Caso di studio: il mercato immobiliare nella contea di King

Beatrice Barzaghi [1,2], Lorenzo Biagiotti [1,3], Guglielmo Muoio [1,4] e Laura Rapino [1,5]

[1] *Università degli Studi di Milano Bicocca - CdLM in Data Science*
[2] *b.barzaghi3@campus.unimib.it - matricola 831829*
[3] *l.biagiotti1@campus.unimib.it - matricola 825850*
[4] *g.muoio@campus.unimib.it - matricola 826029*
[5] *l.rapino@campus.unimib.it - matricola 831346*

18 Gennaio 2021

Sommario

L'obiettivo principale dello studio condotto è quello di creare un modello predittivo che permetta, data una collezione di variabili di *input*, di prevedere il costo di vendita di una casa nella zona di Seattle, più precisamente nella contea di King. Il dataset utilizzato è stato scaricato dal sito Kaggle ed è stato sviluppato un *workflow* KNIME attraverso il quale sono stati applicati diversi approcci e metodi al fine di creare un modello di regressione sufficientemente valido.

Indice

1 Introduzione

Lo scopo dello studio è quello di indagare quali siano i fattori e gli elementi caratteristici di una abitazione che ne aumentano il valore per sviluppare un modello predittivo in grado di stimare il prezzo di vendita di una casa nella città metropolitana di Seattle.

Il dataset scelto per effettuare questa indagine è reperibile sulla piattaforma Kaggle "*House Sales in King County, USA*" [2]. La contea di King, *King County*, si trova negli Stati Uniti, più in particolare, nello Stato di Washington. Ha una popolazione di 1.931.249 abitanti ed una superficie totale di 2.307 miglia quadrate (5.980 km^2) di cui 2.116 miglia quadrate (5.480 km^2) di terreno e 191 miglia quadrate (490 km^2) di acqua [4],[6]. Il capoluogo della contea è Seattle, città più popolosa di tutto lo Stato di Washington. Secondo la rinomata rivista "*U.S. News & World Report*", la King County si classifica al 333-esimo posto su cinquecento in una classifica complessiva basata su punteggi attribuiti a diverse categorie e sottocategorie identificate dal comitato *National Committee on Vital and Health Statistics* [5]. Le categorie scelte dal comitato sono: salute della popolazione, uguaglianza, educazione, economia, mercato immobiliare, nutrizione, ambiente, sicurezza pubblica, vitalità della comunità e infrastrutture. Ogni categoria, che differisce nel peso con il quale il punteggio influirà alla determinazione della classifica, può vedersi attribuire un massimo di 100 punti [7]. Per quanto riguarda la categoria degli immobili, l'assegnamento del punteggio tiene conto della disponibilità, dell'accessibilità economica e della qualità delle case all'interno di una comunità. In questa categoria alla contea di King viene attribuito un punteggio complessivo di 51 punti.

L'impostazione del report è la seguente: le prossime due sezioni riguardano il dataset e la sua esplorazione. Successivamente viene trattato il preprocessing necessario allo studio. I modelli con gli approcci utilizzati e la valutazione dei risultati vengono presentati rispettivamente nella quinta e sesta sezione. A seguito di tutte le osservazioni fatte e delle performance ottenute, nell'ultima sezione vengono tratte le conclusioni dello studio.

2 Dataset

"*House Sales in King County, USA*" è un dataset che tratta i prezzi e le caratteristiche delle case vendute nell'area metropolitana di Seattle nel periodo compreso tra maggio 2014 e maggio 2015.
Prima di condurre lo studio del dataset e delle sue variabili è stato verificato che non ci fossero osservazioni duplicate. Questo non avviene e quindi il dataset oggetto di studio presenta 21.613 osservazioni con 21 variabili descritte come riportato nella Sezione 2.1.

2.1 Variabili

Si riporta l'elenco delle variabili che compongono il dataset con il loro tipo.

D: Double **S**: Stringa **I**: Intero

- **id**: **S** codice identificativo
- **date**: **S** data di vendita
- **price**: **D** prezzo di vendita
- **bedrooms**: **I** numero di camere da letto
- **bathrooms**: **I** numero di bagni
- **sqft_liv**: **D** metratura dello spazio abitativo interno
- **sqft_lot**: **D** metratura del terreno su cui si trova l'abitazione
- **floors**: **I** numero di piani
- **waterfront**: **S** variabile fittizia per stabilire se l'abitazione si trovi in prossimità di uno specchio d'acqua (mare, lago, fiume)
- **view**: **S** indice di qualità della vista che ha la proprietà, compreso tra 0 e 4
- **condition**: **S** indice di qualità della condizione della casa, compreso tra 1 e 5
- **grade**: **S** indice qualitativo compreso tra 1 e 13, dove 1-3 corrisponde ad un livello di costruzione e progettazione di bassa qualità, 7 corrisponde ad un livello medio e 11-13 ad un livello di qualità elevato
- **sqft_above**: **D** metratura dello spazio abitativo interno sopra il piano terra
- **sqft_basmnt**: **D** metratura dello spazio abitativo interno interrato
- **yr_built**: **I** anno di costruzione
- **yr_renov**: **S** anno di restauro, 0 se la casa non è mai stata ristrutturata
- **zipcode**: **S** zipcode, codice postale del luogo in cui si trova l'abitazione
- **lat**: **D** latitudine
- **long**: **D** longitudine
- **sqft_liv15**: **D** metratura media dello spazio abitativo considerando le 15 abitazioni più vicine
- **sqft_lot15**: **D** metratura media dei lotti di terreno delle 15 abitazioni più vicine

Tutte le variabili categoriali ordinali presenti nel dataset sono state trattate come stringhe (S). Per rendere più facile la trattazione è stato deciso di trasformare la variabile *date* in tre differenti variabili: *day*, *month* e *year*. Queste tre nuove variabili sono poi state trattate come variabili categoriali.

3 Esplorazione ed elaborazione dei dati

Il dataset considerato non presenta valori mancanti per nessuna variabile, per cui non si sono avuti problemi di imputazione e/o eliminazione di osservazioni. I prezzi delle case oscillano tra un valore di 75.000$ ed uno di 7.700.000$, andando così a costituire un *range* significativamente ampio. Ciò è dovuto alla presenza di outliers (precisamente 1.140), presenti soprattutto nei percentili più elevati, come si evince dalla Figura 1. Dal punto di vista della correlazione, si è notato un forte legame tra il prezzo di vendita e le variabili relative alla metratura dell'abitazione.

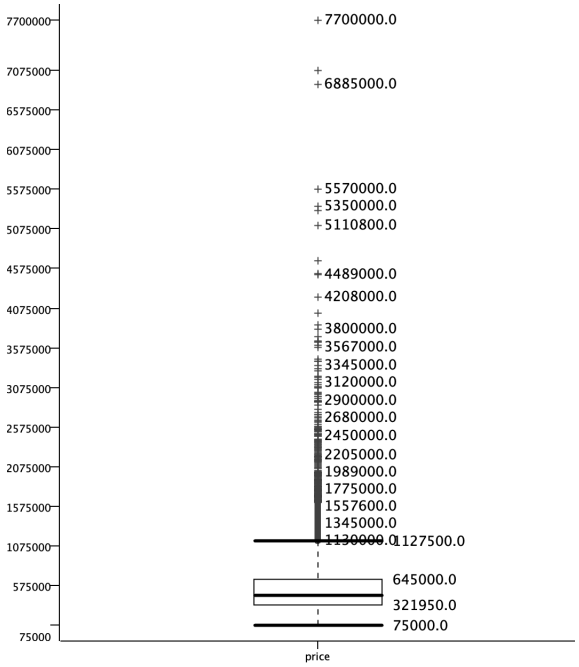

Figura 1: *Boxplot - Distribuzione della variabile risposta price*

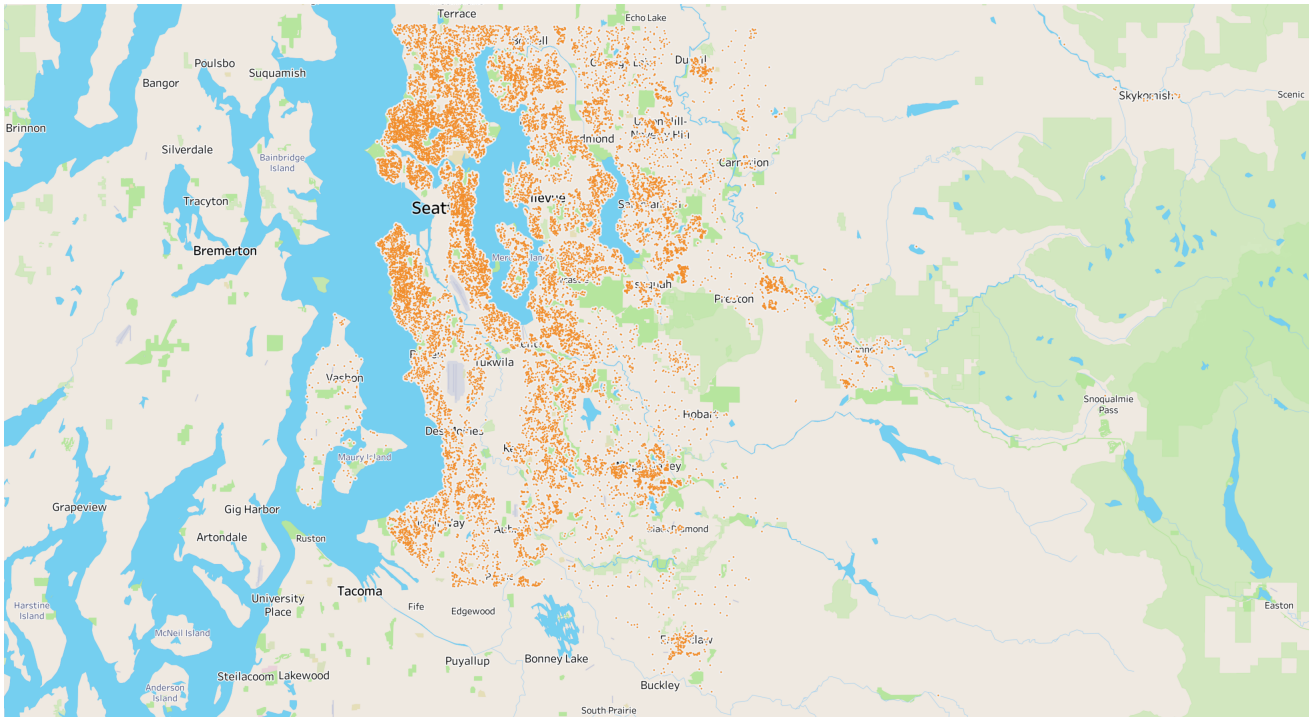

Figura 2: *Mappa - Visualizzazione delle abitazioni presenti nel dataset*

Le abitazioni considerate nel dataset hanno una media di 2080 *square feet* (ovvero $194\ m^2$) di spazio abitativo, 2 bagni e 3 stanze da letto.

4 Preprocessing

In seguito alle considerazioni presentate nel paragrafo precedente, riguardante l'esplorazione dei dati, le seguenti modifiche sono state apportate al dataset.
La prima operazione di preprocessing effettuata sul dataset è stata quella di escludere arbitrariamente la variabile *day* poichè non si tratta di un attributo interessante per lo sviluppo e lo studio dei modelli implementati in seguito. Questa modifica viene realizzata attraverso il nodo *Column Filter* di KNIME. Un'ulteriore modifica riguarda la variabile *yr_renov*. Si è deciso di applicare una ricodifica dove la nuova variabile assumerà valore 0 se la casa non ha subìto ristrutturazioni, valore 1 se la ristrutturazione è stata effettuata prima del 01.01.2000 e valore 2 altrimenti. In seguito si prosegue con una fase di *feature selection*, così da eliminare eventuali attributi non effettivamente utili per l'addestramento dei modelli. Al fine di ottenere questo, sono stati utilizzati due nodi KNIME, *Rank Correlation* e *Correlation Filter*, configurando quest'ultimo con una soglia di correlazione, *Correlation Threshold*, pari a $0,8$. Questa procedura viene effettuata solo su una parte del dataset, ovvero sul training set, in modo tale da evitare lo sfruttamento delle osservazioni appartenenti al test set che vengono considerate ipoteticamente come non note. Si ottiene che dei 21 attributi iniziali in input ne vengono scartati 3: *year*, *sqft_above* e *sqft_lot15*.

Poiché i nodi KNIME di regressione non consentono l'utilizzo di variabili categoriali con un numero relativamente elevato di classi è stato deciso di trasformare la variabile *zipcode* in una variabile binaria grazie all'approccio *one-hot encoding*, ovvero un approccio utilizzato per quantificare dati categoriali. In questo modo ogni codice postale rappresenterà un nuovo attributo, che assumerà valore pari a 1 nel caso in cui la casa appartiene a quel codice postale e valore 0 altrimenti. Per effettuare questa trasformazione in KNIME viene utilizzato il nodo *One to Many*, che aggiungerà 70 nuove colonne, essendo presenti 70 livelli differenti della variabile *zipcode*, al dataset ottenuto in seguito alla *feature selection* precedente. Alla fine di tutte queste operazione si ottiene un dataset di 87 attributi complessivi.

5 Modelli

Al fine di sviluppare un modello di regressione con lo scopo di prevedere il prezzo di vendita di una casa sono stati presi in considerazione 5 algoritmi diversi:

- Simple Linear Regression (SLR)
- Linear Regression (LR)
- Polynomial Regression (POLY)
- Simple Tree Regression (STREE)
- Random Forest Regression (RF)

Nel caso del modello **Simple Linear Regression**, implementato attraverso il nodo *SimpleLinearRegression* (WEKA 3.7), viene costruito un modello di regressione

lineare semplice basato su una sola variabile esplicativa (di tipo numerico) tra quelle disponibili in input. Il criterio a cui si fa riferimento per selezionare il singolo attributo è la minimizzazione dell'errore quadratico.

Il **Linear Regression**, realizzato attraverso il nodo *LinearRegressionLearner* (KNIME), rappresenta un modello di regressione lineare che sfrutta a pieno tutte le variabili esplicative disponibili in input, sia numeriche che categoriali. L'ottimizzazione dell'algoritmo si basa sulla minimizzazione dell'errore quadratico medio (MSE).

Una successiva evoluzione del modello di regressione lineare è il **Polynomial Regression**, eseguito tramite il nodo *PolynomialRegressionLearner* (KNIME), che, oltre alle variabili messe a disposizione, considera, dove possibile, anche l'elevamento di queste ultime ad un grado di potenza massimo scelto (in questo caso, quadratico).

Spostandoci verso algoritmi di tipo euristico, è stato preso in considerazione il modello **Simple Tree Regression**, implementato attraverso il nodo *SimpleRegressionTreeLearner* (KNIME). Questo metodo riproduce, con alcune semplificazioni, l'algoritmo descritto in "*Classification and Regression Trees*" [1].

Per ultimo si è sviluppato il **Random Forest Regression**, realizzato attraverso il nodo *RandomForestLearner(Regression)* (KNIME). Questo modello, rispetto ad un comune algoritmo basato su alberi decisionali, si impegna nella minimizzazione del problema di overfitting sul training set [3].

Ciascuno degli algoritmi presentati viene valutato utilizzando 3 approcci differenti: *holdout*, *cross-validation* e *feature construction and selection*.

5.1 Approcci

1. **Holdout**: l'approccio più semplice prende il nome *holdout*. Esso consiste nella ripartizione del dataset a disposizione in due *subsets*: training set e test test. Si tratta di due insieme disgiunti che, in questo studio, costituiscono rispettivamente l'80% e il 20% del dataset originale. Questa suddivisione è solitamente utilizzata per addestrare il modello sul training set per poi valutarlo su un insieme di dati, test set, la cui variabile di output non è nota (o considerata tale). Un *random seed* specifico è stato scelto a priori per garantire la riproducibilità dei risultati.

2. **Cross validation**: al fine di colmare i limiti e le debolezze dell'approccio precedente, si è voluto utilizzare anche un *K-folds cross validation*. Il dataset viene partizionato in K dataset disgiunti, con un numero quasi costante di osservazioni. In questo modo l'algoritmo si assicura che ogni record del dataset venga incluso nel training set lo stesso numero di volte (K-1) ed esattamente una volta nel test set. L'addestramento e la validazione vengono effettuate K volte, in modo da valutare la volatilità delle prestazioni sui test set. Si è scelto di porre K pari a 5, così da poter effettuare un confronto diretto con i risultati ottenuti tramite *holdout* e confermare o rifiutare eventuali considerazioni fatte.

3. **Feature construction and selection**: l'ultimo approccio utilizzato si basa su alcune assunzioni effettuate durante l'analisi esplorativa del dataset e su un particolare metodo di selezione delle variabili. Infatti, durante una prima visualizzazione grezza della distribuzione della variabile risposta *price*, si è notata una forte asimmetria positiva nell'istogramma (ovvero una coda lunga a destra). Questo fenomeno entra in contrasto con alcune ipotesi di base effettuate da vari algoritmi di regressione, in particolare il *Linear Regression*. Di conseguenza, per puro interesse, si è valutato di trasformare la variabile *price* attraverso una scala logaritmica (Figura 3). Utilizzando come variabile risposta la trasformata *log(price)*, si è implementato un modello di regressione lineare, sfruttando gli approcci sopra esposti, per valutare un confronto con i risultati ottenuti in precedenza.

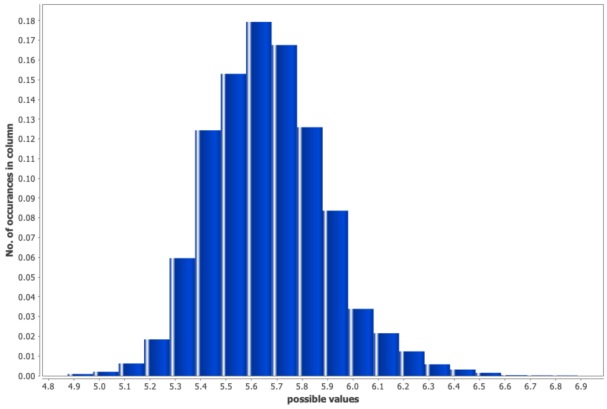

Figura 3: *Histogram Plot - Distribuzione empirica della variabile log(price)*

Infine, impiegando anche in questo caso la regressione lineare, si è voluta effettuare una selezione delle variabili attraverso una *Backward Feature Selection*. Il dataset originario è stato ripartizionato in training set, validation set e test set (seguendo la proporzione 60/20/20). Partendo dal modello "pieno", ovvero contenente tutte le variabili in input, si è implementato un ciclo (*Feature selection loop node*) tale per cui, ad ogni iterazione, viene addestrato il modello sul training set e valutato sul validation set. Successivamente la variabile meno significativa viene rimossa e il processo di addestramento e valutazione viene rieseguito fino all'esaurimento delle variabili. Una volta terminato il processo e collezionato i risultati di R^2, viene selezionato il modello con R^2 maggiore e, in caso di parità, quello con meno variabili. Una volta

riaddestrato il modello tramite fusione di training set e validation set, questo viene valutato sul test set creato precedentemente al ciclo.

6 Valutazione

Una volta definiti gli algoritmi e gli approcci da utilizzare, è stata implementata un'opportuna sequenza di nodi e metanodi sul *workflow* KNIME in modo tale da applicare quanto teoricamente spiegato finora con lo scopo di valutare i modelli proposti e perseguire al meglio l'obiettivo inizialmente posto. L'ordine con cui gli approcci sono valutati è il medesimo della Sezione 5.1.

6.1 Holdout

Un modello di regressione viene considerato buono se la variabile risposta prevista dal modello si avvicina il più possibile al valore reale. Nelle Figure 4 e 5 si osserva che il modello *Simple Tree* è quello che ottiene risultati migliori sul training set, tuttavia ha un elevato errore di generalizzazione che porta l'indice R^2 a diminuire, precisamente da 1 a $0,725$, e il valore di *RMSE* ad aumentare. L'overfitting sembra caratterizzare anche il modello *Polynomial Regression*, il cui valore di R^2 scala di $-0,108$ come si osserva dalla Tabella 1.

Si evince inoltre, dagli stessi due grafici, come i modelli meglio performanti siano *Random Forest* e *Linear Regression*, quest'ultimo in particolare poiché riporta un incremento dell'*RMSE* minore rispetto alla *Random Forest*.

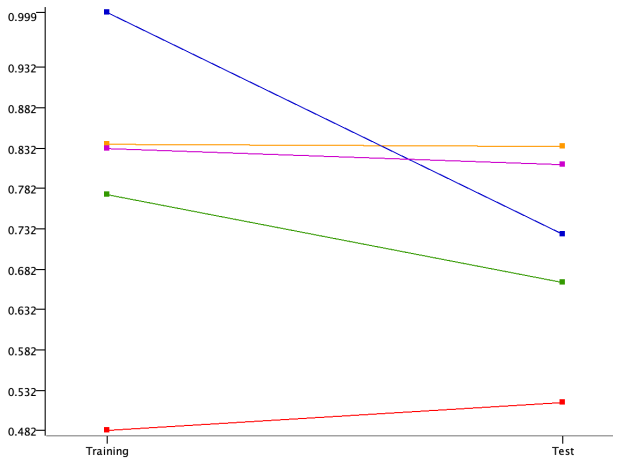

Figura 4: *Lineplot - Differenza dei valori di R^2 nei modelli con training e test.*
Blu: STREE, Arancione: LR, Viola: RF, Verde: POLY, Rosso: SLR

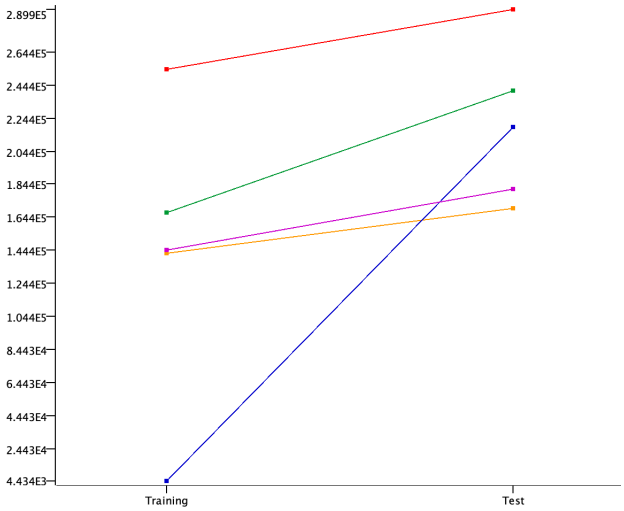

Figura 5: *Lineplot - Differenza dei valori di RMSE nei modelli con training e test.*
Blu: STREE, Arancione: LR, Viola: RF, Verde: POLY, Rosso: SLR

Parallelamente allo sviluppo di questi modelli, si è voluta eseguire una serie di operazioni diagnostiche, come ad esempio la verifica della normalità dei residui, con lo scopo di garantire la validità dei risultati ottenuti, prestando particolare attenzione al *Linear Regression*, algoritmo più performante tra tutti.
Nella Figura 6 viene riportato lo *scatterplot*, grazie al quale è possibile effettuare un confronto tra i valori reali e previsti dal modello *Linear Regression*. Si osserva che i valori sono concentrati sulla bisettrice per prezzi reali minori di $1.000.000$\$, mentre si discostano maggiormente per le case più costose. Infatti, per i prezzi più elevati, il *funnel plot* aumenta la sua ampiezza causando un discostamento maggiore dalla bisettrice per tali osservazioni. Questo avviene poiché le istanze si distribuiscono maggiormente nelle fasce di prezzo minori, per cui il modello tende a classificarle meglio.

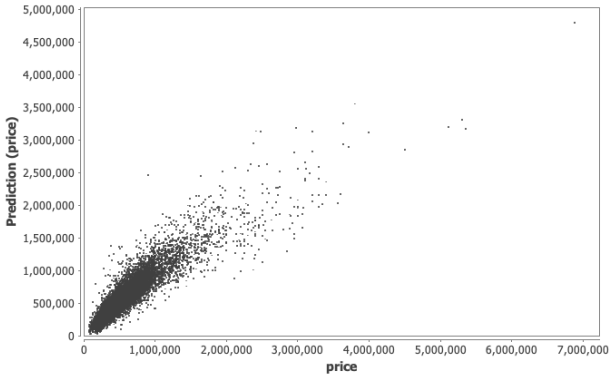

Figura 6: *Scatterplot - Confronto tra i valori reali e i valori previsti nel modello LR*

Al fine di valutare le assunzioni fatte dal modello, sono stati computati i residui e, a seguito di una operazione

di standardizzazione, sono state verificate le principali ipotesi. Nella Figura 7 viene riportato il *Q-Q plot*, che confronta la distribuzione cumulata dei residui con la distribuzione cumulata teorica, in questo caso la normale standard. Se i residui si distribuissero come una normale i punti di questo grafico dovrebbero addensarsi sulla diagonale che va dal basso verso l'alto e da sinistra verso destra. Si può notare come questo non avvenga e, oltretutto, viene riportato anche il *p-value* offerto dal test di Shapiro-Wilk, la cui ipotesi di normalità viene largamente rifiutata.

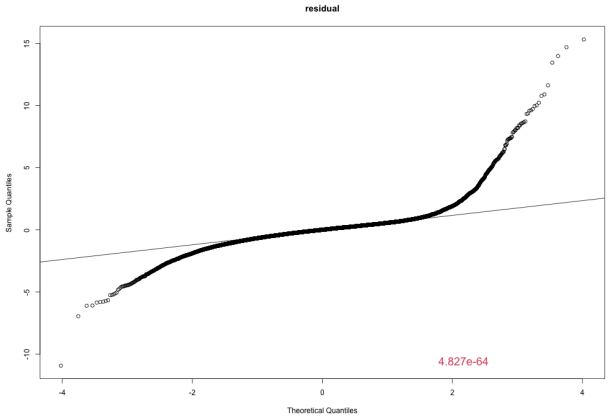

Figura 7: *Q-Q plot dei residui normalizzati (LR)*

In seguito si è verificata graficamente l'ipotesi di omoschedasticità dei residui. Nella Figura 8 si nota come la varianza dei residui dipenda fortemente dalla variabile risposta prevista dal modello. Infatti la forma assunta dai punti nel grafico, "ad imbuto", suggerisce una proporzionalità diretta tra varianza dei residui e il valore previsto dal modello quando contrariamente, le due grandezze dovrebbero risultare indipendenti.

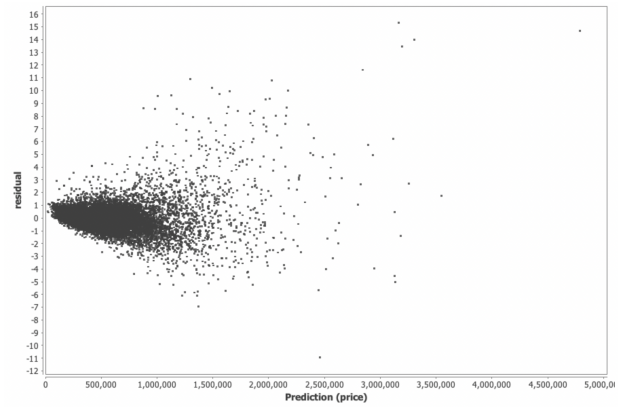

Figura 8: *Scatterplot - Valori previsti vs residui normalizzati (LR)*

Tabella 1: *Valori di R^2 e RMSE dei 5 modelli studiati*

	R^2		$RMSE$	
	Training	Test	Training	Test
SLR	0,483	0,518	254.071	289.966
LR	0,837	0,835	142.530	169.777
POLY	0,775	0,667	167.512	241.134
STREE	1,000	0,725	4.434	218.919
RF	0,832	0,812	144.789	181.228

6.2 Cross Validation

Per ogni *fold* dei modelli "crossvalidati" sono stati calcolati i valori di R^2 e dell'*RMSE*. Nei *boxplot* rappresentati nelle Figure 9 e 10 si visualizzano graficamente i risultati di queste due metriche.

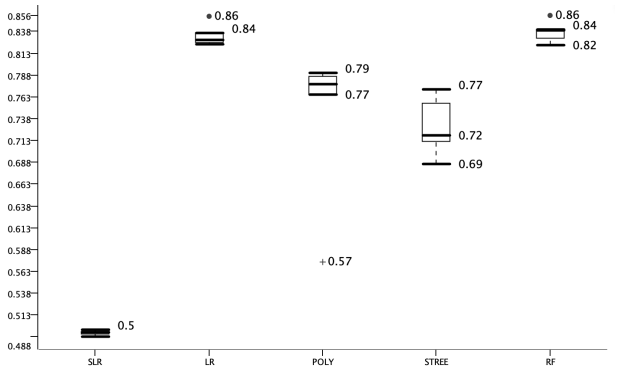

Figura 9: *Boxplot - Confronto dei valori di R^2 nei modelli cross-validati*

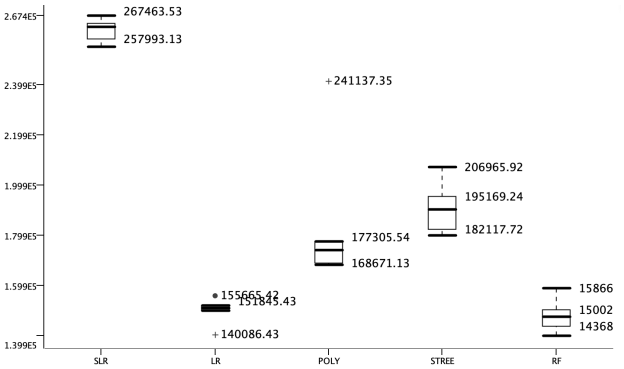

Figura 10: *Boxplot - Confronto dei valori di RMSE nei modelli cross-validati*

I valori ottenuti confermano i modelli *Linear Regression* e *Random Forest* come i migliori fra tutti. Più specificatamente, si nota che la variabilità dei valori di R^2 e *RMSE* è minore per questi modelli; infatti i valori dei due indici per ogni *fold* sono più simili tra loro, provando la bontà di questi modelli. Si osserva che il modello *Polynomial Regression* presenta un outlier con

valore di $R^2 = 0,57$ significativamente minore rispetto agli altri *fold* in cui il valore della metrica è compreso tra $0,77$ e $0,79$.

6.3 Feature construction and selection

Nella Figura 11 vengono confrontati valori reali e previsti dal modello *Linear Regression* a seguito di una trasformazione logaritmica della variabile *price*. Si può notare come le osservazioni si dispongano prevalentemente sulla bisettrice, discostandosi leggermente solo per gli estremi in cui i valori risultano essere meno concentrati.

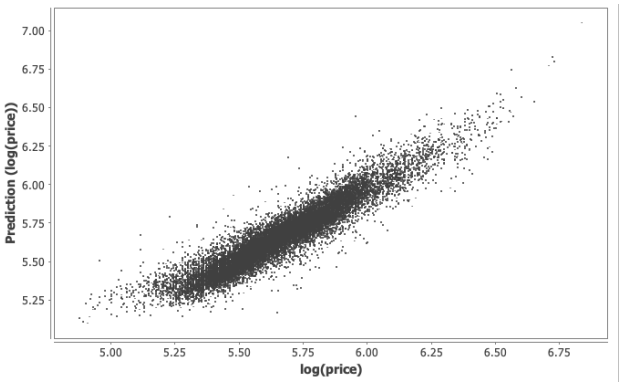

Figura 11: *Scatterplot - Confronto tra i valori reali e i valori previsti nel modello LR con trasformazione della variabile price in* $log(price)$

Il modello *Linear Regression* costruito con la variabile risposta trasformata ha portato a un valore di R^2 nel training set pari a $0,881$ e $0,886$ nel test set, a differenza del modello *Linear Regression* di partenza in cui il valore era pari a $0,837$ (e $0,835$ nel test set). Una successiva implementazione tramite processo di *Cross-Validation*, simile a quello utilizzato in precedenza, ha mostrato una minima variabilità delle metriche relative alla bontà del modello, confermandone i risultati.
Infine, dopo aver creato un opportuno ciclo per effettuare una *Backward Feature Selection* e utilizzati i criteri esposti nella sezione 5.1, è stato selezionato un modello con 13 variabili esplicative. Una volta riaddestrato il modello sull'unione training/validation set e applicandolo al test set, si è ottenuto un valore di R^2 rispettivamente pari a $0,723$ e $0,741$.

7 Conclusioni

Ripartendo dall'obiettivo iniziale e a seguito dei risultati ottenuti, si possono effettuare le seguenti deduzioni. Il modello *Linear Regression*, pur essendo un algoritmo relativamente semplice, si è rivelato essere il migliore sotto la maggior parte degli aspetti considerati ed analizzati. Tuttavia alcune ipotesi di base riguardanti i residui non sono state verificate, come ad esempio la normalità e l'omoschedasticità che non rispettano le assunzioni di partenza del modello. Di conseguenza questo modello non si può considerare pienamente accettabile sotto molteplici punti di vista. La considerevole numerosità di outliers nelle osservazioni corrispondenti alle case con prezzo di vendita relativamente elevato è il fattore principale dell'inadeguatezza del modello. Al fine di affievolire la rumorosità di queste osservazioni, si è valutata la trasformazione logaritmica della variabile risposta *price*. Questo ha comportato un ulteriore miglioramento per quanto concerne gli indici di bontà del modello, tuttavia non sono state completamente risolte le questioni sollevate precedentemente. Per quanto riguarda il metodo di *cross validation*, i risultati ottenuti si allineano conformemente ai risultati conseguiti ed analizzati in precedenza tramite approccio *holdout*.

Un eventuale studio futuro più approfondito pertinente a quanto già introdotto potrebbe innanzitutto considerare modelli di regressione più complessi. Una prima proposta di miglioramento potrebbe essere quella di sfruttare algoritmi più avanzati, come ad esempio le reti neurali, o l'applicazione di metodi statistici maggiormente evoluti come LASSO, che esegue una selezione e regolarizzazione delle variabili con lo scopo di migliorare sia l'accuratezza predittiva che l'interpretabilità del modello risultante. Un secondo suggerimento per il miglioramento dello studio potrebbe essere quello di considerare i dati di vendita per l'intero stato e non solo per la contea di King. È comune pensare che le variabili relative al posizionamento geografico (e.g. *zipcode*) siano rilevanti nella determinazione del prezzo di vendita di una casa. Questo fenomeno potrebbe essere tenuto in considerazione nel momento in cui si valutasse uno studio ad hoc o nel caso in cui si avesse accesso a più dati. Inoltre, non avendo a disposizione un'ampiezza temporale adeguata per valutare un eventuale trend del mercato immobiliare non si è potuta studiare la serie storica dei prezzi delle case, spesso rilevante per esperti di dominio. Queste strade potrebbero migliorare e semplificare la creazione di un modello predittivo.

Software Utilizzati

Per lo studio sono stati utilizzati i seguenti software: Tableau Desktop (v. 2020.3), KNIME (v. 4.3.0) ed R (v. 4.0.3). Per la stesura del report è stato utilizzato LaTeX.

Bibliography

[1] Leo Breiman et al. *Classification and regression trees*. CRC press, 1984.

[2] Harlfoxem. *Dataset Kaggle*. URL: `https://www.kaggle.com/harlfoxem/housesalesprediction`.

[3] Trevor Hastie, Robert Tibshirani e Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[4] *King County, Washington*. URL: https://en.wikipedia.org/wiki/King_County,_Washington.

[5] *National Committee on Vital and Health Statistics*. URL: https://ncvhs.hhs.gov/.

[6] *USNews healthiest-communities*. URL: https://www.usnews.com/news/healthiest-communities/washington/king-county#housing.

[7] *USNews methodology*. URL: https://www.usnews.com/news/healthiest-communities/articles/methodology.