CONFRONTO TRA MODELLI PREVISIVI PER IL MERCATO IMMOBILIARE: APPLICAZIONE AL CASO DEI PREZZI DEGLI IMMOBILI DI MADRID

Romeo Silvestri - 1944738

INTRODUZIONE

Il mercato immobiliare è una parte vitale dell'economia di un paese che si occupa della costruzione, della gestione e della compravendita di beni immobili che possono essere adibiti ad una moltitudine di scopi.

In questo contesto, la capacità di valutare e prevedere correttamente il prezzo di vendita degli immobili è diventata una priorità per i professionisti del settore e per gli investitori. Il prezzo è infatti spesso difficile da stabilire perché è influenzato da una serie di fattori sia interni che esterni all'immobile stesso, come la sua dimensione, lo stato di manutenzione, la posizione geografica e la situazione economica. Per questo motivo i modelli statistici sono un utile strumento a supporto delle finalità previsive.

Nell'ambito di questa tesi si comparano quindi i diversi modelli previsivi per il mercato immobiliare per stabilire il valore delle abitazioni nel modo più accurato possibile. In particolare, si presterà attenzione alla componente spaziale, poiché la posizione geografica dell'immobile è un elemento chiave nel definire il prezzo di vendita.

Si utilizzeranno sia modelli previsivi tradizionali che spaziali. I modelli regressivi tradizionali sono stati ampiamente impiegati in quest'ambito e tengono conto della collocazione geografica attraverso l'introduzione di apposite variabili. D'altra parte, i modelli spaziali incorporano le informazioni sulla vicinanza geografica delle osservazioni direttamente nel metodo di costruzione ed esprimono l'intensità di queste relazioni esplicitamente attraverso appositi parametri.
E' dunque interessante andare a studiare e confrontare questi due diversi approcci.

Un altro problema importante è rappresentato dalla gestione dei dati mancanti.
In questa tipologia di problemi, i dati mancanti sono spesso presenti in grandi quantità, soprattutto per quanto riguarda le informazioni sulle caratteristiche interne all'immobile. Il corretto trattamento dei dati mancanti è quindi parte integrante dell'analisi. Ciò comporta la necessità di utilizzare delle tecniche di imputazione adeguate a stimare i valori mancanti.

La tesi si concentrerà sullo sviluppo di un'applicazione per il mercato immobiliare di Madrid. Questa città rappresenta un'area geografica estremamente interessante con un settore vario, ampio e in costante evoluzione. Oltretutto, quest'area di studio è caratterizzata dalla disponibilità di dati di alta qualità provenienti dai principali portali immobiliari spagnoli. Tali immobili sono infatti descritti da numerose informazioni e, differentemente dai portali di altri paesi, viene spesso indicato l'indirizzo preciso dell'abitazione.

In sintesi, questo lavoro può rappresentare un contributo per migliorare la comprensione del mercato immobiliare e supportare le decisioni di investimento.

Dati Mancanti

I dati mancanti sono molto diffusi quando viene analizzato il mercato immobiliare. Spesso, infatti, i dati in questo settore sono gestiti da privati o da agenzie immobiliari che non sono a conoscenza di parte delle informazioni sulle proprietà. O, ancora, possono trascurare delle informazioni che sono in realtà determinanti al fine di associare un valore di vendita all'immobile.

- Tipologia di Dati Mancanti: per analizzare i dati mancanti è necessario comprendere il meccanismo che regola le probabilità che determinano la "missingness". Ciascun tipo di dato mancante ha infatti bisogno di un trattamento idoneo. Ci sono essenzialmente due metodi per categorizzare questi dati: uno basato su un sistema di classificazione e l'altro basato su una proprietà del meccanismo dei dati mancanti.
  - Classificazione di Rubin: MAR, MCAR e MNAR
  - Ignorabilità

  Per le osservazioni mancanti sulle variabili che definiscono le caratteristiche dell'immobile si è verificata l'appartenenza alla classe MAR e si è stabilita la proprietà di ignorabilità.

- Metodi di Gestione e di Imputazione
  - Metodi Naive: ignorano i dati mancanti, rimuovono le osservazioni incomplete o consistono in forme semplici di imputazione
  - Massima Verosimiglianza e Metodi Bayesiani: imputano implicitamente i dati e si basano sui soli dati disponibili
  - Regressione: imputano i dati tramite i modelli statistici di regressione
  - Matching: imputano i dati associando i valori tramite l'appoggio di unità simili

  Tra queste metodologie analizzate a livello teorico, è stato ritenuto più idoneo utilizzare sugli immobili di Madrid i metodi basati sull'imputazione esplicita dei dati in modo da stimare successivamente i modelli previsivi su un unico dataset completo.

- Imputazione Multipla: si ripete l'imputazione molteplici volte, con lo stesso metodo di stima, generando più dataset completi. L'approccio sottostante è quello bayesiano e prende in considerazione molteplici fonti di incertezza. I risultati delle analisi sui dataset completi vengono combinati tra loro. Questa tecnica è stata impiegata nella tesi come verifica della consistenza di un metodo di imputazione nella fase di selezione.

- Selezione del Metodo di Imputazione: confrontare i vari metodi di imputazione può risultare un problema a livello teorico non indifferente. Non è infatti possibile conoscere la controparte reale dei valori mancanti. Una proprietà desiderabile è la capacità di preservare le distribuzioni dei dati. Si mette in risalto un criterio di selezione basato sugli Imputation Scores, dove si assegna un punteggio ai metodi imputativi in base alla capacità di riprodurre fedelmente le distribuzioni condizionate dei dati osservati. Tra i 6 diversi metodi di imputazione implementati tramite l'algoritmo con equazioni concatenate MICE, il Random Forest risulta il preferibile.

Modelli Previsivi

Per trattare la componente spaziale nei problemi relativi al mercato immobiliare, è possibile utilizzare dei modelli previsivi di regressione che non possiedono esplicitamente i parametri spaziali all'interno della formulazione. Si può invece tener conto della spazialità definendo delle nuove variabili. Ognuno dei modelli ha di base dei vantaggi e degli svantaggi che possono emergere in modo dipendente dal dataset che si va ad analizzare.

- Trattamento della Componente Spaziale: si introducono delle variabili che permettono di misurare l'intensità di alcuni effetti associati all'area geografica di appartenenza dell'immobile.
  - Variabili Spaziali: vanno a definire un particolare tipo di sottomercato spaziale determinando le zone secondo le quali i prezzi sono in relazione tra loro.
    I tre metodi si distinguono prendendo in considerazione una delle seguenti entità:
    - Aree Amministrative
    - Cluster Spaziali
    - Cluster LISA
  - Variabili di Distanza: indicano la lontananza degli immobili rispetto ai luoghi d'interesse. Hanno lo scopo di descrivere la relazione con i punti focali di una città, le comodità e le strutture che aiutano a rendere una zona ambita o da evitare. La vicinanza con i luoghi d'interesse comporterà a modificare il prezzo di un'immobile.

  Il dataset originale non presenta tali variabili e possiede soltanto l'indirizzo di locazione.
  Per questo motivo si sono utilizzati i software ArcGIS e OpenStreetMap per geolocalizzare gli indirizzi (trovare le coordinate geografiche) delle abitazioni e dei luoghi d'interesse di Madrid. Esempi di punti di interesse utilizzati sono i supermercati, le stazioni, i parchi e le scuole. Inoltre, sono state ricavate le distanze minime in linea d'aria.

- Regressione: si presentano gli aspetti teorici sull'analisi dei dati attraverso i modelli di regressione tradizionali che non prevedono l'impiego di parametri spaziali.
  - Regressione Parametrica (modello lineare, glm e regolarizzazione)
  - Regressione Non-Parametrica (KNN regressivo e MARS)
  - Regressione con Alberi (albero regressivo, random forest, AdaBoost e XG-Boost)

- Selezione dei Modelli: si sceglie il modello più accurato in base a degli indicatori di errore di previsione. Le stime vengono fatte attraverso k-fold cross-validation. Il processo si articola in genere nella selezione delle variabili esplicative e nel tuning degli iper-parametri.

Modelli Previsivi Spaziali

Gli effetti spaziali nei dati possono essere descritti attraverso l'esplicitazione di alcuni parametri appositi di natura spaziale. In particolare, supponendo di trovarsi in un contesto di omogeneità spaziale, i termini introdotti in questo lavoro vanno a regolare l'intensità dell'autocorrelazione spaziale nei dati. Il tipo di modello utilizzato è quello autoregressivo simultaneo (SAR) presentato in differenti sue declinazioni perché si è dimostrato efficace con una moderata mole di dati.

- Analisi Esplorativa Spaziale (ESDA): visualizzazione e sintesi dei dati dalla prospettiva spaziale, che aiuta a identificare pattern spaziali e suggerire potenziali modelli statistici da formulare ed applicare per l'inferenza.
  - Struttura di Vicinanza: definisce le relazioni di vicinanza tra le unità spaziali considerate. Questa struttura può essere rappresentata in diversi modi, a seconda del contesto e degli obiettivi dell'analisi. In generale, la struttura di vicinanza può essere definita in base alla distanza euclidea o geodetica tra le unità spaziali, oppure in base alla loro adiacenza topologica. Data una particolare struttura di vicinanza, ogni unità statistica sarà quindi associata ad una matrice dei pesi spaziali.
  - Indicatori di Autocorrelazione Spaziale: misurano una forma globale o locale di autocorrelazione spaziale e possono essere associati a dei test che verificano la presenza/assenza di associazione tra le osservazioni.

- Regressione Spaziale: si espongono i modelli spaziali SAR più comuni. Ciascuno di essi è caratterizzato da uno o più parametri spaziali.
  - Modello di Lag Spaziale
  - Modello di Lag Spaziale sulle X
  - Modello di Errore Spaziale
  - Modelli Spaziali Derivati

- Selezione del Modello Spaziale: si sceglie il modello spaziale utilizzando una strategia mirata che permette di evitare la specificazione di modelli computazionalmente troppo onerosi che verrebbero altrimenti scartati. Questa procedura fa uso dei test collegati agli indicatori di autocorrelazione spaziale, misurati sui residui dei modelli.

Risultati

Lo studio è stato condotto su 6287 immobili di Madrid, tra cui sono presenti appartamenti, attici e case indipendenti. Sono state applicate le tecniche esposte nella parte teorica nei seguenti punti:

- Descrizione del Dataset e delle Variabili
- Pre-Processing: gestione delle variabili e degli errori, dataset split e gestione dei dati mancanti
- Analisi Esplorativa: analisi univariata, bivariata e spaziale
- Regressione e Confronto tra i Modelli

l'XG-Boost che tiene conto dei cluster di tipo LISA come sottomercati spaziali è il modello che meglio prevede i prezzi di vendita degli immobili. Ad esso corrisponde un RMSE pari a 142.000 euro. In termini di accuratezza percentuale (R^2), il modello spiega l'89,9% della variazione nella variabile risposta.

In definitiva, i modelli spaziali non sembrano avere delle buone capacità previsive, ma solo descrittive. E' quindi in genere preferibile optare per un modello tradizionale basato sugli alberi, o, altrimenti, su un modello GLM per bilanciare interpretabilità a capacità previsive.