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

Statistica e Gestione delle Imprese

RELAZIONE FINALE

MODELLO PREVISIVO DEI PREZZI DI OFFERTA DEGLI IMMOBILI

Relatore Prof. Finos Livio
Dipartimento di Psicologia dello Sviluppo e della Socializzazione

Correlatore Prof. Vecchiato Daniel
Dipartimento di Territorio e Sistemi Agro-Forestali

Laureando: Danaj Medi
Matricola N. 1050697

Anno Accademico: 2021/2022

Indice

1. Introduzione

1.1. Che cosa è e come funziona il mercato immobiliare?

Il mercato immobiliare si occupa della costruzione e della compravendita di immobili. È un mercato estremamente attraente, perché l'oggetto di contrattazione (abitazioni, terreni, locali commerciali, industriali, ecc.) ha un elevato valore economico e sociale.

Le società di costruzione sono i soggetti che si occupano, insieme alle società di architettura e di ingegneria, della realizzazione materiale degli immobili residenziali (case, ville, palazzine) e non residenziali (capannoni industriali, prefabbricati industriali, ecc.).

Spesso in questo tipo di mercato è presente una figura di elevata importanza, il mediatore, presente comunque anche in altri mercati. Il mediatore è colui che mette in relazione due o più persone per la conclusione di un affare senza essere legato ad alcuna di essa da rapporti di collaborazione, dipendenza o rappresentanza. In questo tipo di mercato ricopre dunque un ruolo molto importante, il suo aiuto viene richiesto nella maggior parte delle trattative. Il mediatore viene per questo definito in questo mercato: agente immobiliare.

1.2. Il mercato immobiliare italiano

Il mercato immobiliare è anche una forma d'investimento. La liquidità del bene oggetto di scambio (gli immobili) è facile e rapida. Si possono negoziare:

1. case, ville, palazzine uso abitazione e tutti gli immobili residenziali;
2. capannoni, uffici, laboratori, negozi e tutti gli immobili non residenziali e turistici.

La stima del valore dell'immobile invece viene eseguita tenendo conto:

1. collegamenti stradali, navali o aerei della zona di ubicazione;
2. presenza di aree verdi;
3. servizi per gli abitanti;
4. vicinanza ai luoghi strategici (lavoro, turistici, ecc.);
5. crescita demografica, invecchiamento, situazione socio-economica della popolazione contingente;
6. andamento dei tassi di interesse sui mutui relative alle operazioni di acquisto e manutenzione degli immobili.

La crescita della domanda di immobili residenziali può essere alimentata dal flusso di cittadini stranieri.

Importanti per le quotazioni immobiliari sono anche tutte le iniziative di recupero di aree dismesse e in generale di riqualificazione urbana. A beneficiare degli effetti di rivalutazione dei prezzi degli immobili non sono solo i quartieri direttamente interessati, ma anche le aree limitrofe.

2. Raccolta dei dati e data pre-processing

Il dataset su cui si è lavorato è stato fornito gentilmente dal Professor Vecchiato Daniel che ne detiene la proprietà intellettuale.

Tale dataset è stato ottenuto attraverso tecniche di *web scraping*, con uno script in *Python* messo a punto dal Professor Vecchiato Daniel stesso.

Esso aveva come target uno dei principali siti di annunci immobiliari italiano da cui dati in merito ad annunci di compravendita di immobili ad uso residenziale.

Tutti i dati provengono dal sito *immobiliare.it*.

I dati sono stati scaricati/raccolti il 14/02/22, alle 19:25:33.

Il dataset iniziale è composto di 12.315 unità statistiche, immobili situati per la maggior parte all'interno della regione Veneto, a cui si aggiunge la cittadina di Pordenone in Friuli-Venezia Giulia. Le variabili osservate, invece, sono in tutto 76.

In questo studio ci si occupa solo ed esclusivamente degli immobili ad uso residenziale.

Lo scopo di questa tesi è capire quali caratteristiche delle abitazioni influenzano maggiormente il loro prezzo di vendita.

Per farlo, ci affideremo sia ad un'analisi esplorativa, sia all'interpretazione di alcuni modelli previsivi.

Proveremo, pertanto, anche a prevedere il prezzo di un immobile, sulla base delle variabili osservate.

Per il lavoro in esame utilizzeremo il software *R-Studio* (https://www.rstudio.com/).

Dal prossimo capitolo si inizia a prendere in considerazione le variabili, approfondendole. Partiamo!

2.1. Comune

I comuni in totale – nel dataset iniziale – sono 11, di cui 10 situati nella regione Veneto ed uno (Pordenone) in Friuli-Venezia Giulia.

Vediamo le frequenze.

```
##                          N      %
## Bassano del Grappa     371   3.02
## Bolzano Vicentino        1   0.01
## Castelfranco Veneto    551   4.48
## Chioggia              1157   9.41
## Padova                2000  16.27
## Pordenone              335   2.73
## Rovigo                 626   5.09
## Treviso               1859  15.12
## Venezia               2000  16.27
## Verona                1974  16.06
## Vicenza               1417  11.53
## Totale               12291 100.00
## NA                      24   0.00
```

Tabella 1. Frequenze assolute e relative dei comuni. Prima della pulizia.

È stata rimossa l'unità nel comune di Bolzano Vicentino: ininfluente per la creazione del modello, lo avrebbe solo appesantito ulteriormente. Sono stati eliminati i 24 valori nulli (*NA, Not Available*).

Qui di seguito la tabella finale delle frequenze relative e assolute.

```
##                          N       %
## Bassano del Grappa     371    3.02
## Castelfranco Veneto    551    4.48
## Chioggia              1157    9.41
## Padova                2000   16.27
## Pordenone              335    2.73
## Rovigo                 626    5.09
## Treviso               1859   15.13
## Venezia               2000   16.27
## Verona                1974   16.06
## Vicenza               1417   11.53
## Totale               12290  100.00
## NA                       0    0.00
```

Tabella 2. Frequenze assolute e relative dei comuni. Dopo la pulizia.

Si rimane quindi con dieci comuni finali: non sono pochi, ma sarebbe una forzatura farci ulteriori modifiche.

2.2. Prezzo

Si inizia ora ad esaminare la nostra variabile d'interesse, ovvero il prezzo pubblicato sui vari annunci.

Si ricorda che sarà scopo di questa tesi concentrarsi solo sulle vendite degli immobili residenziali.

Si esplicita già ora che, per qualsiasi variabile che verrà in seguito, non sarà preso in considerazione ogni qual volta il dataset di partenza, ma le varie analisi riguarderanno il dataset nella forma più "aggiornata".

Si veda una sintesi iniziale.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1000  130000  225000  265904  350000  999999     749
```

Tabella 3. Sintesi della variabile prezzo, prima della pulizia.

Si decide di eliminare le 749 unità per cui il valore non è disponibile.

Sarebbe stato possibile anche non rimuoverle, per poi provare a verificare se ci sia una qualche motivazione principale che porti a non inserire il prezzo nell'inserzione.

Potrebbe essere uno spunto interessante per un futuro approfondimento a riguardo.

Si escludono anche le unità per cui i prezzi sono minori di € 10.000 (assolutamente irrealistiche per il prezzo di una casa). Si elimina anche il record per

cui il prezzo è € 999.999: verificandolo è un prezzo fittizio e assolutamente non verosimile.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11500  136000  230000  273814  360000  995000
```

Tabella 4. Sintesi della variabile prezzo, dopo la pulizia.

La distribuzione della variabile è fortemente asimmetrica verso i prezzi più alti. Come si può notare dal box plot di seguito.

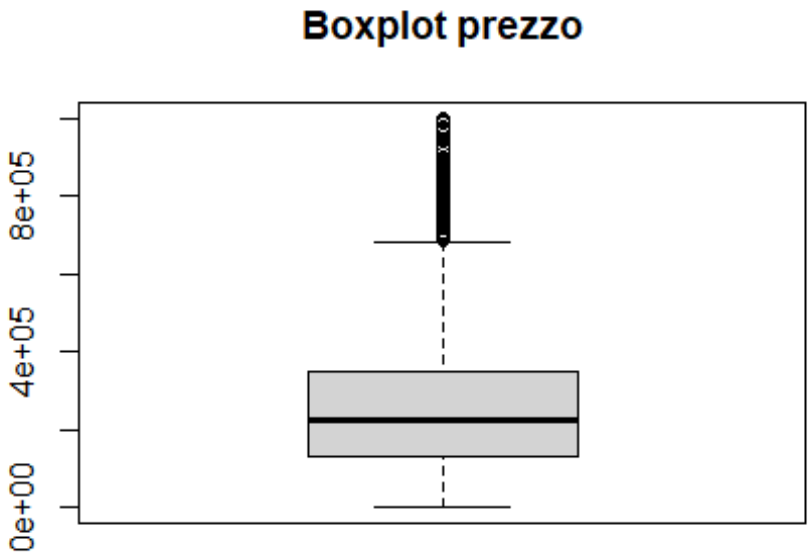

Figura 1. Boxplot del prezzo.

2.3. Superficie

Si passa ora alla variabile superficie. Iniziamo a controllarne alcuni valori sintetici, oltre all'istogramma.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     1.0    88.0   119.0   139.6   165.0   970.0      60
```

Tabella 5. Sintesi della variabile superficie, prima della pulizia.

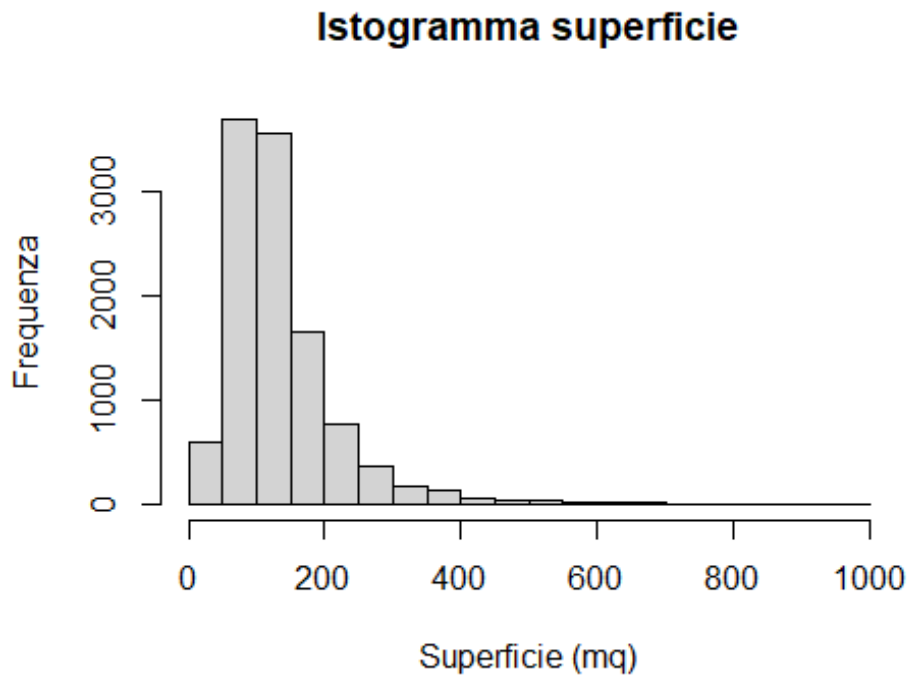

Figura 2. Istogramma della superficie.

Oltre ad eliminare i valori nulli; si filtrano gli immobili anche rispetto al limite di legge[1]. Si imposta quindi la restrizione per cui i mq debbano essere > 14.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    17.0    89.0   120.0   140.3   166.0   970.0
```

Tabella 6. Sintesi della variabile superficie, dopo la pulizia.

A priori ci si aspetta che la superficie sia il primo elemento di significatività per la determinazione del prezzo.

Si prova quindi ora un primo modello lineare molto basilare del prezzo rispetto alla sola superficie, per renderci conto di quale sia la situazione di partenza del coefficiente di determinazione R^2.

Di seguito l'output generato da R-Studio sul modello in questione.

[1] Decreto ministeriale Sanità 5 luglio 1975; art. 2.

```
m1=lm(prezzo~superficie,data)
summary(m1)

##
## Call:
## lm(formula = prezzo ~ superficie, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -851477  -99145  -36427   68569  716195
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 130262.68    2766.03   47.09   <2e-16 ***
## superficie    1019.41      16.66   61.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 155900 on 11101 degrees of freedom
## Multiple R-squared:  0.2522, Adjusted R-squared:  0.2522
## F-statistic:  3744 on 1 and 11101 DF,  p-value: < 2.2e-16
```

La varianza spiegata risulta del 25%; considerando si sta utilizzando solo una variabile, è incoraggiante come inizio.

Qui invece l'output del modello a livello grafico.

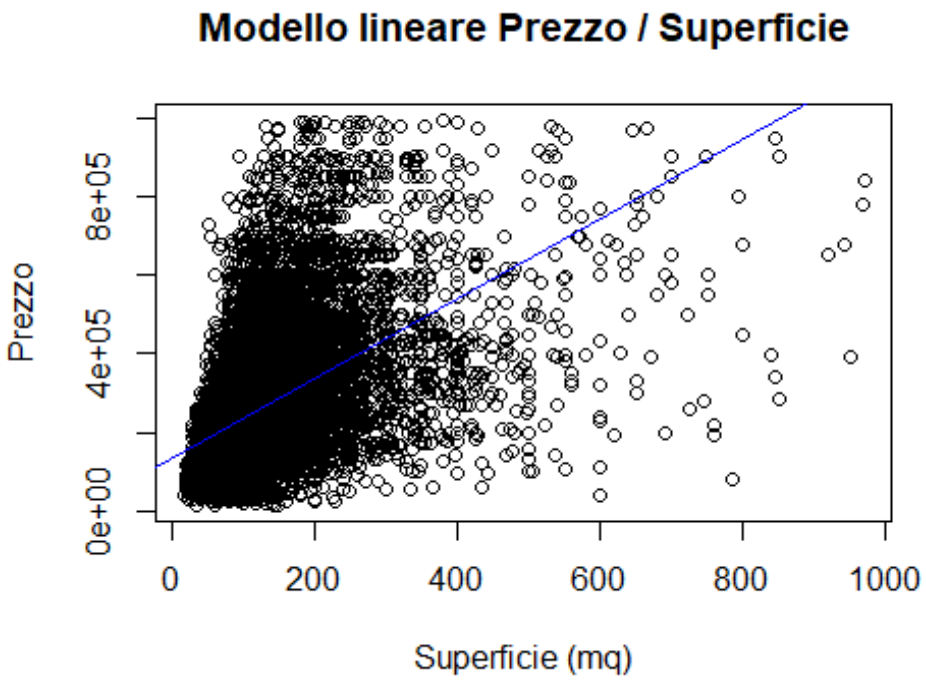

Figura 3. Raffigurazione del modello lineare prezzo / superficie. In blu la retta di regressione lineare che minimizza la somma dei quadrati degli scarti.

2.4. Tipologia

Il dataset aggiornato ha 26 tipologie di immobili. Diamo un'occhiata.

```
##                                N      %
## Appartamento                7847  70.89
## Appartamento in villa         10   0.09
## Attico                       587   5.30
## Attico - Mansarda              1   0.01
## Baita                          1   0.01
## Casa colonica                 17   0.15
## Casa indipendente              6   0.05
## Casale                        15   0.14
## Cascina                        1   0.01
## Loft                          30   0.27
## Mansarda                      52   0.47
## Open space                     6   0.05
## Palazzo - Edificio            87   0.79
## Palazzo - Stabile              1   0.01
## Rustico                      117   1.06
## Rustico - Casale               0   0.00
## Sasso                          1   0.01
## Terratetto plurifamiliare    105   0.95
## Terratetto unifamiliare      506   4.57
## Ufficio                        1   0.01
## Vacanze in appartamento        1   0.01
## Vacanze in Bed & Breakfast     0   0.00
## Villa                         12   0.11
## Villa a schiera              337   3.04
## Villa bifamiliare            535   4.83
## Villa plurifamiliare          94   0.85
## Villa unifamiliare           696   6.29
## Villetta a schiera             3   0.03
## Totale                     11069 100.00
## NA                            34   0.00
```

Tabella 7. Frequenze assolute e relative delle varie tipologie, prima della pulizia.

La variabile tipologia risulta particolarmente problematica viste le numerose modalità. Ad ogni modo si pensa sia troppo importante per poterla eliminare.

Prima del lavoro svolto rispetto alla variabile prezzo e alla variabile superficie, le tipologie erano 28 (sono sparite di fatto le tipologie "Vacanze in Bed&Breakfast" e "Rustico – Casale").

Le 26 variabili rimangono decisamente troppe ai fini della comprensione di qualsivoglia modello.

Si decide, per cominciare, ad eliminare i valori nulli, vista anche la frequenza relativa molto bassa.

Che fare a riguardo? L'obiettivo è ridurre le tipologie creando delle macrocategorie.

I criteri che si prenderanno in considerazione per effettuare quest'accorpamento sono: il significato delle modalità e le frequenze relative.

Vediamo, per approfondimento, anche le medie condizionate del prezzo, relative ad ogni tipologia. Potrebbero tornarci comodo in situazioni di ambiguità.

```
##                         Prezzo medio    N
## Appartamento               243166.2 7847
## Appartamento in villa      306200.0   10
## Attico                     417573.0  587
## Attico - Mansarda          427500.0    1
## Baita                      140000.0    1
## Casa colonica              179882.4   17
## Casa indipendente          225858.3    6
## Casale                     403600.0   15
## Cascina                    260000.0    1
## Loft                       345933.3   30
## Mansarda                   280653.8   52
## Open space                 224500.0    6
## Palazzo - Edificio         430840.2   87
## Palazzo - Stabile          500000.0    1
## Rustico                    262897.4  117
## Sasso                       28000.0    1
## Terratetto plurifamiliare 323647.6  105
## Terratetto unifamiliare   296625.6  506
## Ufficio                    280000.0    1
## Vacanze in appartamento   549000.0    1
## Villa                      292647.8   12
## Villa a schiera            273132.4  337
## Villa bifamiliare          333190.3  535
## Villa plurifamiliare       312510.6   94
## Villa unifamiliare         390407.6  696
## Villetta a schiera         109625.0    3
```

Tabella 8. Prima colonna: media condizionata del prezzo, rispetto alle tipologie; secondo colonna: frequenze assolute.

Ci si è comportati come riportato in seguito a livello decisionale.

Attico - Mansarda	→	Attico
Cascina	→	Rustico
Villa a schiera	→	Villa
Villa bifamiliare	→	Villa

Appartamento in villa	→	Villa

Commento su “Appartamento in villa”: in realtà l’idea iniziale era di accorpare alla categoria “Appartamento”, ma i dati ci mostrano che il prezzo ha un media condizionata molto più vicina a villa.

Villetta a schiera	→	Villa
Terratetto plurifamiliare	→	Terratetto
Terratetto unifamiliare	→	Terratetto
Casale	→	Rustico
Baita	→	Rustico
Casa colonica	→	Rustico
Open space	→	Appartamento
Casa indipendente	→	Appartamento
Loft	→	Mansarda
Palazzo – edificio	→	Palazzo
Palazzo – stabile	→	Palazzo
Sasso	→	Rustico

Ci trova ora in questa situazione.

```
##                       Prezzo medio    N
## Appartamento            243138.8 7859
## Attico                  417589.9  588
## Mansarda                304536.6   82
## Palazzo                 431626.1   88
## Rustico                 265125.0  152
## Terratetto              301269.3  611
## Ufficio                 280000.0    1
## Vacanze in appartamento 549000.0    1
## Villa                   309365.4  991
## Villa unifamiliare      390407.6  696
```

Tabella 9. Prima colonna: media condizionata del prezzo, rispetto alle tipologie; seconda colonna: frequenze assolute.

Si rimuovono le unità con tipologia “Ufficio” e “Vacanze in appartamento”: non sono oggetto di questo studio.

Passo successivo è accorpare le classi con frequenza minore, in una categoria che chiameremo “Altro”.

Vediamo le frequenze relative.

```
## Appartamento               0.710
## Attico                     0.053
## Rustico                    0.014
## Mansarda                   0.007
## Palazzo                    0.008
## Terratetto                 0.055
## Villa                      0.090
## Villa unifamiliare         0.063
```

Tabella 10. Frequenze relative delle tipologie rimanenti.

Quindi, nello specifico.

Rustico	→	Altro
Mansarda	→	Altro
Palazzo	→	Altro

Ecco le sei macrocategorie finali.

```
##                         N      %
## Appartamento         7859  71.02
## Attico                588   5.31
## Terratetto            611   5.52
## Villa                 991   8.96
## Villa unifamiliare    696   6.29
## Altro                 321   2.90
## Totale              11066 100.00
## NA                      0   0.00
```

Tabella 11. Frequenze assolute e relative finali, rispetto alle varie tipologie.

2.5. Anno di costruzione

Iniziamo subito con l'istogramma ed il box-plot per comprenderne un po' la natura.

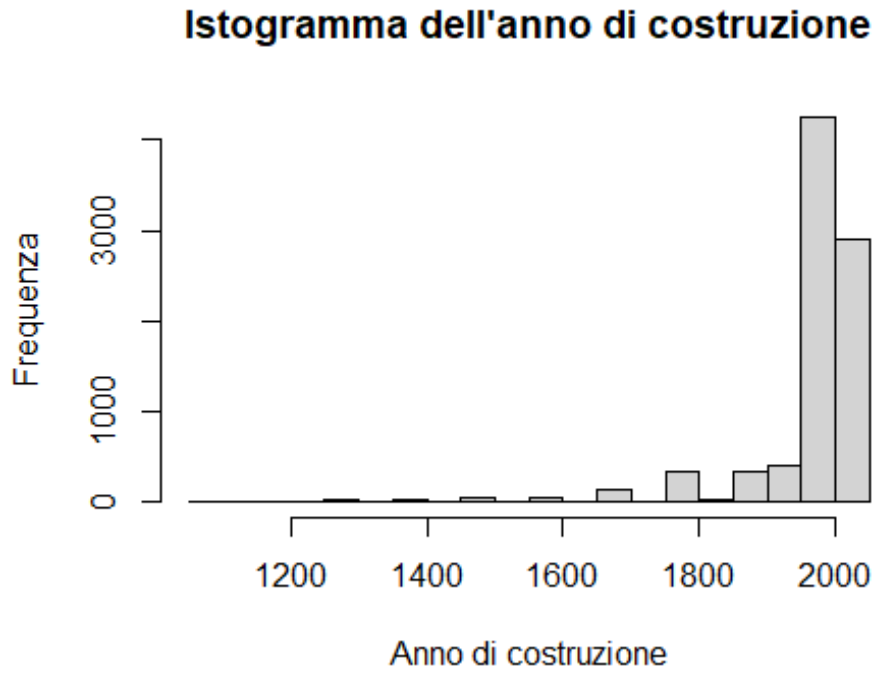

Figura 4. Istogramma dell'anno di costruzione.

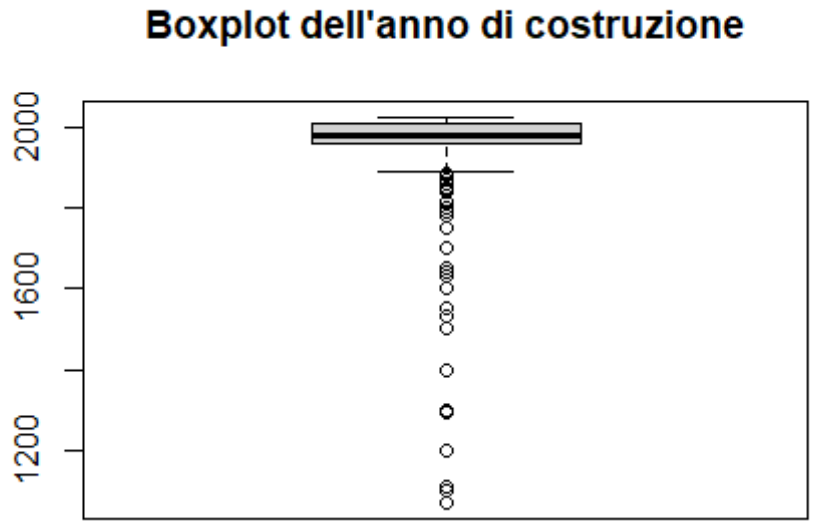

Figura 5. Boxplot dell'anno di costruzione dell'immobile.

La costruzione degli immobili è schiacciata verso gli anni più recenti. La mediana si posiziona all'anno 1978.

Dando un'occhiata alle varie medie del prezzo, per ogni anno, ci si accorge che la storicità dell'immobile crea valore, invece che toglierne (*figura 6*).

Si pensa quindi che un semplice modello lineare non possa cogliere tutti gli aspetti di questa variabile.

Ci serviranno quindi più livelli per poter utilizzare al meglio l'informazione sull'anno di costruzione.

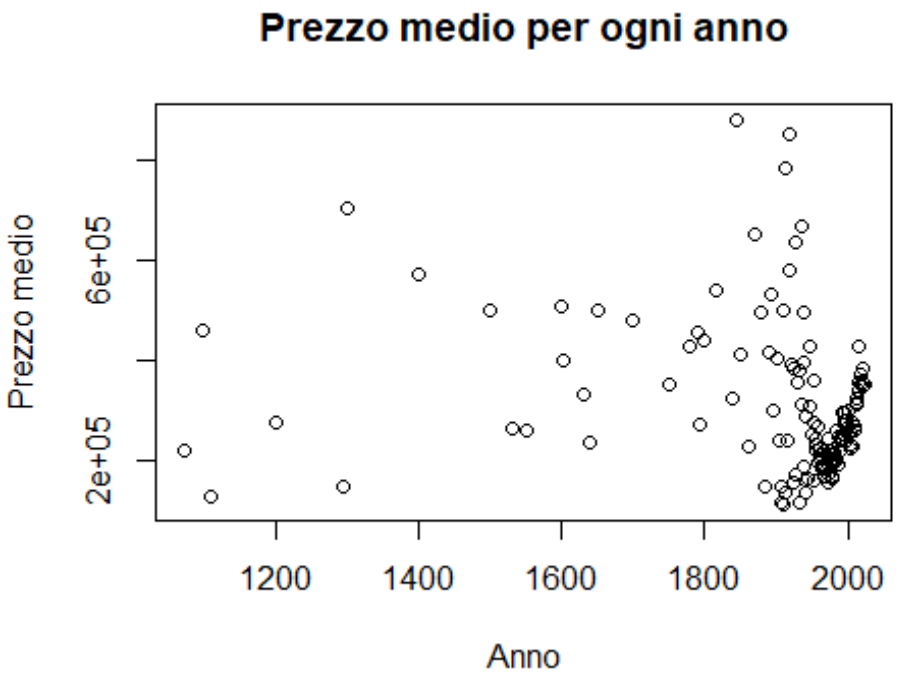

Figura 6. Grafico del prezzo medio condizionato all'anno di costruzione.

Proviamo a creare delle fasce di "storicità", che nello specifico sono:

- ≤ 1850, che chiameremo "Antichissimo";
- 1851 – 1950, che diventerà "Antico";
- 1951 – 1990, come "Vecchio";
- ≥ 1991; prenderà il nome di "Moderno";

- si crea anche la categoria “Sconosciuto” per i dati in cui l’anno di costruzione è mancante (sono 2528 unità).

Guardando il seguente grafico delle medie del prezzo, rispetto alla fascia, si conferma quanto scritto precedentemente.

Figura 7. Prezzo medio per ogni scaglione d’età.

2.6. Ascensore

Sono presenti inizialmente solo i valori “1” ed “NA”.

```
##            N
## 1       4155
## NA      6912
```

Tabella 12. Frequenze assolute rispetto all’ascensore.

Si è deciso di assegnare il valore “0” ai 6912 mancanti.

Con certezza non era possibile inserire l’informazione di “assenza”.

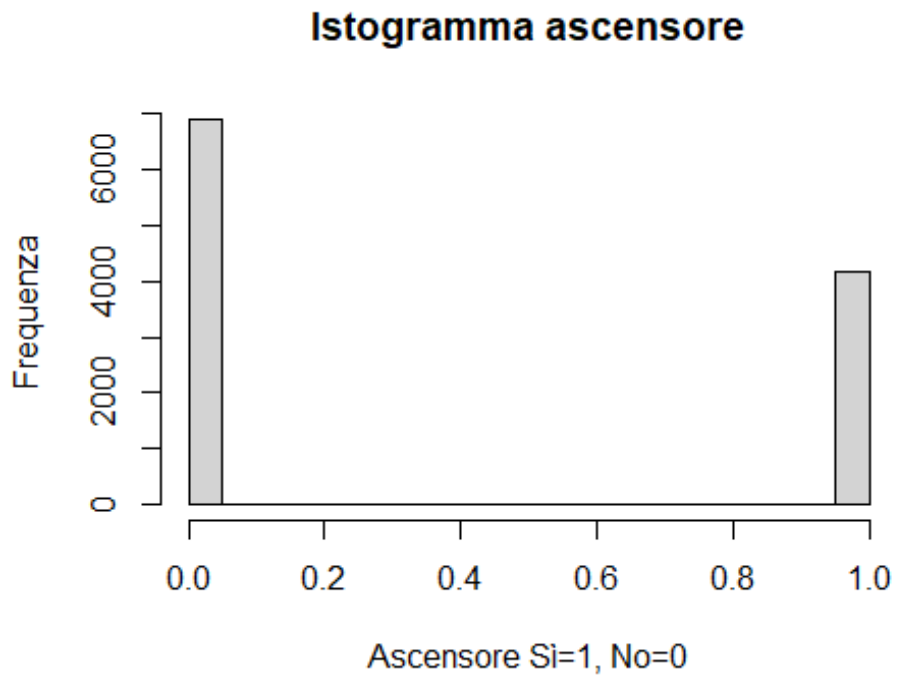

Figura 8. Istogramma ascensore. '1' sta per presenza del carattere, '0' per mancanza.

2.7. Bagni

Vediamo subito una tabella sintetica relativa al numero di bagni.

```
##            N      %
## 2       4729  75.06
## 3       1244  19.75
## 3+       327   5.19
## Totale  6300 100.00
## NA      4767   0.43
```

Tabella 13. Frequenze assolute e relative del numero dei bagni. Dati iniziali.

In questo caso invece si è deciso di assegnare "1" ai valori mancanti, presupponendo il fatto che non sia possibile avere un immobile residenziale senza almeno un bagno.

È sicuramente una considerazione forte, ma viene supportata anche dal fatto che il valore "1" non fosse neppure presente inizialmente.

Le percentuali ottenute risultano più che plausibili e ci confortano.

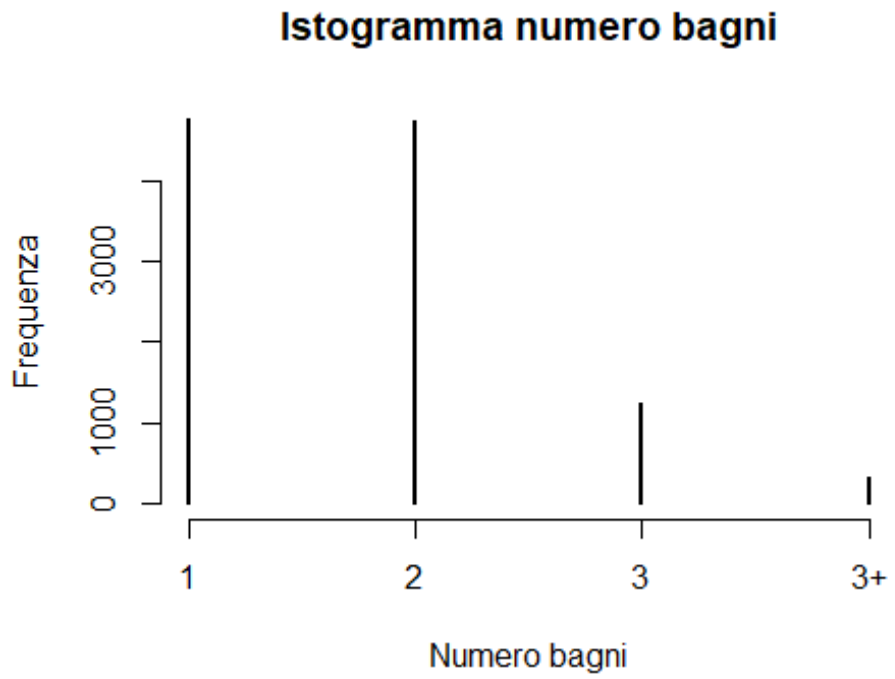

Figura 9. Istogramma del numero dei bagni, dopo modifiche.

```
##             N       %
## 1        4767   43.07
## 2        4729   42.73
## 3        1244   11.24
## 3+        327    2.95
## Totale 11066  100.00
```

Tabella 13. Frequenze assolute e relative del numero dei bagni, dopo modifiche.

2.8. Box Auto

Si è deciso di assegnare il valore “0” ai valori mancanti, la frequenza “0”, anche qui, non era rilevabile.

Ecco la situazione di partenza.

```
##    1    2    3    4    5    6    8   10   12
## 5231 1577   50   31    6    3    2    2    2
```

Tabella 14. Frequenze assolute del numero dei box auto. Situazione iniziale.

Per semplicità poi, si sono raggruppate le classi per cui il valore è >1, denominandola “1+”.

Si sono eliminate anche le unità con box-auto uguali e superiori a 10, poiché insensati in relazione ad un immobile residenziale.

Si poteva anche pensare di mantenere la variabile come numerica, ma avendo creato solo tre livelli si è stabilito di mantenerle come classi.

```
##              N      %
## 0         4163   37.62
## 1         5231   47.29
## 1+        1669   15.09
## Totale   11062  100.00
```

Tabella 15. Frequenze assolute e relative del numero dei box auto. Situazione finale.

2.9. Classe energetica

Rispetto alla classe energetica, la considerazione che iniziale è che, visto può essere vista come una variabile categoriale ordinale, verrà considerata come numerica. Ci aspettiamo un andamento crescente del prezzo al migliorare della classe energetica.

Di seguito il grafico delle frequenze assolute.

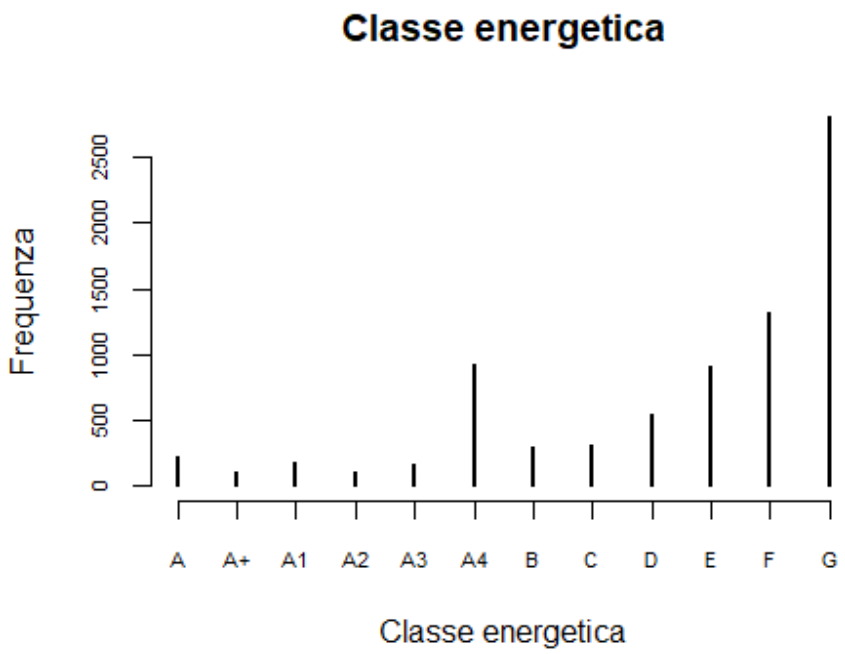

Figura 10. Grafico a bastoncini della classe energetica. Frequenze assolute.

```
##            N     %
## A4       923 11.77
## A3       154  1.96
## A2       100  1.27
## A1       181  2.31
## A+        96  1.22
## A        212  2.70
## B        292  3.72
## C        306  3.90
## D        544  6.94
## E        905 11.54
## F       1317 16.79
## G       2814 35.87
## Totale 7844 99.99
## NA      3218  0.29
```

Tabella 16. Frequenze assolute e relative finali alla classe energetica.

Per quanto riguarda i valori mancanti, essendo ora la variabile di tipo numerica, non si può banalmente creare la categoria "Sconosciuto", poiché andrebbe in contrasto con la natura del tipo di variabile scelta.

Si è deciso di darne il valore 10.5 (che corrisponde all'essere tra la classe energetica E la classe energetica F). Scelta più ragionevole sulla base del prezzo medio, che è di € 244899,6.

È anche sensato pensare con qualcuno possa "omettere" con più facilità il valore, poiché la classe energetica è scarsa.

Qui le varie medie condizionate.

```
##        A       A+       A1       A2       A3       A4        B        C
## 389500.0 348484.4 308365.3 318630.0 355925.3 383069.8 345824.3 294641.8
##        D        E        F        G
## 290036.3 255153.7 242844.3 251057.6
```

Tabella 17. Media del prezzo, condizionata alla classe energetica.

Il grafico è a supporto della tesi.

Sull'asse delle x le classi energetiche sono in ordine decrescente, ovvero: A4, A3, A2, A1, A+, A, B, C, D, E, F, G.

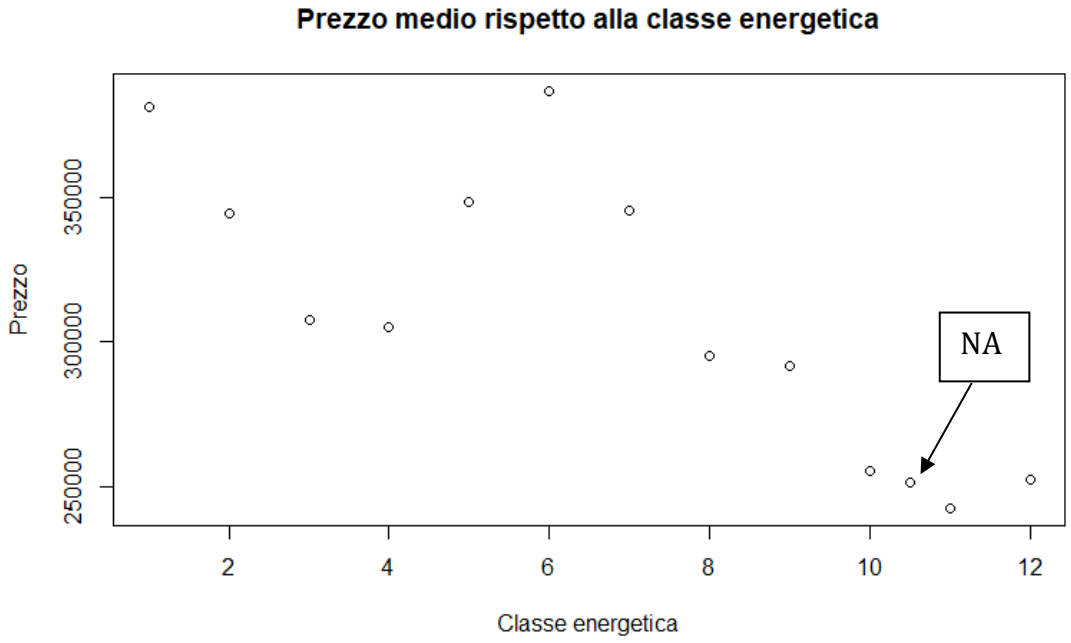

Figura 11. Prezzo medio rispetto alla classe energetica. La scala è decrescente. (A4=1, A3=2, A2=3, A1=4, … , G=12; mentre NA=10.5)

2.10. Clima

Come fatto spesso finora, iniziamo la comprensione di questa informazione con la tabella delle frequenze assolute e relative.

```
##                     N      %
## Assente            91   1.28
## Autonomo         5882  82.42
## Caldo               3   0.04
## Centralizzato     154   2.16
## Freddo            308   4.32
## Predisposizione   699   9.79
## Totale           7137 100.00
## NA               3925  35.48
```

Tabella 18. Frequenze assolute e relative rispetto al "clima". Valori iniziali.

La decisione qui è quella di raggruppare in una categoria 'Altro', i livelli 'assente', 'autonomo', 'caldo', 'centralizzato', 'freddo' e 'predisposizione'. La scelta ricade esclusivamente sulla base delle frequenze relative.

Mentre per i valori mancanti vengono gestiti lasciandoli come tali, ovvero valori sconosciuti.

```
##       Altro    Autonomo Sconosciuto
##        1255        5882        3925
```

Tabella 19. Frequenze assolute finali rispetto al "clima". Valori finali.

2.11. Numero locali

Vediamo subito la tabella delle frequenze in relazione al numero dei locali.

```
##             N      %
## 2        1063   9.84
## 3        2419  22.42
## 4        3046  28.22
## 5        2040  18.90
## 5+       2223  20.62
## Totale 10802 100.00
## NA         271   0.02
```

Tabella 20. Frequenze assolute e relative rispetto al numero di locali.

Si decide di mantenerla come variabile categoriale anziché numerica, poiché i livelli sono solo 5.

Visto che la variabile viene considerata come fattoriale e non numerica, questo facilita la anche la comprensione dei valori mancanti (271 unità), che verranno considerati come la nostra categoria "Sconosciuto".

2.12. Riscaldamento

Qui la variabile categoriale presenta solo due modalità, oltre a 1179 valori nulli. Anche qui si fissano i valori mancanti come la categoria "Sconosciuto".

```
##      Autonomo Centralizzato Sconosciuto
##          7872          2011        1179
```

Tabella 21. Frequenze assolute e relative rispetto al numero di locali.

2.13. Piano

Questa la situazione di partenza:

```
##           N     %
## T      1653 20.35
## 1      1865 22.96
## 2      1744 21.47
## 3      1115 13.73
## 4       471  5.80
## 5       186  2.29
## 6        80  0.98
## 7        33  0.41
## 8        12  0.15
## 9         8  0.10
## 10        2  0.02
## 11        5  0.06
## 12        2  0.02
## 13        4  0.05
## 16        1  0.01
## 18        5  0.06
## A        22  0.27
## R       293  3.61
## S       618  7.61
## S2        2  0.02
## S3        1  0.01
## Totale 8122 99.98
## NA     2940  0.27
```

Tabella 22. Frequenze assolute e relative con riferimento al piano in cui si trova l'immobile. Situazione iniziale.

Vista la numerosità elevata, si è scelto di tenere la classe 'T', '1' e '2'. È stata creata la categoria "Altro", in cui sono state inserite le modalità 'A', 'R', 'S', 'S2' e 'S3'.

Come atto conclusivo sono stati considerati i valori *NA* come categoria "Sconosciuto".

```
##                 N      %
## T            1653  14.94
## 1            1865  16.86
## 2            1744  15.76
## 2+           1924  17.39
## Altro         936   8.46
## Sconosciuto  2941  26.58
## Totale      11063 100.00
## NA              0   0.00
```

Tabella 23. Frequenze assolute e relative con riferimento al piano in cui si trova l'immobile. Situazione finale.

2.14. Stato

Vediamo subito la situazione di partenza.

```
##                             N      %
## € 105.000,00                1   0.01
## € 11.250,00                 1   0.01
## € 187.500,00                1   0.01
## € 27.000,00                 1   0.01
## € 40.800,00                 1   0.01
## € 42.750,00                 1   0.01
## € 43.500,00                 1   0.01
## € 57.000,00                 1   0.01
## € 73.800,00                 1   0.01
## € 85.000,00                 1   0.01
## € 87.525,00                 1   0.01
## Buono / Abitabile        4026  38.43
## Da ristrutturare         1259  12.02
## No                          2   0.02
## Nuovo / In costruzione   1950  18.62
## Ottimo / Ristrutturato   3227  30.81
## Totale                  10475 100.01
## NA                        587   0.05
```

Tabella 24. Iniziali frequenze assolute e relative in riferimento allo stato dell'immobile.

Guardando la tabella e rendendosi conto delle modalità sicuramente errate, si è deciso di raggruppare nella categoria "Altro" tutte le modalità con frequenza sotto il 5%.

Vista la comunque bassissima numerosità, si è provveduto ad inglobare in questa categoria anche tutti i valori mancanti. Chiamando poi la variabile "Altro".

```
##                   Altro      Buono / Abitabile       Da ristrutturare
##                     600                   4026                   1259
## Nuovo / In costruzione Ottimo / Ristrutturato
##                    1950                   3227
```

Tabella 25. Frequenze assolute finali in riferimento allo stato dell'immobile. La categoria "Altro" ingloba anche le 587 con l'informazione mancante.

2.15. Macrozona ID/Macrozona

Si è preferito utilizzare la variabile macrozona ID, che, benché sia meno esplicativa dal punto di vista dell'informazione diretta, presenta molti meno valori mancanti.

Visto che le due variabili rappresentano la stessa informazione, si è pensato di sfruttarle per poter colmare alcuni valori nulli.

Non è stato però possibile, poiché valori *NA* in macrozona corrispondevano sempre in *NA* negli ID.

Sarà poi semplice ricavarsi il nome della macrozona, partendo dall'ID.

Dal punto di vista decisionale, qui si è scelto di raggruppare nella categoria "Altro" tutte le modalità con frequenza relativa minore al 3%, la scelta è assolutamente arbitraria e dettata dalla numerosità davvero molto alta delle varie modalità ID (sono 90 in tutto).

Le uniche macrozone ID con frequenza superiore al 3% risultano le n. 168, 368, 10962 e 10212 che equivalgono rispettivamente alle macrozona "Mestre, Chirignago, Marghera, Catene", "Centro" di Treviso, "Sottomarina" e "Centro" di Verona.

Infine, visto che la categoria dei valori mancanti concentrava già il 6,5%, i valori assenti vengono specificati come categoria “Sconosciuto”.

2.16. Altre variabili

Non saranno considerate al fine della stima dei modelli previsivi, il titolo, la descrizione, la via e il numero della via, i tags, la latitudine, la longitudine, l’url (ovvero il link), la data e l’ora del crawler.

Può essere spunto di riflessione, per uno studio futuro, considerare il titolo e la descrizione dell’annuncio.

Le altre variabili dicotomiche non hanno avuto bisogno di elaborazione, nessuna di queste presenta valori mancanti.

Le adotteremo per come si presentano, lasceremo poi esprimere i modelli a riguardo.

3. Alcuni modelli di previsione

La modellazione predittiva è la fase saliente dell'analisi predittiva che include metodologie e tecniche in grado di estrarre conoscenza da dati a disposizione per fare predizioni su dati o eventi nel futuro.

Bisogna innanzitutto capire se è preferibile un modello che le include tutte le variabili (detto modello completo) o un modello che ne include solo alcune (detto modello ridotto).

Un modello di regressione con molte variabili esplicative è più difficile tra interpretare ed alcune variabili possono risultare ridondanti.

Per confrontare i modelli si possono utilizzare vari metodi.

In generale, è meglio non fare affidamento su un solo metodo ma utilizzane diversi per vedere se ti portano alla stessa conclusione.

Se le verifiche precedenti portano a restringere la scelta tra alcuni modelli molto simili tra loro come potere predittivo, in generale è ragionevole utilizzare il più semplice.

Per la costruzione di ogni modello, si utilizzano il 75% delle osservazioni come insieme di stima (*train*). Il restante 25% verrà utilizzato come insieme di prova (*test*).

Queste porzioni dei dati sono fissate e sono identiche per ogni modello.

Dobbiamo operare cercando un compromesso tra componente di errore e componente di varianza. La trappola da evitare è il sovra-adattamento ai dati.

3.1. Modello di regressione lineare

La forma di regressione lineare è quella basata sul metodo dei minimi quadrati. La prima pubblicazione contenente un'applicazione del metodo nota è datata 1805, a nome di Adrien-Marie Legendre.

Data una certa distribuzione di dati sperimentali, supponiamo che questa possa essere approssimata da una retta che, in qualche modo, sia allineabile fra questi

punti. Si ricorre ad un metodo matematico oggettivo, che prende il nome di metodo dei minimi quadrati.

Il criterio su cui si basa l'individuazione della retta di regressione è abbastanza intuitivo. Questa retta si ottiene costruendo, per ogni punto sperimentale, un quadrato che ha un lato costituito dalla distanza verticale (ordinata) del punto dalla retta. Si ripete quindi il procedimento per ogni punto e si sommano le aree di tutti i quadrati. La retta che approssima meglio la distribuzione dei punti è quella che determina la minore superficie dei quadrati (di qui il termine "metodo dei minimi quadrati").

Il calcolo dei quadrati è necessario perché se il procedimento si basasse sull'uso diretto delle distanze dei punti dalla retta ideale, si perderebbe accuratezza in quanto le ordinate negative si sottrarrebbero a quelle positive.

La regressione lineare multipla è un'estensione dell'analisi della correlazione e della regressione lineare semplice.

Permette di analizzare la relazione lineare tra variabili: si possono verificarne sia la direzione che la significatività. Inoltre, la regressione fa sì che si possa quantificare di quanto in media aumenterà o diminuirà la variabile risposta alla modifica di una variabile esplicativa.

Si specifica che, in questo modello, per non appesantirlo ulteriormente, non verranno prese in considerazioni le interazioni tra le variabili indipendenti.

Si veda l'output del modello di regressione multipla. Una semplice nota sulle variabili fattoriali: la classe di riferimento è inglobata nell'intercetta.

```
## Call:
## lm(formula = prezzo ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -571056  -60208   -7377   45428  628963
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                              298267.47   15746.95  18.941
## anno_costrAntico                         -66804.63    6980.45  -9.570
## anno_costrModerno                       -104450.06    6529.83 -15.996
## anno_costrSconosciuto                    -98663.65    6267.85 -15.741
## anno_costrVecchio                       -139547.35    6203.65 -22.494
## ascensore                                18476.29    3206.85   5.762
```

```
## bagni2                                  55318.73    3209.98  17.233
## bagni3                                 129023.77    5235.99  24.642
## bagni3+                                136395.12    8646.50  15.775
## boxauto(1,13]                            1617.23    3789.44   0.427
## boxauto(13,50]                          18216.76    3188.46   5.713
## classe_en                               -2938.02     602.81  -4.874
## climaAutonomo                            6003.44    4241.14   1.416
## climaSconosciuto                        -5401.49    4607.81  -1.172
## comuneCastelfranco Veneto               -10915.29   12743.06  -0.857
## comuneChioggia                           5121.72    9254.78   0.553
## comunePadova                            49017.56    8348.45   5.871
## comunePordenone                        -10063.83   10786.27  -0.933
## comuneRovigo                           -57632.11    9369.79  -6.151
## comuneTreviso                           42494.45    8442.72   5.033
## comuneVenezia                          158002.54    8892.80  17.767
## comuneVerona                            44982.78    8509.66   5.286
## comuneVicenza                             520.08    8570.06   0.061
## locali3                                 13501.01    4737.92   2.850
## locali4                                 33346.21    4842.93   6.886
## locali5                                 51186.99    5351.24   9.565
## locali5+                                55750.77    5707.07   9.769
## localiSconosciuto                       11006.74    8635.13   1.275
## macrozona_id10962                     -112381.53   10535.35 -10.667
## macrozona_id168                       -268674.38   10022.10 -26.808
## macrozona_id368                        -28680.24   10289.13  -2.787
## macrozona_idAltro                     -150334.35    7148.53 -21.030
## macrozona_idSconosciuto               -116839.81   11705.10  -9.982
## piano2                                  10958.30    4179.00   2.622
## piano2+                                  -932.95    4184.21  -0.223
## pianoAltro                              -9107.42    5185.52  -1.756
## pianoSconosciuto                       -18455.31    6350.61  -2.906
## pianoT                                 -10245.58    4418.71  -2.319
## riscaldamentoCentralizzato              -7044.70    3717.30  -1.895
## riscaldamentoSconosciuto               -14529.80    4507.14  -3.224
## statoBuono / Abitabile                   1297.44    5968.98   0.217
## statoDa ristrutturare                  -33023.92    6712.25  -4.920
## statoNuovo / In costruzione             42787.06    6930.86   6.173
## statoOttimo / Ristrutturato             40525.37    6151.87   6.587
## superficie                                767.56      20.88  36.755
## tipologiaAttico                         52134.86    5740.29   9.082
## tipologiaTerratetto                     -6474.53    7734.06  -0.837
## tipologiaVilla                         -11677.34    7524.12  -1.552
## tipologiaVilla unifamiliare             10616.80    7943.52   1.337
## tipologiaAltro                         -39183.49    8367.67  -4.683
## Fibra.ottica                            -1596.32    3360.81  -0.475
## Balcone                                 -5835.61    2695.92  -2.165
## Terrazza                                -3571.38    2761.48  -1.293
## Impianto.tv.centralizzato                3783.38    4533.45   0.835
## Cantina                                  3026.10    2863.75   1.057
## Giardino.comune                        -18126.94    3322.79  -5.455
## Infissi.esterni.in.vetro...legno        16629.25    5161.07   3.222
## Esposizione.doppia                      -8984.99    4803.17  -1.871
## VideoCitofono                           11559.05    3767.44   3.068
## Cancello.elettrico                      -3960.41    3853.55  -1.028
## Caminetto                               12426.46    4888.54   2.542
## Porta.blindata                           5720.30    3322.39   1.722
## Impianto.tv.con.parabola.satellitare    10371.14    7605.69   1.364
## Giardino.privato                        -2416.37    4051.32  -0.596
## Parzialmente.Arredato                   11580.00    3743.72   3.093
## Infissi.esterni.in.doppio.vetro...PVC    -6282.45    6425.58  -0.978
```

```
## Esposizione.esterna                          -6247.52    4216.27  -1.482
## Armadio.a.muro                                2818.86    3860.28   0.730
## Infissi.esterni.in.doppio.vetro...legno       1581.04    5042.59   0.314
## Taverna                                     -17360.00    6034.84  -2.877
## Impianto.di.allarme                          20425.98    3962.49   5.155
## Impianto.tv.singolo                           -799.11    4477.32  -0.178
## Mansarda                                    -15136.50    4458.50  -3.395
## Esposizione.interna                         -18644.44    6053.46  -3.080
## Idromassaggio                                14476.23    7270.07   1.991
## Infissi.esterni.in.triplo.vetro...PVC        -3931.27    9116.24  -0.431
## Cucina                                      -30598.52   11621.98  -2.633
## Solo.Cucina.Arredata                         -4094.48    6402.76  -0.639
## Esposizione.sud                              30016.49   20898.23   1.436
## ovest                                       -20567.57   20457.94  -1.005
## Arredato                                     -2781.40    4028.52  -0.690
## Piscina                                        289.63   11575.61   0.025
## Infissi.esterni.in.triplo.vetro...legno       3535.47    8895.05   0.397
## Esposizione.nord                             13416.77   31213.60   0.430
## sud                                         -43584.28   32383.63  -1.346
## est                                          -8395.32   22667.36  -0.370
## Infissi.esterni.in.vetro...metallo           -9699.39   12383.85  -0.783
## Infissi.esterni.in.vetro...PVC              -21285.24   12908.35  -1.649
## Infissi.esterni.in.triplo.vetro...metallo    28429.34   17648.19   1.611
## Portiere.intera.giornata                      4628.33    9538.76   0.485
## Infissi.esterni.in.doppio.vetro...metallo   -15967.74   10261.55  -1.556
## Giardino.privato.e.comune                   -42217.22   29729.51  -1.420
## Esposizione.est                              26231.02   35516.03   0.739
## Portiere.mezza.giornata                       5336.89   24224.33   0.220
## Reception                                   180356.57   55583.90   3.245
## Passo.carrabile                            -104094.36   56287.82  -1.849
## Esposizione.ovest                            28362.14   45195.27   0.628
##                                             Pr(>|t|)
## (Intercept)                                  < 2e-16 ***
## anno_costrAntico                             < 2e-16 ***
## anno_costrModerno                            < 2e-16 ***
## anno_costrSconosciuto                        < 2e-16 ***
## anno_costrVecchio                            < 2e-16 ***
## ascensore                                   8.64e-09 ***
## bagni2                                       < 2e-16 ***
## bagni3                                       < 2e-16 ***
## bagni3+                                      < 2e-16 ***
## boxauto(1,13]                                0.66956
## boxauto(13,50]                              1.15e-08 ***
## classe_en                                   1.11e-06 ***
## climaAutonomo                                0.15695
## climaSconosciuto                             0.24113
## comuneCastelfranco Veneto                    0.39171
## comuneChioggia                               0.58000
## comunePadova                                4.49e-09 ***
## comunePordenone                              0.35084
## comuneRovigo                                8.07e-10 ***
## comuneTreviso                               4.92e-07 ***
## comuneVenezia                                < 2e-16 ***
## comuneVerona                                1.28e-07 ***
## comuneVicenza                                0.95161
## locali3                                      0.00439 **
## locali4                                     6.18e-12 ***
## locali5                                      < 2e-16 ***
## locali5+                                     < 2e-16 ***
## localiSconosciuto                            0.20247
```

```
## macrozona_id10962                          < 2e-16 ***
## macrozona_id168                            < 2e-16 ***
## macrozona_id368                            0.00532 **
## macrozona_idAltro                          < 2e-16 ***
## macrozona_idSconosciuto                    < 2e-16 ***
## piano2                                     0.00875 **
## piano2+                                    0.82356
## pianoAltro                                 0.07907 .
## pianoSconosciuto                           0.00367 **
## pianoT                                     0.02044 *
## riscaldamentoCentralizzato                 0.05811 .
## riscaldamentoSconosciuto                   0.00127 **
## statoBuono / Abitabile                     0.82793
## statoDa ristrutturare                     8.83e-07 ***
## statoNuovo / In costruzione               7.00e-10 ***
## statoOttimo / Ristrutturato               4.75e-11 ***
## superficie                                 < 2e-16 ***
## tipologiaAttico                            < 2e-16 ***
## tipologiaTerratetto                        0.40254
## tipologiaVilla                             0.12070
## tipologiaVilla unifamiliare                0.18141
## tipologiaAltro                            2.88e-06 ***
## Fibra.ottica                               0.63481
## Balcone                                    0.03045 *
## Terrazza                                   0.19595
## Impianto.tv.centralizzato                  0.40400
## Cantina                                    0.29068
## Giardino.comune                           5.03e-08 ***
## Infissi.esterni.in.vetro...legno           0.00128 **
## Esposizione.doppia                         0.06143 .
## VideoCitofono                              0.00216 **
## Cancello.elettrico                         0.30411
## Caminetto                                  0.01104 *
## Porta.blindata                             0.08515 .
## Impianto.tv.con.parabola.satellitare       0.17273
## Giardino.privato                           0.55090
## Parzialmente.Arredato                      0.00199 **
## Infissi.esterni.in.doppio.vetro...PVC      0.32824
## Esposizione.esterna                        0.13844
## Armadio.a.muro                             0.46528
## Infissi.esterni.in.doppio.vetro...legno    0.75388
## Taverna                                    0.00403 **
## Impianto.di.allarme                       2.60e-07 ***
## Impianto.tv.singolo                        0.85835
## Mansarda                                   0.00069 ***
## Esposizione.interna                        0.00208 **
## Idromassaggio                              0.04649 *
## Infissi.esterni.in.triplo.vetro...PVC      0.66631
## Cucina                                     0.00848 **
## Solo.Cucina.Arredata                       0.52252
## Esposizione.sud                            0.15095
## ovest                                      0.31475
## Arredato                                   0.48995
## Piscina                                    0.98004
## Infissi.esterni.in.triplo.vetro...legno    0.69103
## Esposizione.nord                           0.66733
## sud                                        0.17838
## est                                        0.71112
## Infissi.esterni.in.vetro...metallo         0.43352
## Infissi.esterni.in.vetro...PVC             0.09920 .
```

```
## Infissi.esterni.in.triplo.vetro...metallo  0.10724
## Portiere.intera.giornata                   0.62754
## Infissi.esterni.in.doppio.vetro...metallo  0.11973
## Giardino.privato.e.comune                  0.15563
## Esposizione.est                            0.46019
## Portiere.mezza.giornata                    0.82563
## Reception                                  0.00118 **
## Passo.carrabile                            0.06445 .
## Esposizione.ovest                          0.53032
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 107200 on 8201 degrees of freedom
## Multiple R-squared:  0.6591, Adjusted R-squared:  0.6551
## F-statistic: 165.1 on 96 and 8201 DF,  p-value: < 2.2e-16
```

È difficile soffermarsi sui singoli numeri per il lettore.

I coefficienti siano in tutto 97, dando un'occhiata alla significatività e alla stima dei parametri si possono tratte già, per iniziare, buone informazioni.

3.2. Modello di regressione lineare con selezione

Per ovviare la questione dei troppi parametri nel modello precedente, si procede ora con una selezione delle variabili esplicative.

La selezione avviene tramite il metodo *stepwise* ibrido. Esso si basa su un algoritmo che automaticamente rimuove (o aggiunge) una variabile alla volta al modello di regressione.

Esso inizia in modalità *forward*, ma se necessario, procede in modalità *backward* e rimuove una variabile precedentemente inserita.

Il modello migliore è scelto in base alla validazione incrociata (*cross-validation*).

Questa è una tecnica statistica che permette di usare in modo alternato i dati sia per il train che per il test.

Spesso viene chiamata *k-fold cross validation* perché diviso il dataset iniziale in una serie di porzioni uguali di dati (k-campi) e ne vengono utilizzate iterativamente per il *train* e per il *test*.

La *k-fold cross validation* segue una logica dinamica, con uno schema come quello riportato nell'immagine qui sotto. Nel nostro caso k è pari a 5.

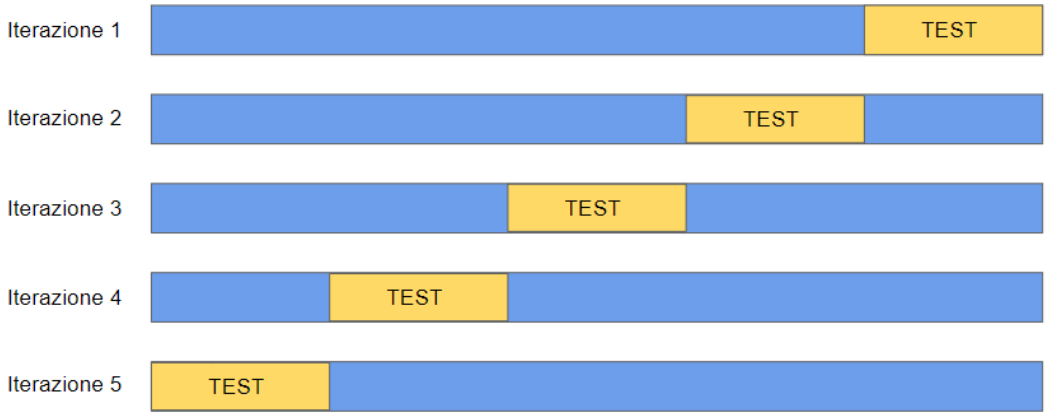

Figura 12. K-fold Cross Validation

Nella cross validation si vede bene come ogni volta si utilizzano porzioni diverse per il *train* e per il *test*. Il *train* è rappresentato dalla parte blu, mentre il *test* da quella gialla.

Il risultato finale sarà poi una media delle *performances* delle varie iterazioni, che, come sopraddetto, nel nostro caso sono 5.

Attenzione a non fare confusione: queste iterazioni avvengono nello stesso ed identico insieme di *train* (75% delle unità) del modello precedente.

Il restante 25% verrà utilizzato sempre e solo al termine, per valutarne la qualità a livello previsivo.

Si aggiunge che non è stata utilizzata la significatività dei parametri, poiché, come si sa, l'errore standard decresce molto rapidamente all'aumentare della numerosità. Il *p-value* ci mostrerà sì l'effetto "significativo", diverso da 0, ma questo potrebbe essere troppo piccolo per risultare interessante (potrebbe essere di fatto non rilevante a livello pratico).

Vediamo l'output di R-Studio.

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -573121   -60437    -7133    45015   636161
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                              295711.07   12443.97  23.763
## anno_costrAntico                         -66653.14    6950.25  -9.590
## anno_costrModerno                       -103779.24    6460.29 -16.064
## anno_costrSconosciuto                    -97832.48    6194.07 -15.795
## anno_costrVecchio                       -139307.83    6158.69 -22.620
## ascensore                                19017.69    3115.04   6.105
## bagni2                                   55593.80    3177.80  17.494
## bagni3                                  129029.56    5192.00  24.852
## `bagni3+`                               136602.91    8574.00  15.932
## `boxauto(13,50]`                         19038.06    3027.98   6.287
## classe_en                                -2874.21     578.68  -4.967
## climaAutonomo                             9638.29    2707.01   3.560
## comunePadova                             47830.96    3979.10  12.021
## comuneRovigo                            -58255.99    5828.08  -9.996
## comuneTreviso                            42876.43    4342.73   9.873
## comuneVenezia                           157268.46    4861.88  32.347
## comuneVerona                             45265.85    4429.33  10.220
## locali3                                  11847.08    4380.32   2.705
## locali4                                  31707.00    4437.53   7.145
## locali5                                  49647.07    4947.68  10.034
## `locali5+`                               54066.18    5275.97  10.248
## macrozona_id10962                      -108788.30    9859.69 -11.034
## macrozona_id168                        -268939.08    9940.44 -27.055
## macrozona_id368                         -29281.27   10244.59  -2.858
## macrozona_idAltro                       -150587.73    7089.22 -21.242
## macrozona_idSconosciuto                 -125065.65    9051.98 -13.816
## piano2                                   11444.80    3616.45   3.165
## pianoAltro                               -9171.02    4672.77  -1.963
## pianoSconosciuto                        -22709.38    4536.54  -5.006
## pianoT                                  -10640.37    3868.78  -2.750
## riscaldamentoCentralizzato               -7020.56    3586.04  -1.958
## riscaldamentoSconosciuto                -14388.56    4344.76  -3.312
## `statoDa ristrutturare`                 -34265.85    4299.88  -7.969
## `statoNuovo / In costruzione`            42089.52    4855.41   8.669
## `statoOttimo / Ristrutturato`            39442.94    3078.95  12.810
## superficie                                 768.92      20.04  38.360
## tipologiaAttico                          52615.63    5619.02   9.364
## tipologiaVilla                           -9051.60    5784.74  -1.565
## `tipologiaVilla unifamiliare`            13463.54    6292.89   2.139
## tipologiaAltro                          -37293.38    7763.73  -4.804
## Balcone                                  -5816.03    2676.02  -2.173
## Terrazza                                 -3871.86    2724.52  -1.421
## Giardino.comune                         -17431.21    3170.79  -5.497
## Infissi.esterni.in.vetro...legno         17489.01    3726.62   4.693
## Esposizione.doppia                       -9011.37    4004.70  -2.250
## VideoCitofono                            11138.29    3508.87   3.174
## Caminetto                                12800.35    4826.55   2.652
## Porta.blindata                            5789.70    3093.54   1.872
## Parzialmente.Arredato                    12894.58    3486.38   3.699
## Esposizione.esterna                      -6553.42    3364.16  -1.948
## Taverna                                 -16276.86    5888.42  -2.764
## Impianto.di.allarme                      20510.37    3899.60   5.260
## Mansarda                                -15008.40    4425.13  -3.392
## Esposizione.interna                     -19084.17    5492.57  -3.475
```

```
## Idromassaggio                                14504.33    7197.41   2.015
## Cucina                                      -23173.32    9192.22  -2.521
## sud                                         -55467.38   12759.34  -4.347
## Infissi.esterni.in.vetro...PVC              -21284.47   12300.30  -1.730
## Infissi.esterni.in.triplo.vetro...metallo    30596.86   17087.89   1.791
## Infissi.esterni.in.doppio.vetro...metallo   -15006.24    9514.96  -1.577
## Giardino.privato.e.comune                   -44570.70   29337.97  -1.519
## Reception                                   184997.01   54549.90   3.391
## Passo.carrabile                            -107895.50   55246.86  -1.953
##                                            Pr(>|t|)
## (Intercept)                                 < 2e-16 ***
## anno_costrAntico                            < 2e-16 ***
## anno_costrModerno                           < 2e-16 ***
## anno_costrSconosciuto                       < 2e-16 ***
## anno_costrVecchio                           < 2e-16 ***
## ascensore                                  1.07e-09 ***
## bagni2                                      < 2e-16 ***
## bagni3                                      < 2e-16 ***
## `bagni3+`                                   < 2e-16 ***
## `boxauto(13,50]`                           3.39e-10 ***
## classe_en                                  6.94e-07 ***
## climaAutonomo                              0.000372 ***
## comunePadova                                < 2e-16 ***
## comuneRovigo                                < 2e-16 ***
## comuneTreviso                               < 2e-16 ***
## comuneVenezia                               < 2e-16 ***
## comuneVerona                                < 2e-16 ***
## locali3                                    0.006852 **
## locali4                                    9.75e-13 ***
## locali5                                     < 2e-16 ***
## `locali5+`                                  < 2e-16 ***
## macrozona_id10962                           < 2e-16 ***
## macrozona_id168                             < 2e-16 ***
## macrozona_id368                            0.004271 **
## macrozona_idAltro                           < 2e-16 ***
## macrozona_idSconosciuto                     < 2e-16 ***
## piano2                                     0.001558 **
## pianoAltro                                 0.049720 *
## pianoSconosciuto                           5.68e-07 ***
## pianoT                                     0.005967 **
## riscaldamentoCentralizzato                 0.050294 .
## riscaldamentoSconosciuto                   0.000931 ***
## `statoDa ristrutturare`                    1.81e-15 ***
## `statoNuovo / In costruzione`               < 2e-16 ***
## `statoOttimo / Ristrutturato`               < 2e-16 ***
## superficie                                  < 2e-16 ***
## tipologiaAttico                             < 2e-16 ***
## tipologiaVilla                             0.117683
## `tipologiaVilla unifamiliare`              0.032426 *
## tipologiaAltro                             1.59e-06 ***
## Balcone                                    0.029779 *
## Terrazza                                   0.155320
## Giardino.comune                            3.97e-08 ***
## Infissi.esterni.in.vetro...legno           2.74e-06 ***
## Esposizione.doppia                         0.024463 *
## VideoCitofono                              0.001507 **
## Caminetto                                  0.008015 **
## Porta.blindata                             0.061305 .
## Parzialmente.Arredato                      0.000218 ***
## Esposizione.esterna                        0.051448 .
```

```
## Taverna                                       0.005719 **
## Impianto.di.allarme                           1.48e-07 ***
## Mansarda                                      0.000698 ***
## Esposizione.interna                           0.000514 ***
## Idromassaggio                                 0.043914 *
## Cucina                                        0.011722 *
## sud                                           1.40e-05 ***
## Infissi.esterni.in.vetro...PVC                0.083596 .
## Infissi.esterni.in.triplo.vetro...metallo 0.073401 .
## Infissi.esterni.in.doppio.vetro...metallo 0.114806
## Giardino.privato.e.comune                     0.128747
## Reception                                     0.000699 ***
## Passo.carrabile                               0.050857 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 11464806717)
##
##     Null deviance: 2.7624e+14  on 8297  degrees of freedom
## Residual deviance: 9.4413e+13  on 8235  degrees of freedom
## AIC: 215816
##
## Number of Fisher Scoring iterations: 2

print(model)

## Generalized Linear Model with Stepwise Feature Selection
##
## 8298 samples
##   61 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 6639, 6640, 6638, 6638, 6637
## Resampling results:
##
##   RMSE    Rsquared   MAE
##   108336  0.6483194  76548.71
```

In tutto i coefficienti ora sono 63, la selezione ha prodotto quindi ben 34 livelli in meno.

3.3. Albero di regressione

Per maggiore completezza abbiamo deciso di indagare anche possibili relazioni tra variabili indipendenti e la variabile risposta, diverse dai legami lineari considerati fino ad ora.

Per farlo si è scelto di utilizzare un albero di regressione.

L'algoritmo *CART (Classification And Regression Tree)* è stato teorizzato nel 1984 da Leo Breiman, Jeron Friedman, Richard Olshen e Charles Stone. Come si può evincere già dal nome, l'algoritmo consente di analizzare sia attributi continui che attributi discreti, producendo – a sua volta – un *output* continuo (regressione) o discreto (classificazione). Sfruttando una procedura non parametrica e il coefficiente di impurità di Gini come misura dell'utilità, l'algoritmo risulta robusto rispetto alla presenza di *outliers* e/o attributi con informazioni mancanti. Inoltre, la procedura del CART, oltre alla fase iniziale di crescita (*building*), prevede una fase di potatura (*pruning*), finalizzata a limitare problemi derivanti dall'*over-fitting*.

L'idea è anche quella di approfondire anche la presenza di possibili interazioni tra variabili (aspetto fino a questo momento trascurato), utilizzando un metodo interpretabile con relativa facilità.

L'albero è costituito da una struttura le cui componenti sono affermazioni di tipo logico, dette *nodi*, del tipo superficie < 139,5.

Si inizia esaminando l'affermazione riportata nel nodo alla radice dell'albero, che è posta in alto; se l'affermazione è vera si segue il ramo sottostante di sinistra, altrimenti quello di destra. Si procede con lo stesso schema, esaminando via via le affermazioni successive, fino a quando non si raggiungono i nodi terminali, dette foglie, che forniscono il valore della funzione in maniera approssimata.

Questo algoritmo sfrutta le variabili per costruire partizioni delle osservazioni, all'interno delle quali la risposta è molto simile (o poco variabile); stimando la risposta con la media delle risposte della partizione in cui cade.

In questo caso, pertanto, non ci si concentra più su effetti marginali delle esplicative, ma su effetti congiunti delle stesse, rappresentati da funzioni costanti a tratti (si stima la media per ogni tratto).

Vista la natura dell'approssimazione ci si aspettano risultati non ottimali qualora la relazione tra risposta e concomitanti sia effettivamente prevalentemente lineare e non siano presenti forti effetti di interazione.

Per evitare di costruire un algoritmo che approssimasse eccessivamente bene il comportamento dei dati nel *train*, inglobandone anche il rumore, dopo aver fatto *crescere* l'albero, anche qui viene utilizzata la convalida incrociata per *potare* l'albero, ovvero per ridurne il numero di foglie in corrispondenza dell'errore minimo. Il *training set*, in questo caso, è stato suddiviso in k=10 parti.

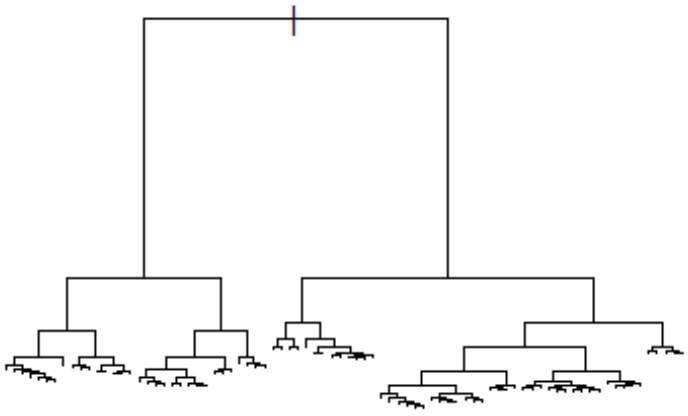

Figura 13. Albero di regressione "cresciuto" con 82 foglie.

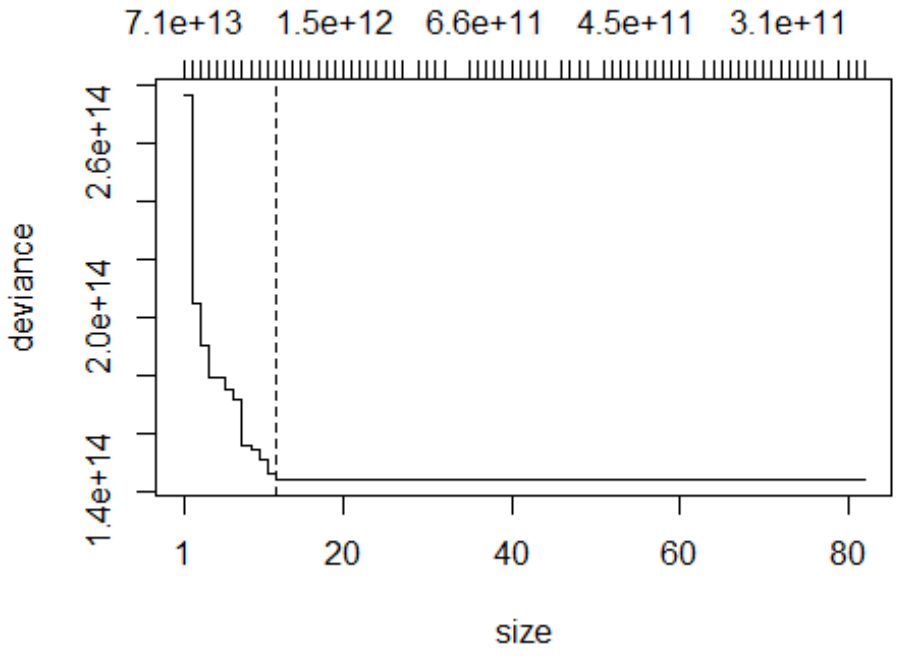

Figura 14. Devianza dell'albero di regressione.

Queste operazioni portano alla costruzione di un albero potato con 12 foglie.

I risultati evidenziati dal grafico sembrano abbastanza concordi con quanto emerso dalle analisi precedenti.

La superficie sembra avere un peso determinante nello spiegare il prezzo e il suo effetto cambia in relazione al comune considerato, al numero di bagni, all'anno di costruzione e alla macrozona.

In particolare, è interessante come il numero di bagni sembri importante soprattutto per le abitazioni con superficie minore.

Mentre per gli immobili più grandi, il primo fattore diventa il comune in cui è ubicato.

Di seguito l'output di R e il grafico dell'albero ottimale.

```
##
## Regression tree:
## snip.tree(tree = mt1, nodes = c(59L, 57L, 21L, 15L, 58L, 11L,
## 8L, 9L, 12L, 13L, 56L, 20L))
## Variables actually used in tree construction:
## [1] "superficie"   "bagni"        "anno_costr"   "macrozona_id" "comune"
## Number of terminal nodes:  12
## Residual mean deviance:  1.647e+10 = 1.365e+14 / 8286
## Distribution of residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -577200  -72330  -16750       0   53110  742700
```

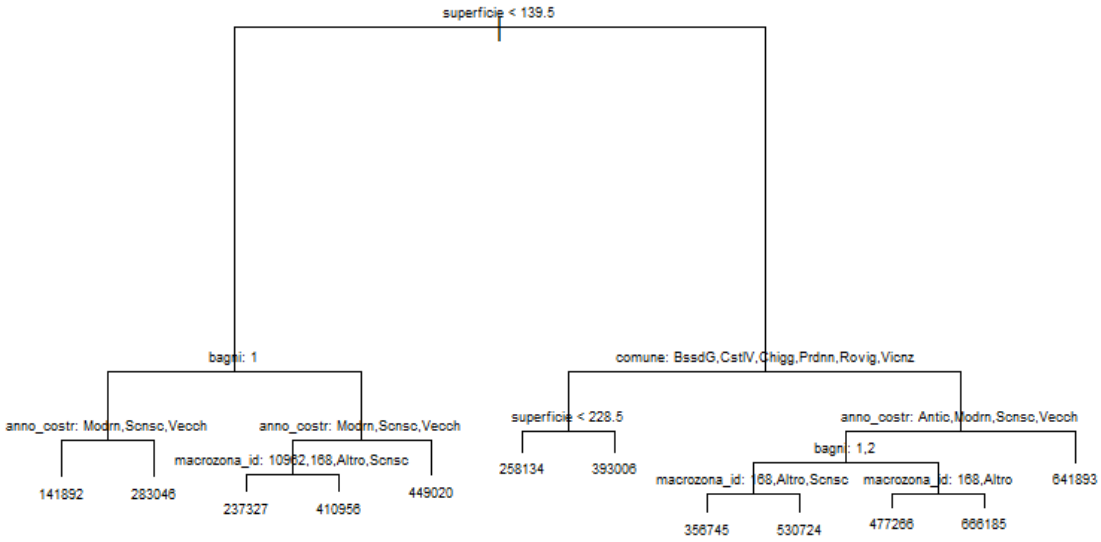

Figura 15. Albero di regressione "potato", con 12 foglie.

4. Conclusioni

Valutiamo la performance dei modelli sulla base dell'errore di previsione.

Abbiamo utilizzato come metriche di confronto lo scarto quadratico medio (*RMSE, Root Mean Squared Error*) e l'errore assoluto medio (*MAE, Mean Absolute Error*). Sono calcolati chiaramente, per ogni modello, sullo stesso test.

Lo scarto quadratico medio è una misura dell'errore assoluto in cui gli errori sono al quadrato per evitare che valori positivi e negativi si annullino a vicenda. Rappresenta la deviazione standard dei residui. Il termine residuo si riferisce alla distanza tra il punto previsto e il punto osservato.

L'RMSE è influenzato da valori anomali (*outliers*), ovvero valori di dati che sono anormalmente distanti dalla vera retta di regressione. Per definizione, l'errore al quadrato per punti così distanti sarà molto alto.

MAE è simile a MSE in quanto restituisce i valori assoluti dei residui, ma senza l'elevazione al quadrato. Non considera la direzione dell'errore, il che significa che non sapremo se gli errori negativi o positivi pesano di più sulla media complessiva. Detto questo, MAE è più robusto ai valori anomali proprio grazie all'assenza dell'elevazione al quadrato dei valori degli errori di previsioni lontane.

Aggiungiamo nella tabella sottostante, nelle ultime due righe, anche due modelli di previsione utilizzando, in uno, solo la variabile superficie come esplicativa e, nell'altro, un modello in cui ci si confronta semplicemente con il prezzo medio.

Tipo modello	*RMSE*	*MAE*
Lineare	**104336**	75380
Lineare con Selezione	104389	**75311**
Albero	128932	94020
Superficie	150705	112448
Prezzo medio	173597	134032

Tabella 26. Modelli previsivi a confronto.

Il modello che risulta migliore a livello previsionale dipende quindi dal metodo utilizzato per valutarne l'errore.

Il modello lineare e il modello lineare con selezione si comportano in maniera simile.

Per la ricerca della semplicità e della parsimonia, si rivela più ragionevole utilizzare il modello lineare con selezione, che ha ben 34 parametri in meno.

L'albero di regressione si comporta in maniera peggiore. Probabilmente la media a tratti semplifica troppo la questione? Ricordiamo anche che i nodi terminali sono 12.

Gli altri due modelli sono stati immessi solo come confronto, da utilizzare quindi, in un certo senso, come base di partenza.

Un possibile miglioramento futuro sarebbe quello di valutare l'inserimento di effetti interazione nel modello lineare, magari anche sfruttando le indicazioni fornite dall'albero (come, ad esempio, il bagno con la superficie, o il comune con la superficie).

I modelli possono ritenersi soddisfacenti, soprattutto tenendo conto delle difficoltà operative nella sua applicazione, date dalle peculiarità del mercato immobiliare, che è di per sé molto eterogeneo.

Approfondire in un prossimo studio la descrizione dell'annuncio potrebbe aiutare nella modellazione. Parole-chiave come "vista mare", "vista montagna" e simili, potrebbero risultare influenti.

Ci sono poi molte caratteristiche estrinseche ambientali che non si ricavano con facilità dal dataset, come la qualità del quartiere in cui si trova l'immobile, la viabilità e l'accessibilità, la presenza di elementi di pregio ambientale (parchi, giardini pubblici e privati, corsi d'acqua, ecc.); o anche l'assenza di elementi di degrado (discariche, cave, traffico, inquinamento, ecc.).

Bibliografia

Azzalini A. e Scarpa B., *Analisi dei dati e data mining*, Milano, Springer, 2004.

Grigoletto M., Ventura L., Pauli F., *Modello lineare. Teoria e applicazioni con R*, Torino, Giappichelli, 2017.

Legendre A. M., *Nouvelles méthodes pour la détermination des orbites des comètes*, F. Didot, Parigi, 1805.

Breiman L., Friedman J. H., Olshen R.A., Stone C.J. *Classification and regression trees*, New York, Chapman and Hall/CRC, 1984.

http://www.immobiliare-italiano.com/

https://www.notiziariofinanziario.com/

https://www.paolapozzolo.it/

https://www.blog.osservatori.net/

https://pulplearning.altervista.org/

https://www.diariodiunanalista.it/

https://www.andreaminini.com/

https://www.linkedin.com/pulse/machinelearning-4-dummies-gli-alberi-decisionali-marco-iannucci/

http://www.galenotech.org/calcoli/nozionidibase.htm