli predittivi adattati alle

caratteristiche locali dei dati, facendo emergere questa metodologia come possibilmente ottimale.

Per quanto concerne la stima del gap annuncio-prezzo di acquisto sul territorio padovano, l'indagine su 281 atti notarili ha consentito di quantificare in modo attendibile il divario mediano, evidenziando come gli immobili al di sotto dei 300.000€ ricevano mediamente sconti del 16,7%. Inoltre, l'analisi a livello sub-comunale ha permesso di far emergere alcune differenze nei comportamenti di mercato a livello locale.

Bibliografia

Azzalini, Adelchi e Bruno Scarpa (2012). *Data analysis and data mining: An introduction*. OUP USA.

Chen, Tianqi e Carlos Guestrin (2016). «Xgboost: A scalable tree boosting system». *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

De Mol, Christine, Ernesto De Vito e Lorenzo Rosasco (2009). Elastic-net regularization in learning theory. *Journal of Complexity* 25.2, pp. 201–230.

Lei, Jing et al. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association* 113.523, pp. 1094–1111.

Lundberg, Scott M e Su-In Lee (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30.

Moussalli, Roger, Mudhakar Srivatsa e Sameh Asaad (2015). «Fast and flexible conversion of geohash codes to and from latitude/longitude coordinates». *2015 IEEE 23rd annual international symposium on field-programmable custom computing machines*. IEEE, pp. 179–186.

Park, Byeonghwa e Jae Kwon Bae (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert systems with applications* 42.6, pp. 2928–2934.

Romano, Yaniv, Evan Patterson e Emmanuel Candes (2019). Conformalized quantile regression. *Advances in neural information processing systems* 32.

Shapley, Lloyd S et al. (1953). A value for n-person games.

Tibshirani, Ryan J et al. (2019). Conformal prediction under covariate shift. *Advances in neural information processing systems* 32.