# House sales prediction

## Analisi di regressione e validazione dei modelli.

Stefano Luca Balu, Paola Bruno, Anna De Angeli, Raffaele Tornatora

Università degli Studi di Milano Bicocca.

ABSTRACT

Obiettivo di questa analisi è prevedere, attraverso una regressione, il prezzo di vendita delle case nella King County, tenendo in considerazione diverse variabili che fanno riferimento alle caratteristiche specifiche di ogni abitazione. Particolare attenzione è stata rivolta alla selezione degli attributi e alla validazione dei modelli sviluppati attraverso l'analisi dei residui.

Si propone, infine, un'alternativa al modello di regressione attraverso un approccio di classificazione multi-class.

INDICE



## INTRODUZIONE


Situata lungo il Puget Sound nello Stato di Washington nord-occidentale, la HMA (Seattle Housing Market Area) consiste nelle contee King e Snohomish. L'HMA è riconosciuta come centro per la progettazione e produzione aeronautica, oltre che per la progettazione software e il commercio, con la presenza di aziende rinomate in tutto il mondo fra le quali Microsoft Corporation e Amazon.

Secondo un rapporto del *Department of Housing and Urban Development of U.S., 1 Settembre 2015* in quest'area, le condizioni del mercato immobiliare di vendita, presentano un tasso di sfitto stimato dell'1,2%, in calo rispetto al 2,6% di aprile 2010.

Tale riduzione riflette l'aumento della domanda,

poiché le finanze delle famiglie e l'accesso al credito erano migliorate e gran parte dell'inventario in eccesso derivante dalla recente crisi di preclusione era stata assorbito.

Durante i 12 mesi terminati ad agosto 2015, le vendite di case nuove ed esistenti sono risultate 54.700, con un aumento del 14% rispetto ai 12 mesi precedenti; inoltre la media del prezzo di vendita è aumentata del 6%.

Questi effetti sul mercato immobiliare sono stati possibili grazie a condizioni economiche favorevoli, con il tasso di crescita del lavoro che ha superato la crescita nella nazione dal 2011. (1)

# 1 DATASET E PREPROCESSING

Lo studio è condotto sul dataset "House sales in King County, USA", pubblicato sulla piattaforma Kaggle.com. [1] Il dataset contiene i prezzi di 21613 case vendute tra maggio 2014 e maggio 2015, nella King County, una contea dello Stato di Washington, il cui capoluogo è Seattle.

## 1.1 TRASFORMAZIONE DELLE VARIABILI

Delle 21 variabili contenute nel dataset, alcune sono state trasformate ai fini dell'analisi di regressione.

Innanzitutto è stato trasformato il formato della data in YYYY-MM-DD. Sono poi state binarizzate le variabili "sqft_basement" e "yr_renovated" poiché abbiamo ritenuto che l'informazione rilevante fosse la presenza o meno di una cantina, per la prima, e l'avvenuta ristrutturazione o meno, per la seconda.

Le variabili categoriali di tipo ordinale "view", "condition" e "grade" sono state binarizzate.

A differenza delle prime due, per "grade" è stato necessario il raggruppamento in 4 categorie per rappresentare i livelli di costruzione e progettazione (basso, medio-basso, medio-alto e alto), prima di effettuare la binarizzazione.

Sono state eliminate le variabili "Date" e "Yr_built" poiché riassunte da un nuovo attributo "age", ottenuto come differenza tra le due.

Durante la fase di esplorazione si è notato che alcune case sono state vendute due volte nel periodo temporale di riferimento; la seconda volta ad un prezzo maggiore della prima, nonostante le caratteristiche ad esse associate fossero le medesime.

Usando come subset tali osservazioni, si può notare come il prezzo della seconda vendita aumenti in media di circa il 27% rispetto alla prima.

Per questo motivo abbiamo creato la variabile "sell_period" che può assumere quattro valori a seconda che la casa sia stata venduta nel periodo 1 (da 05-2014 a 07-2014), 2 (da 08-2014 a 10-2014), 3 (da 11-2014 a 01-2015), 4 (da 02-2015 a 05-2015).

## 1.2 OUTLIERS

Analizzando la distribuzione dei valori assunti dalle diverse variabili, abbiamo riscontrato la presenza di possibili valori anomali:

- Records che riportano valori di bagni e camere da letto pari a 0;
- Record con numero di camere pari a 33, valore non ragionevole se viene considerato lo spazio vivibile ad esso associato;

Queste osservazioni sono state eliminate, in quanto rappresentano solo lo 0,12% dei dati.

# 2. METODOLOGIE E PROBLEMI AFFRONTATI

## 2.1 FEATURE SELECTION

Per la stima dei parametri, un'assunzione fondamentale riguarda l'indipendenza lineare tra le variabili esplicative.

Se una variabile esplicativa dovesse essere linearmente dipendente dalle altre (collinearità perfetta), significherebbe che le informazioni in essa contenute sono in realtà già presenti nel dataset attraverso altre variabili. Pertanto

[1] Il dataset è disponibile al seguente link: https://www.kaggle.com/harlfoxem/housesalespredicti on

l'eliminazione di tale attributo non comporterebbe una perdita di informazione.

Oltre ad esaminare il valore della matrice di correlazione, anche la presenza di coefficienti di correlazione elevati (in valore assoluto) è sintomo di multicollinearità. Una tipologia di analisi statistica utilizzata per individuarne la presenza è l'insieme dei fattori di crescita della varianza (VIF).

Il VIF indica quanto una variabile esplicativa risulti spiegata dalle altre. Nel caso in cui il VIF risulti uguale a 1, tale variabile non è coinvolta in nessuna situazione di multicollinearità.
Non esiste un criterio universalmente riconosciuto per stabilire il valore soglia del VIF al di sopra del quale considerare la presenza di una forte multicollinearità, perciò in questo caso sono stati considerati come indicativi VIF superiori a 10.

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

Le variabili eliminate secondo questo criterio sono: "condition" e "sqft_above".

Il VIF per alcune categorie della variabile "grade" è risultato maggiore della soglia; nonostante questo le stesse non sono state eliminate poiché ai fini dell'analisi di regressione sarebbe risultato poco esplicativo tenerne solo alcune.

Infine è stata valutata la possibilità di calcolare l'importanza delle variabili selezionate con il VIF nel modello di regressione, attraverso la tecnica di Feature Ranking (applicata attraverso la funzione VarImp della libreria caret in R). In particolare tale funzione può essere applicata a diversi modelli, tra cui il modello lineare; in tal caso viene utilizzato il valore assoluto della statistica t per ogni parametro del modello.

Secondo questa tecnica le variabili eliminate sono: "sqft_lot", "floors", "sqft_lot15", "has renovated?" , "has_basement?" , "sell_ period" . Per quanto riguarda "sell_period", i coefficienti di regressione ad essa associati sono significativi, nel senso che la differenza tra le categorie è statisticamente significativa; tale variabile non è però risultata rilevante ai fini della costruzione del modello.

## 2.2 ANALISI DEI RESIDUI

Premessa: le ipotesi alla base del modello di regressione sono le seguenti:

- il modello deve essere appropriato;
- i termini di errore devono essere indipendenti;
- i termini di errore devono essere distribuiti approssimativamente come una Normale;
- i termini d'errore devono avere tutti la stessa varianza.

L'assunto di omoschedasticità (la varianza dei residui deve essere costante) è quello che viene violato più spesso. Le stime dei coefficienti continuano ad essere non distorte anche in presenza di eteroschedasticità; risulterà però sottostimata la varianza di tali stime e di conseguenza non saranno validi i test di ipotesi.

Queste violazioni delle ipotesi classiche sono diagnosticabili attraverso un grafico dei residui.

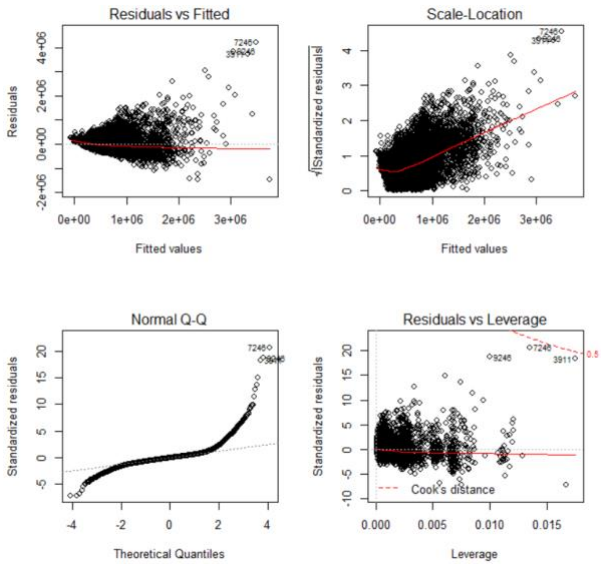


**Figura 1**

1. Il primo riquadro in alto a sinistra della fig.1 mostra gli errori residui contro i loro valori stimati. I residui devono essere distribuiti in modo casuale attorno alla linea orizzontale che rappresenta un errore residuo pari a 0, cioè non dovrebbe esserci una netta tendenza nella distribuzione dei punti.
2. Il riquadro in basso a sinistra della fig.1 è un plot QQ standard, ovvero confronta i valori dei residui standardizzati (quantile reale) rispetto la linea che individua la loro distribuzione normale (quantile teorico);

ovvero il grafico rappresenta una figura per cui se i punti si distribuiscono sulla linea, la distribuzione dei residui risulta normale e quindi la regressione rappresenta un modello adeguato.

3. Il riquadro in alto a destra della fig.1 mostra la radice quadrata dei residui standardizzati in funzione dei valori stimati. Anche in questo caso non ci deve essere alcuna tendenza evidente in questo grafico.
4. Il riquadro in basso a destra nella fig.1 mostra il valore di leverage dei punti che rappresenta una misura dell'importanza dell'osservazione nel determinare il risultato di regressione (forti deviazioni dalla tendenza nei punti iniziali o finali della serie hanno molto peso sul modello).
   I punti spostati a destra e in alto sono quelli che hanno peso maggiore.
   Sovrapposte al plot ci sono linee di contorno per la distanza di Cook, che è un'altra misura dell'importanza di ciascuna osservazione sulla regressione.

Valori di distanze di Cook bassi per un punto indicano che la rimozione della rispettiva osservazione ha poco effetto sui risultati della regressione, ovvero l'osservazione in particolare non ha valori devianti dalla tendenza. Invece valori di distanze di Cook superiori a 1 sono sospetti ed indicano la presenza di un possibile outlier. (2)

In questo caso nessuna osservazione sembra essere troppo importante nell'analisi.

Guardando la fig.1 come output della regressione possiamo notare che i residui non si distribuiscono casualmente intorno al valore 0 e che la loro distribuzione non risulti normale (riquadro in basso a sinistra della Fig. 1). Per questo motivo procediamo con la trasformazione del prezzo nel suo logaritmo.

Come si può notare dalla fig. 2 adesso le condizioni fondamentali per l'analisi di regressione vengono rispettate.

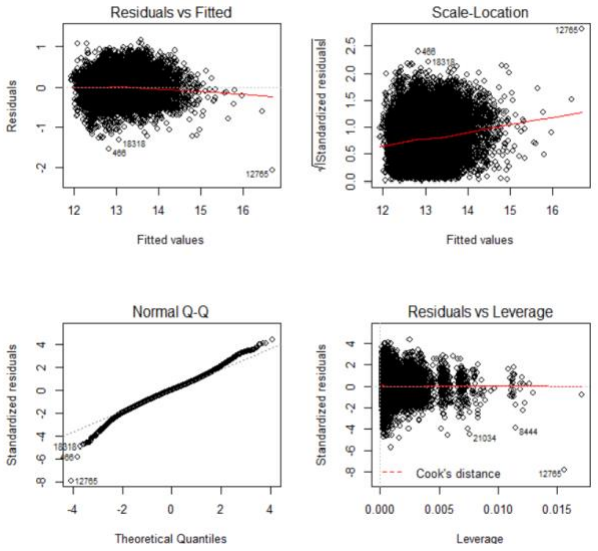


**Figura 2**

In questo caso, calcolando il VIF, le variabili eliminate sono: "Condition" e "sqft_above".

Per la variabile "grade" valgono le stesse considerazioni fatte in precedenza.

Anche in questo caso, è stata calcolata l'importanza delle variabili selezionate con il VIF attraverso il Feature Ranking. Adesso le variabili eliminate sono: "sqft_lot ", " sqft_lot15", "has renovated?", "has_basement?", "bedrooms" , "sell_ period" .

Per quanto riguarda "sell_period", valgono le stesse considerazioni fatte precedentemente.

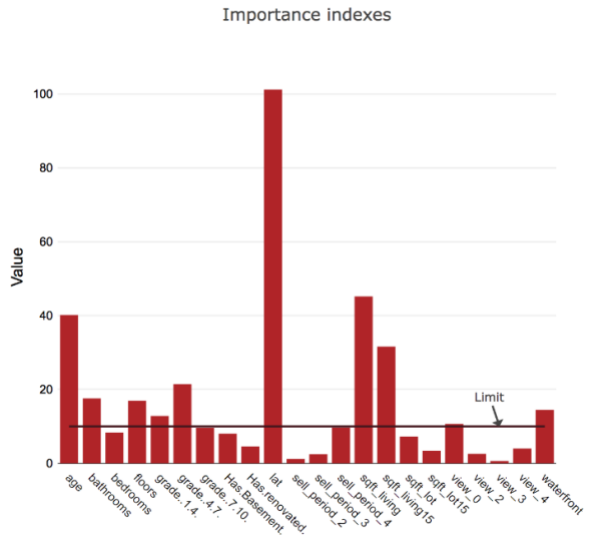


**Figura 3**

La fig.3 mostra l'importanza di ogni variabile nella regressione. La soglia al di sopra del quale considerare importanti le variabili è 10.

## 2.3 CLASSIFICAZIONE MULTI-CLASS

Si propone un approccio diverso, che consiste nello sviluppo di un problema di classificazione del tipo multi-class (one versus all). Tale analisi è motivata dalla necessità di ripartire le osservazioni in diverse fasce di prezzo, prevedendo quest'ultime in alternativa alla predizione attraverso il modello visto in precedenza. In questo modo sarà possibile discriminare tra abitazioni a basso, medio-basso, medio-alto e alto costo.

In aggiunta alle operazioni di preprocessing descritte in precedenza, è stato necessario procedere con la discretizzazione della variabile continua "price" in quattro classi sulla base dei quartili, in modo tale che ogni categoria abbia circa lo stesso numero di osservazioni.

Nonostante il livello di accuratezza più alto ottenuto tra i classificatori provati sia riconducibile al Gradient Boosting Trees, l'algoritmo di classificazione scelto in questa fase è il Random Forest. Il problema principale con il modello a gradiente è che nella fase di feature selection (wrapper) e di apprendimento si riscontra un forte rallentamento del processo.

Una foresta casuale è un classificatore d'insieme che è composto da molti alberi di decisione (nel nostro caso 10) e che dà in uscita la classe che corrisponde all'uscita delle classi degli alberi presi individualmente.

La validazione del classificatore è stata effettuata con la cross-validation attraverso la divisione del dataset in 10 fold.

Per quanto riguarda la feature selection, in questo caso si è scelto di effettuare la selezione degli attributi rilevanti in modo automatico utilizzando il metodo wrapper: nonostante sia computazionalmente costoso, permette di selezionare le variabili attraverso l'algoritmo di classificazione scelto per l'analisi mediante l'ottimizzazione di misure di performance quali accuracy ed errore. A tal proposito sono stati confrontati due metodi: Feature Elimination e Forward Feature Selection. La differenza tra i due risiede nel fatto che il primo consiste nell'addestramento del classificatore prima con tutti gli attributi e, con i cicli successivi, vengono rimosse le caratteristiche in modo tale che la performance del modello, in termini di errore, risulti migliorata. Nel secondo caso si applica il processo inverso: ad ogni ciclo vengono aggiunte variabili input su cui viene addestrato il modello di classificazione in modo tale che la sua performance, in termini di accuracy, risulti migliore della precedente. In entrambi i casi il processo di feature selection è ripetuto fino a quando non è osservato alcun miglioramento sulla rimozione/ aggiunta delle variabili esplicative.

In questa sede è stato selezionato il modello di Backward Feature Elimination in quanto permette una selezione di attributi che consente al classificatore di raggiungere un livello di accuracy più alta.

# 3. SOMMARIO DELLE ANALISI

## 3.1 DESCRIZIONE DEL FLUSSO KNIME

Il lavoro è stato svolto sostanzialmente in 3 fasi:

- Preprocessing
- Modelli di regressione
- Classificatore multi-class.

Nella fase di preprocessing è stata necessaria la trasformazione di alcune variabili (binarizzazione, aggregazione), come precedentemente descritto; la valutazione di alcune osservazioni come possibili outliers e la loro rimozione.

Terminata questa fase abbiamo valutato la possibilità di predire il prezzo attraverso analisi di regressione, con relativa rimozione di alcuni attributi, tenendo in considerazione diverse tecniche quali VIF e feature ranking.

Questo lavoro è svolto attraverso il nodo "R snipped" in quanto è stato preferibile effettuare questo tipo di analisi con "R", che fornisce strumenti più adeguati per le analisi statistiche,

in particolare per la regressione.[2]

L'output ottenuto dagli R snipped è un nuovo dataset privo delle variabili scartate tramite la feature selection.

A questo punto prediciamo il prezzo con il nodo "Gradient Boosted Regression Learner" e "Gradient Boosted Regression Predictor"; ne abbiamo accertato l'efficacia e l'accuratezza attraverso "Cross Validation", con 10 validazioni.

Il gradient boosting tradotto letteralmente "potenziamento del gradiente" è una tecnica di machine learning di regressione e problemi di Classificazione statistica che producono un modello predittivo nella forma di un insieme di modelli più deboli, tipicamente alberi di decisione. Costruisce un modello in maniera simile ai metodi di boosting, e li generalizza permettendo l'ottimizzazione di una funzione di perdita differenziabile arbitraria. (3)

Questa operazione è stata ripetuta trasformando il prezzo nel suo logaritmo, per i motivi già descritti in precedenza. Infine sono stati confrontati i risultati ottenuti.

Per quanto riguarda il problema di classificazione del tipo multi-class, gli algoritmi utilizzati sono il Gradient Boosting Trees e il Random Forest. Anche in questo caso la validazione del classificatore è stata effettuata con la cross-validation attraverso la divisione del dataset in 10 fold.

Per la feature selection, si è scelto di effettuare la selezione degli attributi rilevanti in modo automatico utilizzando il metodo wrapper, in particolare la Backward Feature Selection.

# 4. RISULTATI

## 4.1 RISULTATI ANALISI DI REGRESSIONE

La misura utilizzata per valutare i vari modelli di regressione è l'$R^2$; in particolare abbiamo utilizzato tale coefficiente di determinazione per confrontare il modello di stima del logaritmo del prezzo con quello del prezzo reale.

Avendo un range di valori molto ampio, il logaritmo del prezzo è stato utile per avere una distribuzione più uniforme e di conseguenza un coefficiente di determinazione più elevato. In particolare nel modello di predizione del prezzo reale il risultato ottenuto è il seguente:

| | |
|---|---|
| $R^2$: | 0,813 |
| Mean absolute error: | 87.434,394 |
| Mean squared error: | 25.231.094.319,163 |
| Root mean squared error: | 158.842,986 |
| Mean signed difference: | -8.845,44 |

Mentre nel modello di predizione del logaritmo del prezzo:

| | |
|---|---|
| $R^2$: | 0,845 |
| Mean absolute error: | 0,152 |
| Mean squared error: | 0,043 |
| Root mean squared error: | 0,207 |
| Mean signed difference: | 0,001 |

Applicando l'esponenziale al logaritmo del prezzo, otteniamo una nuova stima del prezzo reale con i seguenti risultati:

| | |
|---|---|
| $R^2$: | 0,82 |
| Mean absolute error: | 85.989,207 |
| Mean squared error: | 24.229.488.049,036 |
| Root mean squared error: | 155.658,241 |
| Mean signed difference: | -14.091,125 |

## 4.2 RISULTATI CLASSIFICAZIONE MULTI-CLASS

Nonostante il livello di accuratezza più alto ottenuto tra i classificatori provati sia riconducibile al Gradient Boosting Trees, l'algoritmo di classificazione scelto in questa fase è il Random Forest, per i motivi spiegati in precedenza. Confrontando i metodi per la selezione degli attributi è stata scelta la tecnica di Backward Feature Elimination.

| Row ID | D Acc |
|---|---|
| RF (forward) | 0.743 |
| RF (backward) | 0.746 |
| GB (forward) | 0.763 |
| GB (backward) | 0.766 |
| NB (backward) | 0.615 |
| NB (forward) | 0.704 |

[2] Per eseguire il nodo "R snipped" è necessario installare le seguenti librerie sul proprio software R: caret, pacman e mlbench.

La Random forest è abbastanza performante per quanto riguarda accuracy ed errore e permette di ottenere valori di AUC che si avvicinano a quelli ottenuti dal Gradient Boosted Trees: le curve Roc per le diverse classi di prezzo risultano "più alte" rispetto a quelle ottenute per altri classificatori (Naive Bayes). I valori ottenuti sono:

- price [78,000; 322,000]: AUC→ 0.956
- price (322,000; 450,000]: AUC → 0.873
- price (450,000; 645,000]: AUC → 0.882
- price (645,000; 7,700,000]: AUC → 0.96

Nelle figure 4, 5, 6 e 7 sono mostrate le curve Roc riferite alle quattro classi precedentemente discusse per il modello di classificazione Random Forest (backword).

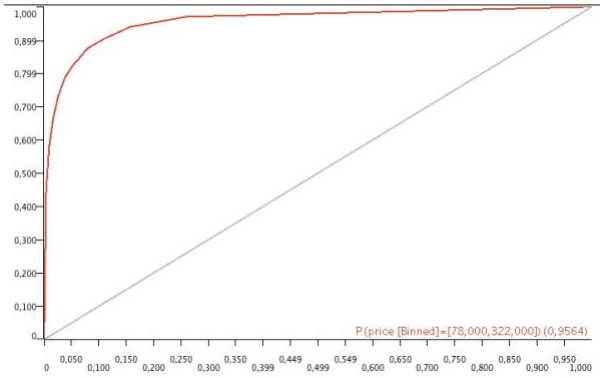


**Figura 4**

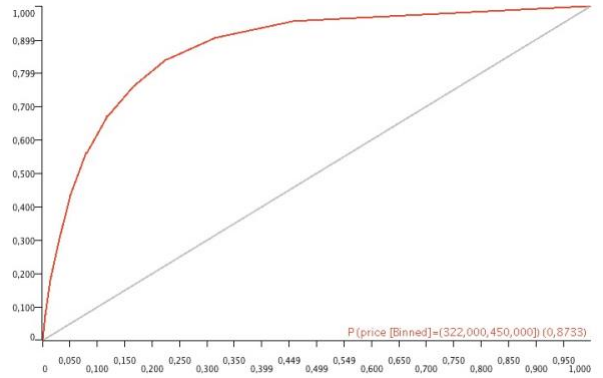


**Figura 5**

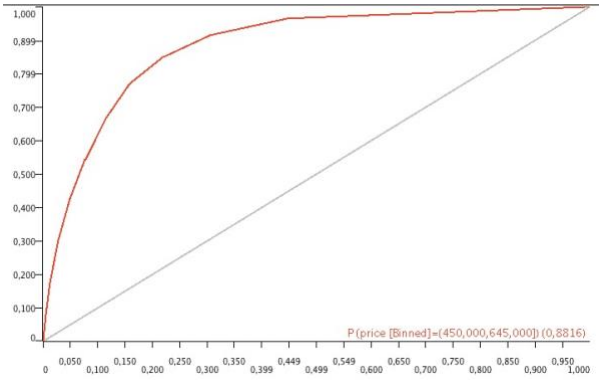


**Figura 6**

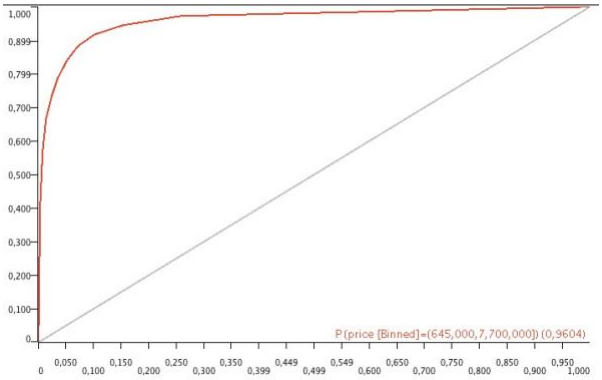


**Figura 7**

## 5. CONCLUSIONI

Nell'analisi presentata, l'obiettivo è quello di sviluppare un modello di previsione del prezzo per le case della contea di King, Washington. Il focus dell'elaborato consiste nell'analisi di regressione: con un coefficiente di determinazione dell'82% affermiamo che gli attributi esplicativi, scelti attraverso il meccanismo di feature selection, sono dei buoni regressori per predire il valore della variabile target, "price". In alternativa a tale metodo, è stato sviluppato un modello di classificazione multi-class al fine di predire la fascia di prezzo di appartenenza di ciascuna abitazione.

L'elaborato mostra come problemi di Machine Learning possano essere risolti in modi diversi. In questo caso l'obiettivo perseguito è la previsione del valore esatto del prezzo e, per questo motivo, il modello di regressione sviluppato rappresenta il punto focale dell'analisi. La classificazione multi-class è una proposta alternativa che è stata trattata al fine

di mostrare come sia possibile distinguere abitazioni appartenenti a diverse fasce di prezzo, rispondendo all'esigenza di discriminare tra case a basso, medio basso, medio alto o alto costo.

È importante notare che gli attributi a disposizione nel dataset "House sales in King County, USA", fanno riferimento esclusivamente alle caratteristiche specifiche delle abitazioni. Il prezzo delle case, però, oltre a dipendere dalle variabili disponibili, è molto influenzato anche da altri fattori di tipo economico come l'andamento del mercato immobiliare, la disponibilità di accesso al credito, l'andamento dell'economia nazionale, il potere d'acquisto e così via. La disponibilità di tali variabili aggiuntive potrebbe migliorare in modo significativo il modello elaborato.

**Possibili sviluppi futuri**

Nella disciplina estimativa, essendo i prezzi il fondamento di ogni giudizio di stima, particolare importanza riveste l'epoca a cui si riferisce la stima che può essere riferita ai tempi, in particolare:

- presente: stima attuale
- passato: stima retrospettiva
- futura: stima prospettiva

La metodologia estimativa trova come unico fondamento logico della valutazione la comparazione tra il bene oggetto di stima e altri beni di prezzo (o costo) noto con caratteristiche simili presi a confronto.
La comparazione dei beni di confronto, in relazione ai diversi gradi di eguaglianza o diseguaglianza, può essere la seguente:

- beni uguali
- beni simili
- beni intermedi
- beni dissimili
- beni diseguali (4)

In un possibile studio futuro si potrebbe classificare le abitazioni in base a queste comparazioni.
È importante chiarire che, quando parliamo di immobile comparabile, definiamo un bene appartenente allo stesso segmento di mercato; ne deriva l'importanza di un'accurata identificazione del particolare segmento.
Si dovrà quindi tener conto della destinazione d'uso, tipologia e localizzazione del bene.
Il principio di comparazione si può concretizzare tramite il confronto sulla base di prezzi o valori concorrenti (confronto attuale) o passati (confronto intertemporale).
Può accadere che nel segmento di mercato di riferimento le transazioni siano poche, non rilevabili e/o inattendibili per una serie innumerevole di circostanze ricorrenti nel tempo. (5)

Per questo motivo potrebbe essere interessante sviluppare un modello che stimi il prezzo attuale attraverso un confronto intertemporale, per poi essere comparato con il modello di stima ottenuto con un confronto attuale, che è stato oggetto di questo studio.

## 6. ELENCO E DESCRIZIONE DELLE VARIABILI:

- "Date": Data di vendita delle abitazioni;
- "bedrooms": numero di camere;
- "bathrooms": numero di bagni; la presenza di frazioni significa:
  - ✓ 3/4 bagno: doccia, lavello, toilette,
  - ✓ 1/2 bagno: lavello, toilette,
  - ✓ 1/4 bagno: toilette;
- "sqft_living": piedi quadri delle abitazioni;
- "sqft_lot": piedi quadri dello spazio esterno;
- "Floors": numero di piani;
- "Waterfront": variabile dummy che indica la presenza di vista sul lago o meno;
- "View": può assumere valori da 0 a 4 e indica la bontà della vista dall'abitazione;
- "Condition": può assumere valori da 1 a 5 e indica le condizioni dell'appartamento;
- "grade": può assumere valori da 1 a 13 e indica il livello di costruzione e progettazione dell'abitazione;
- "sqft_above": piedi quadri dei piani al di sopra del terreno;
- "sqft_basement": piedi quadri dei piani al di sotto del terreno;
- "yr_built": anno di costruzione dell'abitazione;
- "yr_renovated": anno di ristrutturazione dell'abitazione;
- "sqft_living15": media dei piedi quadri di spazio interno vivibile delle 15 abitazioni più vicine;
- "sqft_lot15": media dei piedi quadri dello spazio esterno delle 15 abitazioni più vicine

- “zipcode”;
- “lat” e “long”: latitudine e longitudine.

## RIFERIMENTI


1. huduser.gov. *huduser.gov.* [Online] 01 09 2015. https://www.huduser.gov/portal/publications/pdf/SeattleWA-comp-16.pdf.

2. [Online] file:///C:/Users/PC-HP/Downloads/930001934Analisigraficare%20(1).pdf.

3. [Online] https://it.wikipedia.org/wiki/Gradient_boosting.

4. Moncelli, Massimo. *Il tecnico estimatore nell'esecuzione immobiliaree nelle procedure concorsuali.* s.l. : Maggioli, 2014.

5. monitorimmobiliare.it. *www.monitorimmobiliare.it.* [Online] http://www.monitorimmobiliare.it/gli-asking-price-nelle-valutazioni-immobiliari_20178151151.