Predizione del prezzo delle auto usate nel mercato indiano

Team_24: Lorenzo Bellotti, Giovanni Tosi, Daniele Quattrocchi, Mohamed Helmi Ben Hassine

Sommario

Il dataset selezionato è stato scaricato da Kaggle ed è costituito dalle caratteristiche di un campione di veicoli usati appartenenti al mercato dell'auto indiano.

Lo scopo di questo progetto è stato quello di individuare il modello migliore per prevedere il prezzo delle auto usate in funzione delle loro caratteristiche. Inoltre, si è cercato di individuare gli attributi con maggiore influenza nel processo di previsione.

Indice

Introduzione

Il dataset UsedCars è formato dalle caratteristiche tecniche di 6019 automobili provenienti dal mercato indiano delle auto usate dell'anno 2019. Il mercato di riferimento si caratterizza per una crescita costante dovuta alle condizioni economiche favorevoli che hanno spinto il numero delle vendite ad un costante aumento negli ultimi anni.

Il dataset iniziale si compone di 6019 righe, ciascuna delle quali rappresenta una singola auto usata, e 13 colonne, che descrivono le caratteristiche tecniche di ogni osservazione.

In particolare:

- Name (Qualitativa - Nominale): marca e modello dell'auto osservata;
- Location (Qualitativa – Nominale): ubicazione di vendita dell'auto;
- Year (Qualitativa – Ordinale): anno di immatricolazione;
- Kilometers_Driven (Quantitativa – Continua): chilometraggio percorso;
- Fuel_Type (Qualitativa – Nominale): tipo di carburante;
- Trasmission (Qualitativa – Nominale): tipo di trasmissione;
- Owner_Type (Qualitativa – Nominale): numero del proprietario;
- Mileage (Quantitativa – Continua): consumo medio;
- Engine (Quantitativa – Continua): cilindrata;
- Power (Quantitativa – Continua): potenza espressa in bhp;
- Seats (Quantitativa – Discreta): numero di posti a sedere;
- New_Price (Quantitativa – Continua): prezzo (in INR Lakhs[1]) di listino del medesimo modello nuovo;
- Price (Quantitativa- Continua): prezzo di vendita (in INR Lakhs[1]);

[1] Il Lakh è un'unità del sistema di numerazione indiano pari a 10^5 rupie.

Le fasi dello studio comprendono i seguenti step:

1. **Esplorazione Dati:** analisi del dataset di partenza.
2. **Preprocessing:** gestione delle criticità del dataset, come ad esempio valori nulli, eventuali conversioni delle unità di misura e creazione di colonne funzionali alla previsione del prezzo.
3. **Modelli:** selezione dei modelli predittivi da sottoporre a successiva valutazione.
4. **Validazione:** valutazione delle performance dei modelli precedentemente selezionati al fine di individuare il modello maggiormente efficace per la predizione del prezzo.

1 - Esplorazione Dati

Analizzando il dataset iniziale sono emerse le seguenti problematiche:

- **Presenza di valori nulli nella colonna New_Price:** si è riscontrata la mancanza di dati per l'86% delle osservazioni.
- **Ambiguità dell'attributo Name:** è presente la marca e il modello dell'automobile all'interno della stessa colonna.
- **Tipo di dato non coerente:** si è osservata la presenza di attributi numerici trattati come stringhe (es. Mileage, Engine e Power).

Analizzando la distribuzione dei prezzi, si evince come questi siano compresi tra un valore minimo di 0,44 e un valore massimo di 160 INR Lakhs, con il 75% delle osservazioni considerate che arriva ad un prezzo massimo pari a 9,95 INR Lakhs.

Inoltre, altri due aspetti particolarmente significativi sono il valore della mediana, pari a 5,64 INR Lakhs, e la numerosa presenza di outlier all'interno della visualizzazione della distribuzione. Quest'ultimo fattore influenza in maniera negativa la visualizzazione del box plot.

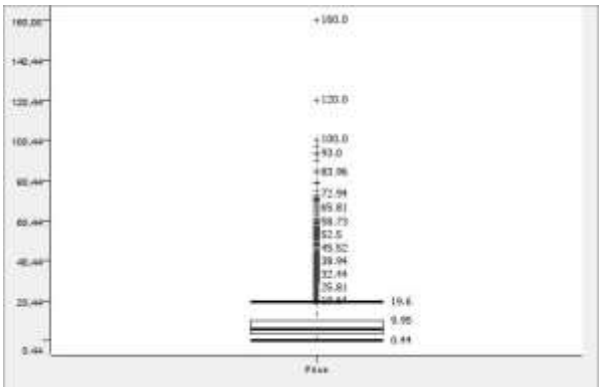

Figura 1: Distribuzione dei valori Price

2 - Preprocessing

A causa del numero di colonne limitate, non è stato necessario eliminare attributi superflui. L'unico intervento per quanto riguarda la selezione di attributi è stato quello di eliminare la colonna New_Price a causa dell'elevata percentuale di valori nulli (86%). Proprio questo aspetto rendeva impossibile l'applicazione di altre tecniche che consentissero la sostituzione dei valori nulli con altri valori.

2.1 Feature transformation

L'obiettivo di questa sezione è stato quello di rendere i dati di partenza utilizzabili per lo scopo di questo studio. Il primo attributo sottoposto a trasformazione è stato Name, si è provveduto a suddividere la stringa contenente marca e modello del veicolo in due attributi distinti. In questo modo si è creata la colonna Brand popolata col nome della marca del produttore dell'automobile (30 modalità) eliminando contestualmente la colonna contenente il modello, la quale avrebbe introdotto un numero eccessivo di variabili ai fini della regressione. Infine, è stata uniformata la nomenclatura dell'attributo brand. (metanodo *Preprocessing Name*)

Inoltre, a causa della difficoltà di conversione e dell'esigua numerosità di valori, si è deciso di filtrare l'attributo Fuel_Type, eliminando i record contenenti CNG e LPG che rappresentano rispettivamente l'alimentazione a metano e a GPL. In questo modo si è conservata l'integrità del dataset. (metanodo *Preprocessing Fuel_Type*)

Per gli attributi Mileage, Engine e Power è stata eseguita una conversione del data type da stringa a intero, escludendo le rispettive unità di misura. (metanodi *Preprocessing Mileage, Preprocessing Engine e Preprocessing Power*)

Al fine di poter utilizzare l'attributo Year come variabile quantitativa, si è provveduto alla trasformazione dell'anno di immatricolazione in vehicle_age che rappresenta l'età in anni del veicolo. (metanodo *Preprocessing Year*)

Infine, si è interevnuto sull'attributo Seats, raggruppando il numero di posti dell'auto in 4 bin (2 posti, 4 posti, 5 posti, maggiore di 5 posti), riducendo in questo modo la variabilità della colonna e rendendola utilizzabile ai fini dell'applicazione dei modelli predittivi, che verranno esposti di seguito. (Contenuto nel metanodo *Preprocessing Seats*)

Conclusa la fase di trasformazione si è proceduto all'eliminazione dei missing values, portando il dataset da 6019 a 5780 righe, non compromettendo tuttavia l'integrità del dataset iniziale. Infatti, le righe contenenti i valori mancanti ammontano a solo il 4% del totale.

Al fine di poter applicare i modelli di regressione, è stato necessario procedere con la binarizzazione di ciascun attributo qualitativo, portando il numero di colonne da 13 a 62. (Contenuto nel metanodo *Preprocessing* nel nodo *One to Many*).

Una volta conclusa questa fase, è stato eseguito il calcolo delle eventuali correlazioni tra le variabili numeriche presenti nel data set processato.

Figura 2: Correlogramma

Si noti come le variabili Engine e Power siano altamente correlate tra di loro (0.87). Aspetto curioso è la poca correlazione tra le variabili Price e Kilometers_Driven. Quest'ultima, per essere significativa deve essere analizzata in funzione del marchio del veicolo.

3 – Modelli

Dato che il class attribute è una variabile quantitativa continua, si è deciso di utilizzare dei modelli di regressione. In particolare, si sono selezionati i seguenti modelli:

- Simple Linear Regression (SLR): questa tipologia di regressione si basa sulla predizione del class attribute, che utilizza la variabile con lo squared error minore. Tale modello è il più semplice per una regressione lineare e permette di comprendere la variabile esplicativa che spiega al meglio il class attribute.
- Linear Regression (LR): modello di regressione basato sull'utilizzo di più variabili esplicative. Tale modello rappresenta l'applicazione basilare del modello di regressione lineare.
- R Linear Regression (R): modello di regressione che sfrutta le potenzialità di R utilizzando come input tutte le variabili a disposizione. Questo modello è stato selezionato con lo scopo di effettuare un confronto con i risultati ottenuti dall'applicazione del modello precedente.
- Multi Layer Perceptron (MLP): algoritmo di regressione basato su rete neurale in grado di sviluppare una funzione di regressione non lineare. Quest'ultimo aspetto lo differenzia dai precedenti modelli fornendo una valida alternativa per la predizione del prezzo.
- Polynomial Linear Regression (POLY): modello di regressione che sfrutta una funzione di regressione di secondo grado. Come il modello precedente, esso si basa su una struttura non lineare, con la possibilità di impostare a priori il grado della funzione.

4 – Validazione

4.1 Holdout

Successivamente alla fase di preprocessing e di selezione dei modelli, si procede alla validazione degli stessi utilizzando l'approccio di holdout.
In particolare, è stata eseguita una partizione del dataset in training set (67% delle righe) e test set (33% restante). La tecnica di campionamento utilizzata è stata il campionamento casuale semplice, con random seed pari a 1234. Una volta ottenuto il risultato della partizione, si è addestrato il learner sulle righe del

training set. Successivamente, l'algoritmo sviluppato è stato testato tramite l'Inducer che riceve in input il test set, concludendo in tal modo lo sviluppo del modello.

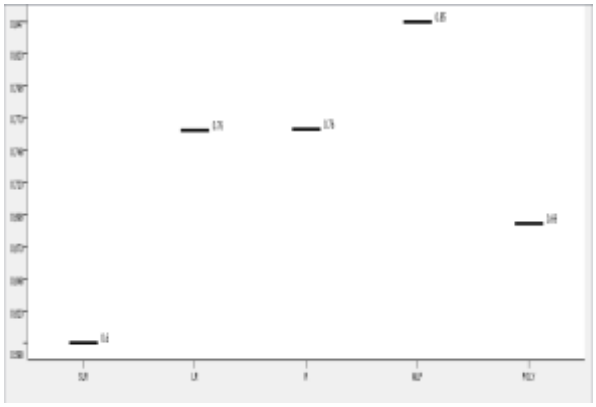

Figura 3: Confronto R^2

Il modello con R^2 migliore è dato da MLP con un valore di 0.85. Questo valore è la prova di un ottimo adattamento ai dati. Il modello SLR fornisce il risultato peggiore con un R^2 pari a 0.6. Per questa ragione, si deduce che un solo attributo (Power) abbia una capacità esplicativa del 60% nella previsione del prezzo.

Il Mean Squared Error (MSE) indica la discrepanza quadratica media tra i valori dei dati osservati ed i valori dei dati stimati.

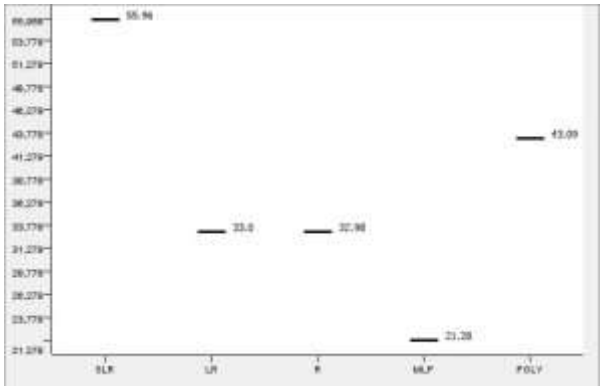

Figura 4: confronto MSE

Come atteso, il modello MLP restituisce il valore minimo di MSE, mentre il modello SLR assume il valore massimo. La differenza tra i due valori è spiegata dalla differenza di R^2 dei modelli.

Poiché il modello MLP è risultato essere quello più performante, si è deciso di approfondire ulteriormente l'analisi dei valori predetti rispetto a quelli reali mediante uno scatter plot.

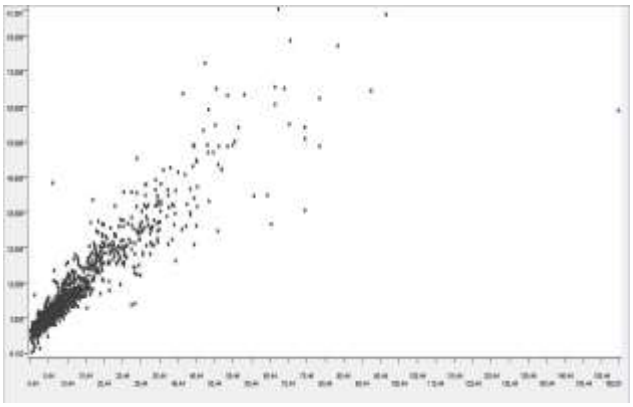

Figura 5: Scatter Plot valori predetti VS valori reali, MLP

La densità delle osservazioni lungo la retta di regressione indica un buon fitting del modello in funzione dei dati. Come da aspettativa, si nota una dispersione delle osservazioni dovute alla presenza di numerosi outlier, che andrebbero opportunamente trattati.

Inoltre, si è deciso di visualizzare l'andamento dei residui mediante un istogramma.

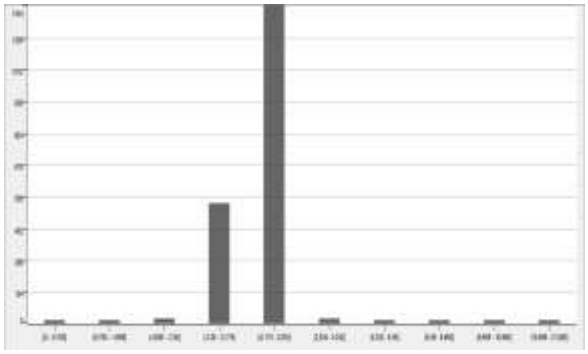

Figura 6: Istogramma residui, MLP

Da tale grafico, si evince un andamento non normale. In particolare, il 23% delle osservazioni ha residui compresi tra -0,174 e 2,034. Tuttavia, non si riscontrano anomalie, dovute alla non normalità, per quanto riguarda la distribuzione dei residui.

4.2 Cross Validation

Al fine di avere una panoramica più ampia per quanto riguarda la validazione dei modelli, si è scelto di eseguire la cross validation.

Nell'applicazione di questo metodo si è scelto di utilizzare un K=10, consentendo la suddivisione del dataset iniziale in 10 partizioni. Tale metodo risulta particolarmente dispendioso dal punto di vista computazionale, in particolare se applicato al modello MLP.

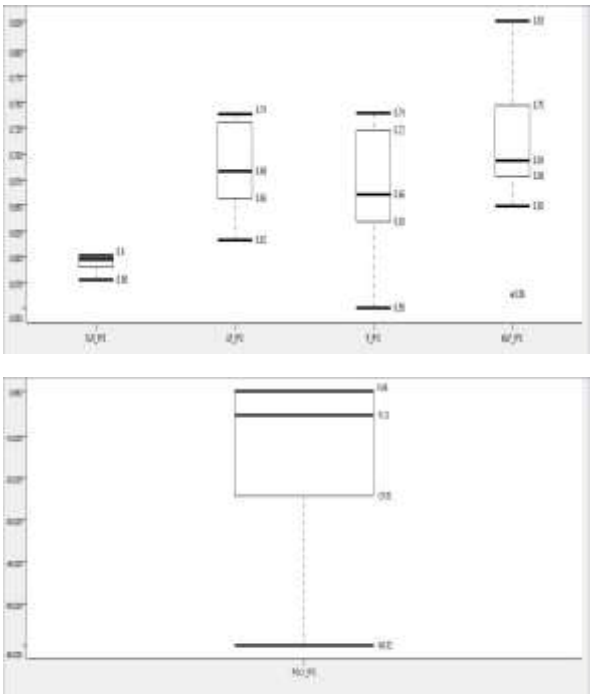

Figura 7: Confronto R^2

Come riscontrato nel holdout, dai box plot emerge che il modello MLP è quello più performante, nonostante sia caratterizzato da un range tra il valore massimo e il minimo pari a 0,27 al contrario del modello SLR che mostra una variabilità ridotta. A differenza dei risultati ottenuti in precedenza, si nota un netto peggioramento delle performance del modello polinomiale. Quest'ultimo si caratterizza per la presenza di valori negativi di R^2 che indicano un pessimo adattamento del modello ai dati.

Analogamente al metodo di validazione di holdout, si illustrano i grafici contenenti i valori del MSE.

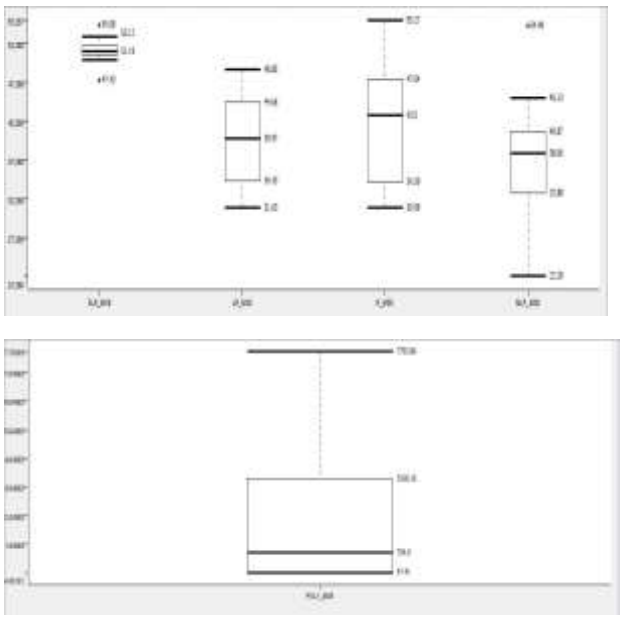

Figura 8: Confronto MSE

Come atteso, il modello MLP risulta avere i valori minori di MSE. Si può constatare che il modello polinomiale, valutato con la tecnica di cross validation, risulti inutilizzabile per lo scopo dello studio poiché i risultati forniti sono pessimi.

4.3 Feature Selection

Lo scopo della feature selection è quello di ottimizzare le performance dei modelli ottenendo il miglior compromesso tra numero di variabili coinvolte e miglior R^2.

Si è scelto di applicare tale procedura al modello di regressione lineare e al modello di regressione polinomiale in quanto, grazie alla loro struttura, supportano questo tipo di tecnica.

Per quanto riguarda il primo modello, con 51 features, si ottiene un R^2 pari a 0,775. Grazie al processo di feature selection, questo risultato può essere ottimizzato riducendo drasticamente il numero di colonne, utilizzandone 19 rispetto alle 51 iniziali, perdendo solamente circa il 2% del valore di R^2 ottenendo così un risultato pari a 0,75.

L'applicazione del metodo al modello di regressione polinomiale è risultato inefficace. Infatti, la variazione del numero di features non incide sull'andamento del suo valore di R^2.

Conclusioni

Alla luce dei risultati ottenuti durante il processo di valutazione dei modelli, è stato riscontrato che il modello MLP risulti essere il migliore per quanto riguarda le performance, con un R^2 pari a 0,85 e un MSE pari a 21,28. Il compromesso da accettare per ottenere un risultato così soddisfacente consiste in un costo computazionale elevato.

Una valida alternativa a questo modello è rappresentata dal modello di regressione lineare che risulta essere meno efficace nei risultati ma decisamente meno computational demanding.

Per quanto riguarda quest'ultimo modello è particolarmente interessante la possibilità di applicare la feature selection evidenziando gli attributi che più influiscono nel processo di predizione del prezzo dei veicoli.

Nella fattispecie tale modello è influenzato maggiormente dai seguenti attributi:

- Power
- Car_Age

Questi due attributi incidono sull'andamento del prezzo per un valore di R^2 pari a 0,66.
Un aspetto particolarmente curioso è il fatto che i due marchi automobilistici che influenzano maggiormente la retta di regressione siano Land Rover e Honda.
In conclusione, il risultato di questo studio si può ritenere soddisfacente, nonostante sia ulteriormente migliorabile mediante l'opportuna gestione dei valori influenti e degli outliers.

Fonti

- https://www.kaggle.com/avikasliwal/used-cars-price-prediction?select=train-data.csv
- https://scikit-learn.org/stable/modules/neural_networks_supervised.html
- https://www.mordorintelligence.com/industry-reports/india-used-car-market
- https://hub.knime.com/lisovyi/spaces/Public/latest/Examples/Cross%20Validation%20example~ph4V4jcg5pNTRZfb
- https://nodepit.com/node/org.knime.base.node.meta.feature.selection.FeatureSelectionLoopStartNodeFactory2