Classificazione sulla ristrutturazione degli immobili e predizione del prezzo delle case nella contea di King County.

Andrea Afify[a], Alessandro Risaro[a], Marta La Franca[a], Alessandra Maggipinto[a]

[a]*DISCo, Università degli Studi di Milano-Bicocca, VialeSarca 336, 20126 Milano, Italy.*

Abstract

In questo lavoro si utilizza un dataset della piattaforma Kaggle che include gli immobili venduti tra maggio 2014 e maggio 2015 nella contea di King County. Da una parte si applica al dataset un modello di classificazione per prevedere l'eventuale ristrutturazione di una casa; dall'altra si propongono modelli regressivi per prevedere il prezzo degli immobili. Il modello di classificazione viene effettuato tenendo conto del problema della class imbalance.

Keywords: Pricing, Classification Models, Class Imbalance, Regression models.

1. Introduzione

Riuscire ad elaborare modelli per stimare il prezzo delle case è di importanza fondamentale sia per chi deve vendere un immobile sia per chi lo deve acquistare. Sono stati infatti proposti in letteratura numerosi modelli predittivi per risolvere il problema del pricing degli immobili(si veda e.g. Brown et al. (1997),Limsombunchai (2004), Mohd et al. (2020)). La ristrutturazione di un immobile aumenta o preserva il suo valore e per questo è un elemento fondamentale per determinarne il prezzo. In questo lavoro si utilizza un dataset di Kaggle (*https* : *//www.kaggle.com/harlfoxem/housesalesprediction*) in cui sono presenti i prezzi delle case della contea di King County (USA) per elaborare modelli di classificazione e predittivi sul prezzo.

Il primo obiettivo di questo lavoro per il ruolo strategico che assume nel pricing di un immobile è quello di sviluppare un modello di classificazione in grado di prevedere quali immobili siano stati ristrutturati.

Infine il secondo obiettivo è stato quello di sviluppare modelli regressivi in grado di prevedere il prezzo degli immobili. Per raggiungere questo scopo si sono utilizzati e confrontati diversi modelli:

- regressione lineare multipla
- regressione logaritmica

Il lavoro è organizzato come segue: Nella Sez. 2 si descrive il dataset a nostra disposizione evidenziando le caratteristiche degli attributi. Nella Sez. 3 si descrive l'attività svolta per filtrare e assicurare una buona qualità dei dati per le analisi seguenti. Nella Sez. 4 si discute come si è ottenuta la classificazione delle case ristrutturate affrontando il problema della class imbalance. Nella Sez. 5 si discutono in dettaglio i risultati ottenuti mediante regressione lineare e non lineare per la previsione del prezzo degli immobili. Inoltre se ne valuta comparativamente l'efficacia. Il lavoro termina con le nostre note conclusive nella Sez. 6.

2. Descrizione del dataset

Il dataset in esame è composto da 21600 righe e da 21 variabili, le quali descrivono le caratteristiche delle case vendute:

- ID (String): ID unico per ogni proprietà venduta;
- Date (String): data in cui è stata venduta la proprietà;
- Price (Double): prezzo a cui è stata venduta ogni proprietà;
- Bedrooms (Integer): numero di camere da letto presenti nella proprietà;
- Bathrooms (Double): numero di bagni presenti nella proprietà (dove .5 rappresenta un bagno con WC ma senza doccia);
- Sqft_living (Integer): superficie in metri quadri dello spazio abitativo interno a gli appartamenti;
- Sqft_lot (Integer): superficie in metri quadri del lotto di terreno su cui è edificata la proprietà;
- Floors (Integer): numero di piani della proprietà;
- Waterfront (Integer): variabile dummy che indica se la proprietà si affacci o meno sul lungomare;
- View (Integer): indice compreso tra $[0, 4]$ che indica quanto sia buona la vista della proprietà;
- Condition (Integer): indice compreso tra $[1, 5]$ che esprime la condizione della proprietà;
- Grade (Integer): indice compreso tra $[1, 13]$ che esprime il livello di costruzione e di progettazione della proprietà;
- Sqft_above (Integer): superficie in metri quadri dello spazio abitativo interno che si trova sopra il livello del suolo;
- Sqft_basement (Integer): superficie in metri quadri dello spazio abitativo interno che si trova sotto il livello del suolo;
- Yr_built (String): L'anno in cui la proprietà è stata costruita;
- Yr_renovated (String): L'anno dell'ultima ristrutturazione della proprietà;
- Zipcode (String): Zipcode dell'area in cui si trova la proprietà;
- Lat (Double): latitudine;
- Long (Double): longitudine;
- Sqft_living15 (Integer): La media della superficie in metri quadri dello spazio abitativo interno delle 15 proprietà più vicine;
- Sqft_lot15 (Integer): La media della superficie in metri quadri dei lotti di terreno delle 15 proprietà più vicine.

3. Preprocessing

3.1. Data inconsistency

Durante la fase di esplorazione del dataset si è riscontrato un problema di inconsistenza dei dati:

- una proprietà presenta un valore della variabile *bathrooms* pari a 33 che confrontato con il valore della variabile *sqft_living* segnala un' inconsistenza dei dati;
- diciotto proprietà presentano valori delle variabili *bedrooms* e/o *bathrooms* pari a 0.

Dato lo scarso numero di righe coinvolte si è proceduto alla rimozione delle stesse dal dataset.

3.2. *Feature creation e feature elimination*

Si è proceduto poi alla creazione di tre nuove variabili ritenute utili ai fini del lavoro svolto:

- Year_sale (Integer): è stato estratto l'anno dalla variabile Date;
- Age_house (Integer): dalla differenza tra le variabili Year_sale e Year_built abbiamo estratto l'età dello stabile;
- Renovated (String): una variabile binaria che assume valori {y,n} a seconda che lo stabile sia stato ristrutturato o meno.

Dei 21 attributi iniziali si è deciso di eliminare gli attributi *lat*, *long* e *zipcode*, in quanto non utili ai fini della nostra analisi.

4. Classificazione

4.1. *Esplorazione dei dati*

In questa fase della analisi ci si è concentrati sull'esplorazione dei dati, in particolare, si è controllato come fosse distribuita la variabile "*renovated*":

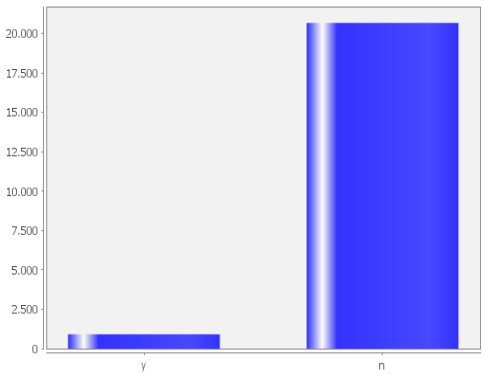


Figura 1: In figura si mostra lo sbilanciamento della classe.

La Fig.(1) mostra chiaramente lo sbilanciamento della classe, pertanto si è reso necessario adottare le tecniche per trattare una classe sbilanciata. Di seguito, dopo aver descritto come si è effettuata la fase di feature selection, si propongono le tre strategie adottate con gli eventuali problemi e rischi riscontrati.

4.2. *Modelli utilizzati*

Per approciare il problema della classificazione si è deciso di confrontare vari modelli e tecniche con lo scopo di individuare la più adatta al nostro problema. I modelli confrontati sono i seguenti:

- Modelli euristici tra cui Decision Tree (J48 e il Knime Decision Tree) e Random Forest che prevedono l'attributo di classe attraverso la costruzione di alberi di decisione.
- Modelli probabilistici tra cui il Naive Bayes (NBTree) il quale si basa sul teorema di Bayes per prevedere le classi.
- Regressione logistica.

4.3. *Valutazione della perfomance*

In quanto il nostro dataset risulta essere sbilanciato rispetto alla variabile di classe (renovated) l'accuratezza non viene utilizzata nella nostra analisi. Perciò si sono tenuti in considerazione altri indicatori come:

- Recall: indica la porzione di record positivi classificati correttamente dal nostro modello.Un alto valore della recall indica che la maggior parte delle osservazioni con classe positiva sono state previste in modo corretto;
- Precision: indica la frazione di record che sono effettivamente positivi tra tutti quelli previsti come tali;
- F1 measure: indica la media armonica tra precision e recall, consente di fornire un'interpretazione delle due metriche più ragionevole, infatti un

valore elevato di questo indice garantisce un giusto bilanciamento tra le due.

4.4. Feature selection

Per prima cosa si è effettuata la feature selection al fine di selezionare gli attributi più significativi e migliorare l'interpretabilità del modello. La feature selection è stata condotta separatamente per le tre strategie adottate e in base alla complessità computazionale e all'efficienza si sono usati filtri o wrapper.

In particolare quando si è utilizzato l'approccio dell'undersampling, essendo diminuite notevolmente le istanze a disposizione, si è scelto di utilizzare il wrapper perchè questa tecnica sebbene pesante computazionalmente è generalmente più efficace dei filtri. Utilizzando il wrapper J48 si è trovato il seguente sottoinsieme di attributi: *bathrooms*, *floors*, *waterfront*, *view*, *grade*, *yr_built*, *age_house*

Per l'approcio dell'oversampling, essendo aumentate notevolmente le istanze a disposizione, si è scelto di confrontare alcuni filtri e scegliere quello che minimizzasse l'errore, si è quindi trovato che il più conveniente fosse il filtro multivariato J48(Cfs). Utilizzando questo filtro si è trovato il seguente sottoinsieme di attributi: *bedrooms*, *bathrooms*, *floors*, *condition*, *grade*, *yr_built*, *age_house*, *yr_sale*.

Infine si è riscontrata una inefficienza nella procedura di feature selection per il cost sensitive approach; la strategia utilizzata è stata dunque quella di selezionare manualmente gli attributi che si rivelano essere i più sensati per valutare una potenziale ristrutturazione. Gli attributi selezionati sono dunque i seguenti: *bedrooms*, *bathrooms*, *floors*, *condition*, *grade*, *yr_built*, *age_house*, *yr_sale*.

4.5. Undersampling

Tramite il metodo dell' Holdout è stato suddiviso il dataset iniziale in due partizioni disgiunte (assegnando il 70% delle osservazioni al training set e il 30% delle osservazioni al test set). Successivamente ai dati di training è stato applicato il campionamento stratificato equal size (utilizzando il nodo knime equal size sampling). Le criticità relative all'applicazione di questo metodo riguardano la riduzione della dimensionalità del training set che potrebbe portare ad una situazione di overfitting. Al fine di valutare quale modello fosse più adatto per l'analisi in esame si sono valutate le misure di performance (accuracy, precision, recall e F1 measure) per la previsione della classe rara. Di seguito si riportano i risultati ottenuti:

Tabella 1: Under Sampling Approach: precision, recall, F1 measure. Il pedice n indica le previsioni per le istanze "no", il pedice y indica le previsioni per le istanze "yes" della variabile "renovated".

Algoritmo	Precision	Recall	F1 Measure
Decision Tree_n	0.763	0.99	0.862
Decision Tree_y	0.818	0.132	0.228
J48_n	0.77	0.99	0.866
J48_y	0.821	0.136	0.234
Random Forest_n	0.767	0.988	0.864
Random Forest_y	0.788	0.13	0.223
Logistic_n	0.797	0.988	0.882
Logistic_y	0.777	0.145	0.244
Simple Logistic_n	0.998	0.96	0.979
Simple Logistic_y	0.055	0.577	0.1
NBTree_n	0.701	0.989	0.821
NBTree_y	0.828	0.109	0.193

Gli algoritmi di classificazione non hanno presentato valori di F1 Measure significativi.

4.6. Oversampling

Tramite il metodo dell' Holdout è stato suddiviso il dataset iniziale in due partizioni disgiunte (assegnando il 60% delle osservazioni al training set e il 40% delle osservazioni al test set). Successivamente ai dati di training è stato applicata la tecnica conosciuta in letteratura come: "Synthetic Minority Oversampling TEchnique" (SMOTE) (si veda e.g. Chawla et al. (2002)) utilizzando il nodo Knime SMOTE. Le criticità relative all'applicazione di questa tecnica riguardano la possibilità concreta di perdita di informazione dovuta alla creazioni di nuovi record articiali per bilanciare la classe meno numerosa. Di seguito si riportano i risultati ottenuti:

Tabella 2: Over Sampling Approach: precision, recall, F1 measure. Il pedice n indica le previsioni per le istanze "no", il pedice y indica le previsioni per le istanze "yes" della variabile "renovated".

Algoritmo	Precision	Recall	F1 Measure
Decision Tree_n	0.971	0.973	0.972
Decision Tree_y	0.38	0.365	0.372
J48_n	0.98	0.971	0.976
J48_y	0.347	0.438	0.387
Random Forest_n	0.981	0.972	0.976
Random Forest_y	0.355	0.453	0.398
Logistic_n	0.836	0.989	0.906
Logistic_y	0.792	0.176	0.288
Simple Logistic_n	0.835	0.989	0.906
Simple Logistic_y	0.792	0.176	0.288

In questo caso il classificatore random forest presenta delle performance più elevate rispetto agli altri modelli, portando ad un livello di F1 measure pari a circa 0,4.

4.7. Cost Sensitive Approach

Il terzo metodo utilizzato per affrontare il problema della classe sbilanciata è il Cost Sensitive Approach, che permette di bilanciare il dataset applicando pesi specifici ad ogni valore della matrice di confusione. In particolare sono stati assegnati i seguenti costi: TN = -10, FN = 100, FP = 30, TP = -1. Il costo di classificare un falso negativo è stato considerato maggiore rispetto a quello di classificare un falso positivo; è stato aggiunto un costo negativo per i veri positivi ed i veri negativi. Di seguito si riportano i risultati ottenuti sia in termini di Recall, Precision, F1-measure, sia in termini di costo:

Tabella 3: Cost Sensitive Approach: precision, recall, F1 measure. Il pedice n indica le previsioni per le istanze "no", il pedice y indica le previsioni per le istanze "yes" della variabile "renovated".

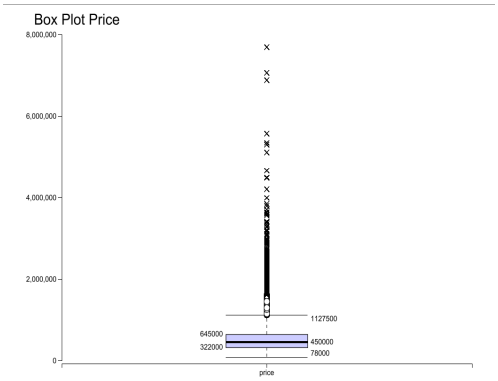
Algoritmo	Precision	Recall	F1 Measure
Decision Tree_n	0.969	0.973	0.971
Decision Tree_y	0.388	0.355	0.371
J48_n	0.983	0.973	0.978
J48_y	0.377	0.496	0.429
Random Forest_n	0.985	0.971	0.978
Random Forest_y	0.339	0.504	0.405
Logistic_n	0.983	0.97	0.977
Logistic_y	0.322	0.461	0.379
Simple Logistic_n	0.985	0.97	0.977
Simple Logistic_y	0.311	0.471	0.375
NBTree_n	0.986	0.967	0.977
NBTree_y	0.246	0.437	0.315

I costi trovati per Decsion Tree, J48,Random Forest, Logistic, Simple Logistic, NBTree sono rispettivamente pari a 127639, 132318, 128206, 128226, 127894 e 127582.

Il classificatore J48 presenta delle performance più elevate rispetto a gli altri modelli in termini di F1 Measure, ma risulta avere un costo elevato. Mentre il classificatore NBTree risulta essere quello con costo minore.

5. Predizione del prezzo

L'obiettivo di questa sezione è predire il prezzo delle case nella contea di King County basandosi su un modello di regressione. La tecnica dell' Holdout è stata eseguita partizionando il Dataset in due parti (Training Set e Test Set), utilizzando la divisione 2/3 e 1/3 rispettivamente con il metodo Linear Sampling. Durante la fase di esplorazione è stata osservata una forte presenza di outliers per la variabiale *Price*.



Tuttavia agli outliers per la variabile *Price* corrispondono outliers per le variabili *Bedrooms*, *Bathrooms*, *Sqft_living*, *Grade*, *Condition* e *View*. (fig.2)

Per questo motivo si è ipotizzato che gli outliers presenti siano rappresentativi e non siano dovuti ad errori di compilazione del Dataset, si è dunque deciso di non escluderli dal modello. Inizialmente è stata analizzata la matrice di correlazione lineare da cui è possibile evincere come gli attributi a disposizione nel dataset siano correlati con la variabile *Price*. In particolare le variabili con coefficiente di correlazione $\rho > 0.5$ sono risultate essere: *Bathrooms*, *Sqft_living*, *Grade*, *Sqft_above*, *Sqft_living* (fig 3).

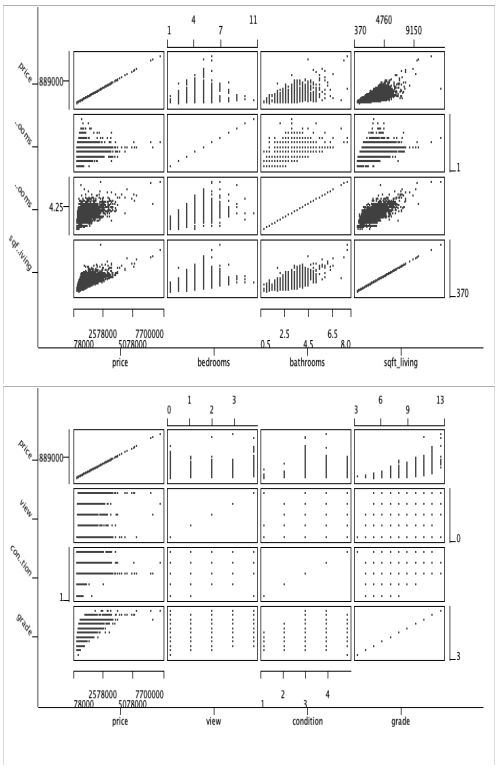


Figura 2: Scatter Plot

Si è dunque sviluppato un modello di regressione lineare multivariato utilizzando questi attributi. Per valutare la qualità e l'accuratezza predittiva del modello ci siamo soffermati sul coefficiente Adjusted R-Squared che risulta essere pari a 0.5522 (fig 4).

Non soddisfatti del modello ottenuto si è deciso di costruire un nuovo modello di regressione lineare utilizzando un nodo del software R prendendo in considerazione tutti gli attributi disponibili. Dal summary si evince che gli attributi *age_house*, *sqft_basement*, *sqft_lot* non danno un grande contributo (presentano un P-Value alto), inoltre *sqft_above*, oltre ad essere poco significativo, è fortemente collineare con *sqft_living*, quindi si è

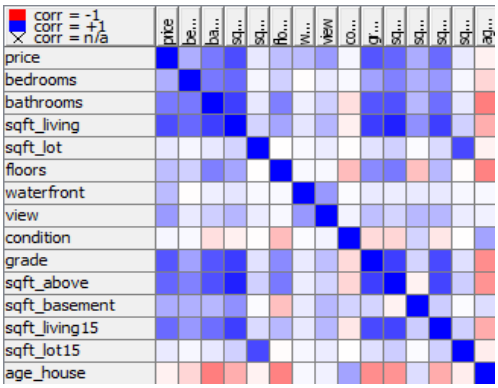


Figura 3: Correlazione

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
bathrooms	-34,845.594	4,285.5258	-8.131	4.44E-16
sqft_living	261.1335	5.6224	46.4455	0.0
grade	110,653.824	3,078.4907	35.9442	0.0
sqft_above	-85.2792	5.5405	-15.3919	0.0
sqft_living15	11.5764	5.0593	2.2881	0.0221
Intercept	-646,228.3999	16,870.7092	-38.3048	0.0

Multiple R-Squared: 0.5421
Adjusted R-Squared: 0.5419

Figura 4: Primo modello di Regressione Lineare

deciso di escluderli. In questo modello si è ottenuto un Adjusted R-squared pari a 0.6537. Sono state analizzate graficamente le ipotesi di normalità e indipendenza dei residui. (fig 5)

Il normal Q-Q plot dei residui standardizzati verifica l'assunzione della normalità della componente erratica del modello lineare. Quanto più i punti che rappresentano i residui ordinati giacciono in prossimità della linea Q-Q, tanto più plausibile è tale assunzione. In questo grafico i residui non presentano un andamento lineare, ma presentano un andamento esponenziale. Si è deciso di procedere allora con una regressione logaritmica cambiando il prezzo con il corrispettivo valore logaritmico. Nel modello ottenuto sono stati utilizzati gli attributi di quello precedente, con l'aggiunta di *sqft_lot* che è invece risultato significativo. Il coefficiente Adjusted

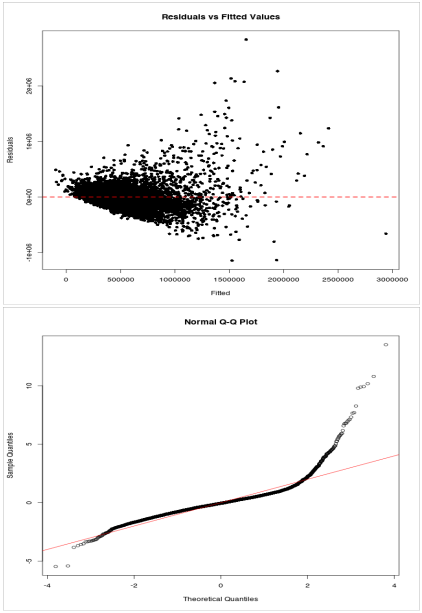


Figura 5: Grafici del modello di regressione applicato sul Test Set

R-squared è risultato pari a 0.6538, i grafici dei residui sono invece visibilmente migliorati (fig. 6).

Si è allora deciso di considerare tale modello il più valido tra quelli considerati, inoltre il coefficiente Adjusted R-squared compreso tra 0.5 e 0.7 rappresenta un effect size di moderata entità. Infine, una volta scelto il modello, si è proceduto a verificare l'ipotesi effettuata inizialmente sugli outliers, per fare ciò è stata utilizzata *la distanza di Cook* che misura l'influenza di ciascuna osservazione sulla stima dei parametri del modello (fig. 7). Osservando che nessun valore risulta essere maggiore dell'unità si è deciso di mantenere gli outliers come precedentemente ipotizzato.

6. Conclusioni

Riuscire a determinare, utilizzando modelli di classificazione, quali case siano state ristrutturate è stato complicato a causa del forte sbilanciamento della classe. Le tecniche utilizzate, le più frequenti tra quelle trovate in

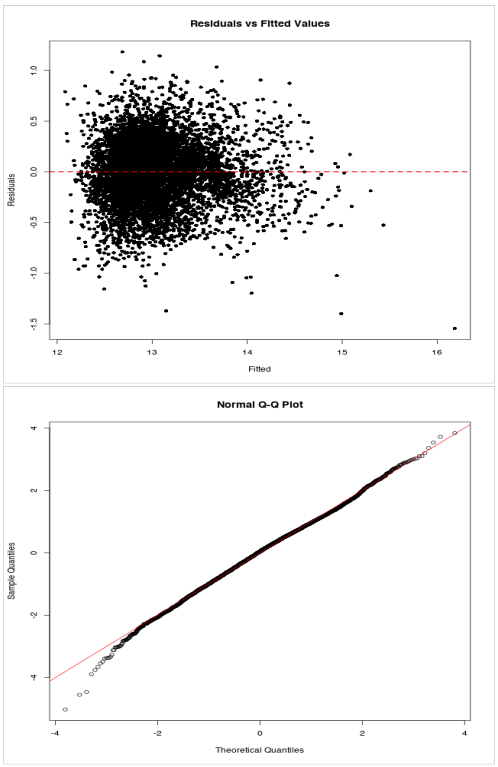


Figura 6: Grafici del modello di regressione applicato sul Test Set

letteratura, hanno rivelato alcuni punti deboli che hanno inficiato sulla bontà dei modelli. In particolare l'undersampling e l'oversampling si sono rilevati utili a ribilanciare artificialmente la classe rara, ma avendo allenato l'algoritmo su una classe ribilanciata, in fase di previsione sul test set si ha avuto una conseguente e notevole perdita di efficienza. Il terzo approccio adottato pur portando a migliori risultati sul test set, grazie ai termini di penalizzazione introdotti per tener conto dello sbilanciamento, ha comportato costi molto alti. Inoltre cercando di aumentare la precision si ottiene una conseguente diminuzione della recall e viceversa. La soluzione implementata è stata quella che ha permesso di bilanciare i due effetti.

In secondo luogo si è cercato di sviluppare un buon modello per predire il prezzo di una casa in vendita. Si

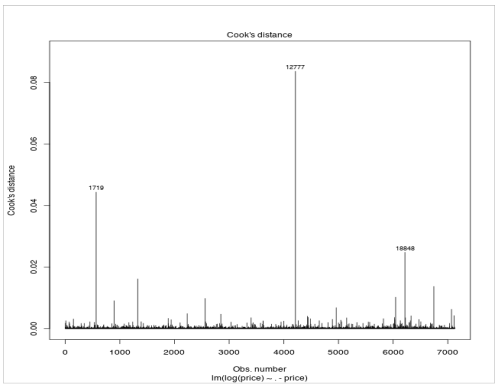


Figura 7:

sono creati due diversi modelli di regressione arrivando alla conclusione che la migliore predizione del prezzo si ottenesse con il modello di regressione logaritmico. Da questo modello si evince che il prezzo di una casa in vendita non è influenzato dall'età della casa o dalla presenza o meno di una cantina o di un seminterrato. Inoltre si è notato che un numero elevato di bagni e di stanze, un'ottima qualità della vista, i metri quadri interni, un'ottima condizione della casa e un'alta qualità di progettazione portano ad un aumento particolarmente significativo del prezzo.

Riferimenti bibliografici

Brown, J.P., Song, H., McGillivray, A., 1997. Forecasting uk house prices: A time varying coefficient approach. Economic Modelling 14, 529–548.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 321–357.

Limsombunchai, V., 2004. House price prediction: hedonic price model vs. artificial neural network, in: New Zealand agricultural and resource economics society conference, pp. 25–26.

Mohd, T., Jamil, N.S., Johari, N., Abdullah, L., Masrom, S., 2020. An overview of real estate modelling techniques for house price prediction. Charting a Sustainable Future of ASEAN in Business and Social Sciences , 321–338.