

PROGETTO INGEGNERIA DELLA CONOSCENZA

A.A. 2022/23

House PRICE PREDICTION- DOCUMENTAZIONE

22/06/2023

Bernardino Cristallo - Matricola: 676879

b.cristallo3@studenti.uniba.it

Lorenzo Levanto – Matricola: 735404

l.levanto@studenti.uniba.it

INDICE DEI CONTENUTI

- 01. INTRODUZIONE
- 02. DETTAGLI E SCELTE IMPLEMENTATIVE
- 03. GUIDA ALL' UTILIZZO

01. INTRODUZIONE

House Price Prediction è un progetto software che si pone come obiettivo la predizione dei prezzi delle case americane. Date in input le caratteristiche della casa, il programma fornisce una predizione del prezzo.

02. DETTAGLI E SCELTE IMPLEMENTATIVE

Sviluppato in Python 3.7

Librerie utilizzate:

- Pandas
- Numpy
- Sklearn
- Matplotlib
- Tqdm
- Tkinter

Si è utilizzato un dataset estrapolato dal sito [kaggle.com](https://www.kaggle.com), chiamato "data.csv", il file contiene i campi **'date'** (data di registrazione), **'price'** (prezzo), **'bedrooms'** (numero camere da letto), **'bathrooms'** (numero bagni), **'sqft_living'** (piedi quadrati vivibili), **'sqft_lot'** (piedi quadrati totali), **'floors'** (piani), **'waterfront'** (facciata su mare),

'condition' (condizioni), **'sqft_above'** (piedi quadrati calpestabili), **'sqft_basement'** (piedi quadrati seminterrato), **'yr_built'** (anno di costruzione), **'yr_renovated'** (anno di ristrutturazione), **'street'** (via), **'city'** (città), **'statezip'** (CAP), **'country'** (nazione).

Al dataset sono state uniti i 2 campi, **'bedrooms'** e **'bathrooms'** in un unico campo **'rooms'** (numero delle stanze), per diminuire il numero delle features e incrementare il peso delle stanze a livello di predizione.

Il software genera un file "house.csv" che contiene il dataset risultante dalle operazioni appena descritte. Le istruzioni che creano tale file sono commentate, in quanto non più utili una volta che si crea il file. Tornerebbero utili in caso di aggiornamenti dei dataset utilizzati.

Per avere dati più accurati abbiamo deciso di eliminare i duplicati e i campi con i valori nulli dal csv, inoltre abbiamo deciso di convertire i 'piedi quadri' in 'metri quadri' poiché il sistema metrico decimale in Italia è quello dei metri.

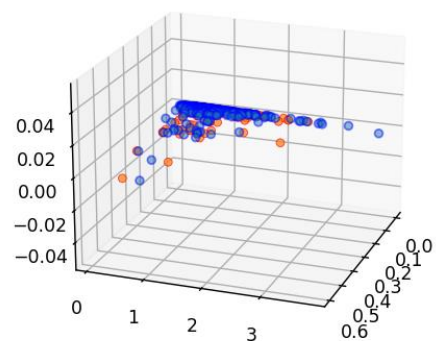
Per ottimizzare la predizione del prezzo abbiamo scalato i valori del dataframe tra 0 e 1 e prima dell'apprendimento supervisionato, viene utilizzato l'apprendimento non supervisionato con operazioni di clustering.

I modelli (di tipo random forest regressor) utilizzano come training-set una porzione (80%) del dataset, l'altra porzione è usata come test-set. La scelta di non utilizzare tutta la base di conoscenza come training-set è atta ad evitare l'overfitting e a valutare i modelli appresi utilizzando dati (test-set) non utilizzati per l'apprendimento. Successivamente vi è l'utilizzo del K-fold con cross-Validation per suddividere il training set in k fold e addestrare/valutare il modello su ogni fold.

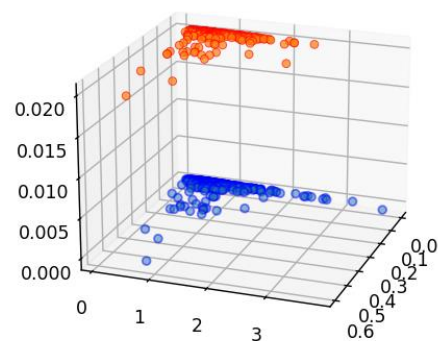
La valutazione della qualità dei modelli si basa sul calcolo del punteggio medio MAE attraverso la cross-validation (Mean Absolute Error '134422.63'), sul MAE('110438.91') calcolato tra i prezzi predetti per le case del test-set e quelli reali, e sul calcolo del r2-score ('0.74563' = 74.5%) in grado di calcolarci la percentuale di accuratezza delle predizioni.

Le features selezionate sono i campi **'sqft_living'**, **'sqft_lot'**, **'floors'**, **'waterfront'**, **'view'**, **'condition'**, **'sqft_above'**, **'sqft_basement'**, **'yr_built'**, **'yr_renovated'**, **'street'**, **'city'**, **'country'**, **'rooms'**, i campi **'date'** e **'statezip'** si sono mostrati irrilevanti per la qualità della predizione di Price.

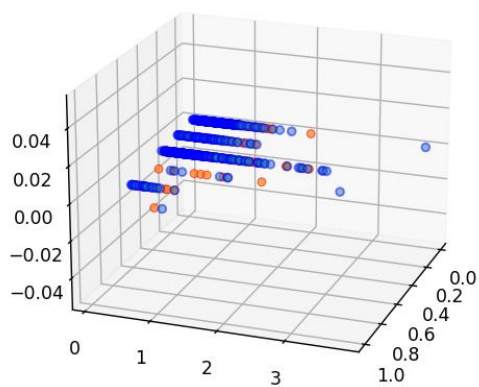
Per mostrare l'accuratezza della predizione per **ogni singola features** abbiamo inserito la stampa di grafici che dimostrano tale accuratezza (in blu i valori reali delle features e in rosso la loro predizione), le istruzioni che creano tale grafico sono commentate, in quanto non più utili una volta che si crea il grafico, di seguito qualche esempio, rispettivamente le seguenti features (sqft_living, floors, condition, yr_build, street):



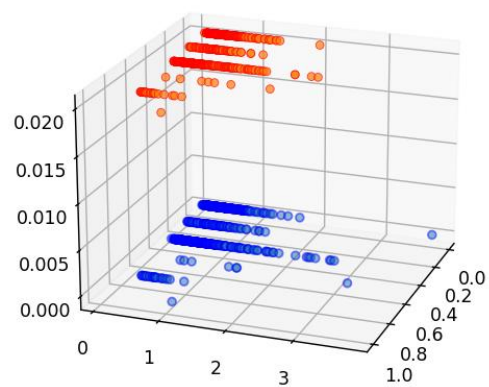
100



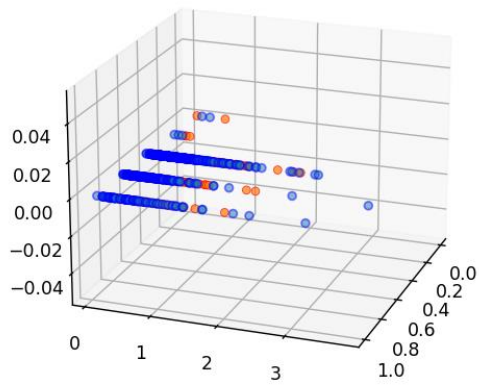
100



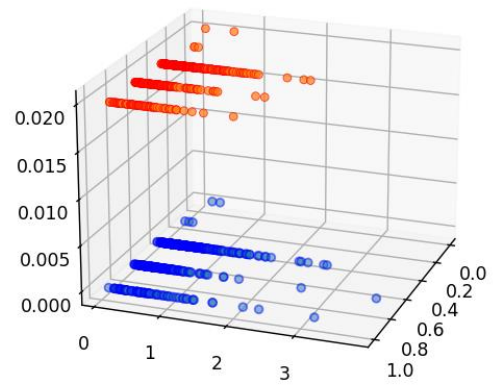
100



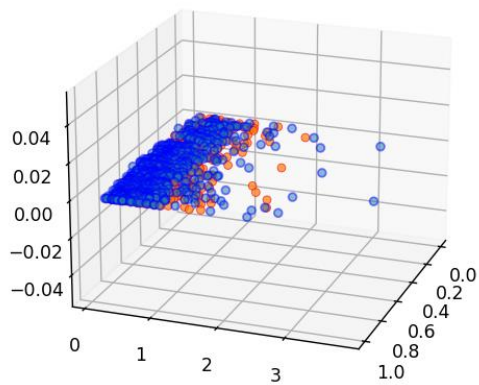
100



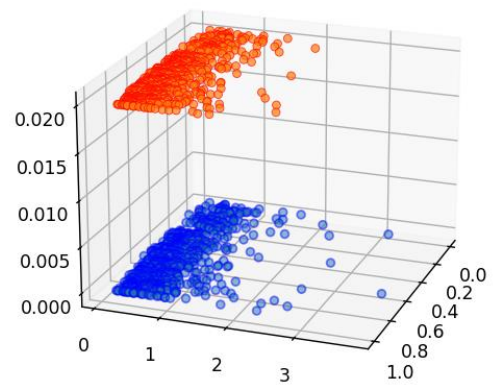
100



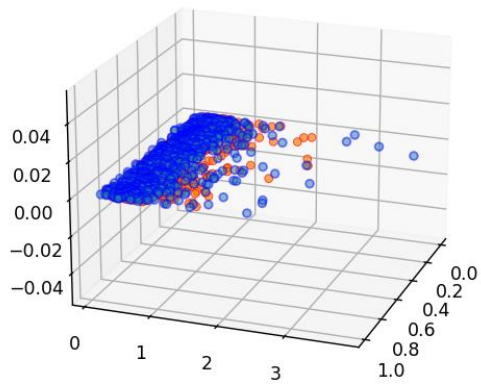
100



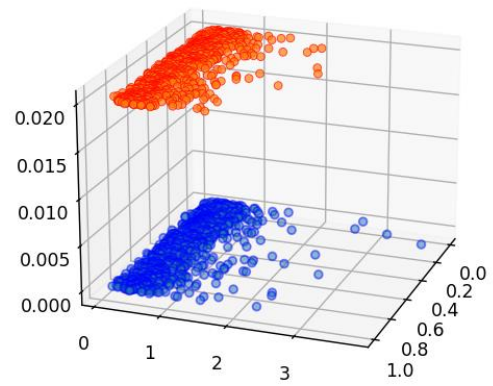
100



100



100



100

03. GUIDA ALL' UTILIZZO

Aprire i file 'model.py' e 'house_prediction.py', presenti nella cartella del progetto caricata sul repository, con un qualsiasi ambiente di programmazione (ad es. Visual Studio) che supporti Python (dalla versione 3.7).

Eseguire house_prediction, si visualizzerà tale interfaccia:

| Seleziona i valori | |
|-----------------------------------|--------------------|
| Nazione: | USA |
| Citta': | Algona |
| Via: | 1 View Ln NE |
| Metri quadri Vivibili: | 34.37383872166481 |
| Metri quadri Lotto: | 59.271646228167974 |
| Metri quadri Seminterrato: | 0.0 |
| Metri quadri Calpestabili: | 34.37383872166481 |
| Anno di costruzione: | 2014 |
| Anno di restauro: | 2014 |
| Piani: | 1.0 |
| Affaccio sul mare: | 0 |
| Stanze: | 2.0 |
| Vista: | 0 |
| Condizione: | 0 |
| <button>Avvia Predizione</button> | |

Selezionare i valori dei campi e premere 'avvia predizione' per visualizzare il prezzo predetto.

I campi dei metri quadri (ex. 'metri quadri vivibili', 'metri quadri lotto'...) sono gli unici campi che possono essere avvalorati senza dover per forza selezionare uno tra i valori presenti. I campi restanti vanno avvalorati scegliendo tra i valori disponibili, in modo da evitare che l'utente inserisca una nazione, regione, via ecc. inesistenti nel mondo reale.

Di seguito qualche screen di qualche predizione fatta usando i valori di alcune case reali presenti nel csv:

Valore casa su house.csv = 235000€

House Prediction in USA


Seleziona i valori

| | |
|----------------------------|--------------------|
| Nazione: | USA |
| Citta': | Seattle |
| Via: | 7542 21st Ave SW |
| Metri quadri Vivibili: | 112.41174284652546 |
| Metri quadri Lotto: | 873.2813080639169 |
| Metri quadri Seminterrato: | 0.0 |
| Metri quadri Calpestabili: | 112.41174284652546 |
| Anno di costruzione: | 1949 |
| Anno di restauro: | 0 |
| Piani: | 1.0 |
| Affaccio sul mare: | 0 |
| Stanze: | 4.00 |
| Vista: | 0 |
| Condizione: | 2.0 |

Avvia Predizione

Il prezzo predetto è: 273890.53

Valore casa su csv = 335000€

 House Prediction in USA


Seleziona i valori

| | |
|----------------------------|-------------------------------|
| Nazione: | <div>USA</div> |
| Citta': | <div>Redmond</div> |
| Via: | <div>2616 174th Ave NE</div> |
| Metri quadri Vivibili: | <div>125.4180602006689</div> |
| Metri quadri Lotto: | <div>237.82980304719436</div> |
| Metri quadri Seminterrato: | <div>0.0</div> |
| Metri quadri Calpestabili: | <div>125.4180602006689</div> |
| Anno di costruzione: | <div>1976</div> |
| Anno di restauro: | <div>0</div> |
| Piani: | <div>1.0</div> |
| Affaccio sul mare: | <div>0</div> |
| Stanze: | <div>4.00</div> |
| Vista: | <div>0</div> |
| Condizione: | <div>3.0</div> |

Avvia Predizione

Il prezzo predetto è: 379132.26

Valore casa su csv = 550000€

 House Prediction in USA

Seleziona i valori

| | |
|----------------------------|---|
| Nazione: | <input type="text" value="USA"/> |
| Citta': | <input type="text" value="Redmond"/> |
| Via: | <input type="text" value="9105 170th Ave NE"/> |
| Metri quadri Vivibili: | <input type="text" value="180.23039762170197"/> |
| Metri quadri Lotto: | <input type="text" value="975.4738015607581"/> |
| Metri quadri Seminterrato: | <input type="text" value="74.32181345224824"/> |
| Metri quadri Calpestabili: | <input type="text" value="105.90858416945375"/> |
| Anno di costruzione: | <input type="text" value="1976"/> |
| Anno di restauro: | <input type="text" value="1992"/> |
| Piani: | <input type="text" value="1.0"/> |
| Affaccio sul mare: | <input type="text" value="0"/> |
| Stanze: | <input type="text" value="6.50"/> |
| Vista: | <input type="text" value="0"/> |
| Condizione: | <input type="text" value="4.0"/> |

Il prezzo predetto è: 511205.50