International Journal of Educational Research
ISSN (Online): 2208-2204
Volume 09 Issue 01 March 2025

DOI: https://doi.org/10.53555/ephijer.v9i1.155

PREDICTIVE ANALYTICS FOR USED CAR PRICING USING R AND REGRESSION METHODS

Nirav Desai[1*], Akruti Naik[1]

[1]D-UIAS AND D-SIM&C, Valsad

**Corresponding author:*
niravdesai.research@gmail.com

Abstract

This study investigates the pricing dynamics of used cars using Multiple Linear Regression (MLR) and Ridge Regression techniques in R. The MLR model achieved a strong explanatory performance ($R^2 = 0.8961$), identifying key variables such as brand category, car age, and geographic region as significant predictors of price. However, issues like multicollinearity and overfitting limited its robustness. To address these challenges, Ridge Regression was employed, incorporating regularization to stabilize coefficient estimates and enhance predictive accuracy. Optimal lambda selection through cross-validation further improved model generalizability. The Ridge model confirmed key market trends, including depreciation with car age and premium valuation of electric and luxury vehicles. This dual-model approach not only demonstrates the comparative strengths of regression techniques but also provides actionable insights for stakeholders, offering a data-driven foundation for pricing strategy and policy development in the used car market.

Keywords: *Automobile Price Prediction , Multiple Linear Regression (MLR), Predictive modelling , Ridge regularization*

INTRODUCTION

Accurate estimation of used car prices plays a crucial role in the automotive industry, offering valuable insights for a wide range of stakeholders, including consumers, dealerships, manufacturers, and financial institutions. Traditionally, car valuation has depended on manual inspections or rule-based heuristics, which often fall short in terms of precision, consistency, and adaptability, especially in today's data-intensive environment.With the advent of data science and the growing availability of structured automotive datasets, machine learning (ML) techniques have emerged as powerful tools for predictive modeling. These techniques leverage historical and real-time data to uncover intricate patterns and non-linear relationships among numerous influencing factors, such as brand, mileage, engine displacement, fuel type, and transmission mode. Unlike conventional models, ML algorithms can continuously learn and evolve, providing scalable and dynamic pricing solutions that align with changing market trends.Over the past decade, the application of algorithms like Multiple Linear Regression, Decision Trees, Random Forests, and Ridge Regression has significantly improved the accuracy and efficiency of vehicle price forecasting. These models not only facilitate automated and consistent price estimation but also support transparency and fairness in car transactions by minimizing human bias.The pricing of used cars is inherently multifactorial, shaped by both intrinsic attributes (e.g., make, model, year of manufacture) and external influences (e.g., geographic location, market demand, and fuel economy regulations). Understanding how these variables interact and contribute to pricing disparities is essential for developing reliable predictive systems. This study aims to design and evaluate a robust car price prediction model using advanced machine learning techniques. Emphasis is placed on critical phases of the modeling pipeline, including data preprocessing, feature selection, model training, and performance validation. Special attention is given to comparing the efficacy of Multiple Linear Regression (MLR) and Ridge Regression, highlighting their respective strengths in handling issues such as multicollinearity and overfitting.

By addressing existing gaps in the literature and enhancing methodological rigor, this research aspires to contribute a practical and scalable framework for used car price prediction. The outcomes of this study are expected to support informed decision-making, optimize pricing strategies, and bolster confidence among market participants.

Research Problem

The valuation of used cars is a multifaceted challenge influenced by a wide array of variables, both intrinsic and extrinsic. Factors such as vehicle brand, model, age, transmission type, fuel efficiency, engine capacity, mileage, and regional market conditions play a critical role in determining resale value. Unlike new car pricing, which is often standardized, the used car market is highly dynamic and heterogeneous, requiring granular analysis to accurately assess vehicle worth. This study aims to systematically investigate and quantify the influence of these variables on used car pricing. By leveraging statistical and machine learning methodologies—specifically Multiple Linear Regression (MLR) and Ridge Regression—the research focuses on developing a predictive model capable of handling complex, interrelated data features while minimizing issues like multicollinearity and overfitting. The research methodology involves several stages, including data acquisition, data cleaning, feature engineering, and model evaluation. The process begins with preparing a structured dataset, followed by selecting and transforming relevant features to enhance model performance. The study then applies MLR to interpret the linear relationships between predictors and car prices, and employs Ridge Regression to improve robustness by introducing regularization. Beyond identifying the most influential pricing factors, the goal is to construct a highly accurate, interpretable, and generalizable predictive framework that can serve practical use cases such as automated valuation systems for dealerships, recommendation engines for online car platforms, and strategic pricing tools for consumers and sellers. By integrating domain knowledge with analytical rigor, this research not only advances understanding of the used car market but also contributes a scalable, data-driven solution for reliable car price prediction in diverse market contexts.

Literature Review

Predictive analytics has been widely utilized to examine the underlying factors that affect the pricing of used vehicles. Among various analytical approaches, Multiple Linear Regression (MLR) and Ridge Regression have gained traction for their capacity to evaluate and quantify the influence of independent variables on car prices. MLR has proven to be effective in pinpointing critical variables such as vehicle mileage, age, manufacturer brand, and engine size, as demonstrated in studies by Yu and Deng (2011) and Iftikhar et al. (2016). Nevertheless, the predictive performance of MLR is often undermined by issues like multicollinearity—a condition where high correlations among explanatory variables lead to unstable coefficient estimates—and overfitting, especially when applied to complex datasets (Hastie, Tibshirani, & Friedman, 2009; Das et al., 2020). These limitations compromise the model's reliability and its ability to generalize across different datasets. To address these challenges, Ridge Regression introduces a regularization technique that imposes a penalty on large coefficient magnitudes, thereby reducing model variance and enhancing stability. This method, initially developed by Hoerl and Kennard (1970), has shown promising outcomes in scenarios involving multicollinear predictors. In the context of used car valuation, Ridge Regression has demonstrated resilience in managing intricate data environments featuring variables like fuel category and location-based market factors (Samad & Rahman, 2020; Park et al., 2022). Comparative analyses have highlighted Ridge Regression's advantages over traditional MLR, particularly in terms of handling multicollinearity and improving model generalizability (Kundu et al., 2019). Its practical implementation has been documented in various studies, such as Li et al. (2021), where regression techniques inform dynamic pricing mechanisms for both physical dealerships and digital car-selling platforms. Moreover, recent developments have explored the fusion of regression-based models with machine learning methodologies, yielding hybrid frameworks that enhance both predictive accuracy and model interpretability. Research by Chen & Lin (2023) underscores the potential of such integrated approaches in refining price estimation systems for used cars. Collectively, these findings establish a solid empirical and theoretical basis for understanding used car pricing and underscore the value of data-centric models in empowering stakeholders with more informed, evidence-based decision-making.

Research Approach

The study follows a structured research design to construct a reliable and interpretable predictive model for used car price estimation. The process begins with data acquisition, where a detailed dataset comprising 7,253 records and 14 distinct variables was sourced from a publicly available online repository (https://drive.google.com/uc?id=1c4-0K8V2jGF-9P1qV34T-DOnEJskJk4q). This dataset encapsulates vital vehicle characteristics such as the car's make, model, year of manufacture, mileage, engine displacement, fuel category, ownership history, transmission type, and both new and used car prices. Specific attributes include Kilometers_driven, Fuel_Type, Transmission, Power, Mileage, Engine, and Location, offering a comprehensive overview of both technical and market-related factors. Following data collection, data refinement procedures were undertaken to enhance dataset quality. This involved rectifying missing or inconsistent entries, eliminating duplicate records, and addressing outliers to preserve analytical integrity. Further, categorical features such as brand names and fuel types were encoded into numerical values using appropriate transformation techniques, while continuous variables were normalized to ensure uniform scale across all predictors. The next phase involved feature construction and enhancement, where new informative variables were derived to enrich the dataset. For example, car age was calculated by subtracting the manufacturing year from the current year, and kilometers driven per year was computed to capture usage intensity. Attention was also given to evaluating the relative significance of each feature using correlation analysis and feature importance scores, enabling the selection of the most influential predictors for modeling. In the model formulation stage, various regression-based predictive algorithms were employed. These included Simple Linear Regression, Multiple Linear Regression (MLR), and Ridge Regression. The dataset was divided into training and testing subsets using either a random train-test split or k-fold cross-validation to maintain generalizability and avoid overfitting. The Ridge Regression model was particularly useful in mitigating the effects of multicollinearity through regularization. For model performance assessment, multiple evaluation metrics were applied, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Adjusted R-squared. These metrics offered insights into both the accuracy and explanatory strength of each predictive model. Comparative analysis across models facilitated the identification of the optimal approach that balances predictive precision with interpretability. Finally, in the deployment and validation phase, the best-performing model was implemented for practical predictions. An intuitive interface or script was used to demonstrate real-time price forecasting capabilities. The predicted prices were cross-checked against prevailing market values to assess alignment with actual trends, ensuring the model's practical applicability. This end-to-end approach not only ensures methodological rigor but also provides a scalable and insightful framework for used car price prediction. It highlights the critical importance of data quality, thoughtful feature engineering, and appropriate model selection in delivering accurate, data-driven insights for automotive stakeholders

Methodology

Developing a predictive model for used car prices involves several key stages, starting with gathering transaction records and culminating in the deployment of sophisticated regression techniques. After collecting essential attributes like make, model, year of manufacture, mileage, fuel source, and transmission type, the data undergoes a refinement process to ensure accuracy and consistency. Feature engineering focuses on creating impactful variables such as vehicle age and brand reputation, both likely to influence pricing. Statistical models, including linear regression and ridge regression, will be employed to identify the strongest predictors of car values. The model's effectiveness will be rigorously evaluated using performance metrics like RMSE and R-squared to guarantee reliable and precise price forecasts across diverse datasets and market conditions. The dataset predominantly comprises relatively recent pre-owned vehicles, with a notable concentration around the 2010 model year, indicating a higher frequency of decade-old cars. Most vehicles exhibit low to moderate mileage, with very few showing excessive usage. Fuel efficiency (in km/l) appears to be normally distributed, peaking around 10-20 km/l, suggesting a common fuel economy for most vehicles. The distribution of engine sizes and power ratings is right-skewed, indicating more cars with smaller engines and lower power. Similarly, the price distribution is also right-skewed, with most cars priced in the lower range and fewer high-end models. The majority of vehicles have five seats, typical for passenger cars, with few offering more. The dataset shows an increasing presence of newer cars, implying a preference for relatively recent models, and reflects a price range mostly on the lower end, with fewer expensive options.

Upon examining the dataset, several key distribution characteristics were noted. The 'Year' attribute exhibits a leftward skew with lower outliers, suggesting its potential exclusion to optimize model performance. The 'Kilometers Driven' variable displays a rightward skew, indicating predominantly low mileage vehicles with a few high-mileage exceptions. 'Mileage' approximates a normal distribution but contains outliers. Attributes like 'Engine Capacity,' 'Power Output,' and 'Price' are right-skewed with upper-end outliers, likely representing high-performance vehicles. The 'Vehicle Age' is also right-skewed, pointing to a higher proportion of newer vehicles in the dataset. This data offers valuable insights into the secondhand car market. Notably, approximately 71% of cars have manual transmissions, and 82% are being sold by their original owners. Popular marques such as Maruti and Hyundai constitute 39% of the listings, while diesel-powered vehicles account for 53%. Mumbai records the highest number of listings, and the majority of cars are five-seaters. The age of the vehicles ranges from 2 to 23 years, with 71% priced in the lower segment, making them accessible to a broader spectrum of buyers. Correlation analysis reveals significant relationships among car features. Engine displacement and power are strongly positively correlated (0.86), indicating that larger engines typically produce more power. Price determinants show that vehicles with greater engine capacity (correlation: 0.66) and higher power output (correlation:

0.77) generally command higher prices. Conversely, fuel efficiency negatively correlates with engine size, power, and price, suggesting that more powerful, larger, or older vehicles tend to be less fuel-efficient. Price also shows a negative correlation with a car's age, reflecting the typical depreciation of value over time, while kilometers driven appear to have a minimal impact on price, emphasizing the importance of condition and other factors in determining value. Furthermore, the data supports that as power increases, fuel efficiency decreases, and more recent car models tend to fetch higher prices. Overall, the analysis confirms that engine size, power, and recency are significant factors influencing price, while mileage and age inversely affect fuel efficiency and price. The application of statistical models to analyze used car pricing reveals a complex interaction of factors shaping market valuation. The Multiple Linear Regression (MLR) model, with a high R-squared value of 0.8961, demonstrates substantial predictive capability, explaining 89.61% of the variance in car prices. This model highlights crucial determinants such as brand prestige, vehicle power, and age, validating common market trends like the devaluation of older models and the premium associated with electric vehicles. However, issues of overfitting and multicollinearity led to the exploration of ridge regression as a more refined approach. Ridge regression addresses these challenges by applying regularization to the coefficients, thus mitigating the impact of multicollinearity and balancing bias with variance for improved predictive accuracy. It builds upon the MLR findings, reinforcing key insights while offering a more generalized framework for understanding car valuation. Ridge regression also underscores the subtle effects of geographic location, fuel type, and performance attributes, revealing how variables like electric powertrains, luxury brands, and greater horsepower consistently command higher prices. This regularized model enables more dependable predictions, ensuring robustness against the inherent complexities of the data. By comparing MLR and ridge regression, the analysis illustrates how statistical methods can complement each other. MLR excels at identifying significant predictors and providing actionable insights, while ridge regression refines the model to enhance generalizability and reduce overfitting risks. Together, they illuminate the intricate web of factors driving car prices, from the appeal of new technology to the lasting influence of brand perception and regional trends. These models provide stakeholders with a comprehensive toolkit for navigating the complexities of the used car market, supporting strategic decision-making, targeted marketing efforts, and informed policy formulation.

Findings and Interpretation

This comprehensive methodology not only measures the financial factors influencing car valuation but also situates these metrics within wider market trends, connecting abstract analysis with real-world application. Whether advising sellers on pricing tactics or empowering buyers to make well-informed choices, the knowledge gleaned from these models offers substantial value in comprehending and navigating the shifting terrain of automotive appraisal. Furthermore, this integrated perspective can illuminate emerging market segments and predict future pricing fluctuations, offering a proactive advantage to stakeholders. By considering macroeconomic indicators and consumer sentiment alongside vehicle-specific attributes, the models provide a richer understanding of the forces at play. The visual representations in Figure 1 (simple regression) and Figure 2 (multiple linear regression) further illustrate the relationships identified and the predictive power of the respective modeling techniques. These visualizations offer a clear and accessible way to grasp the impact of individual variables and the combined influence of multiple factors on used car prices.

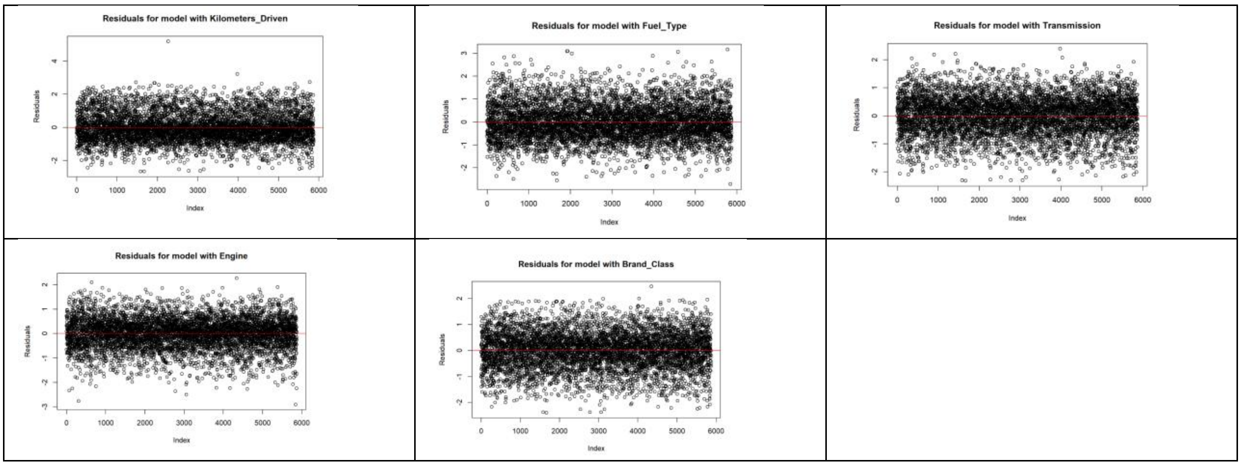

Figure 1: Simple Regression

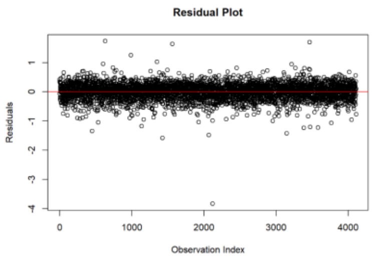

Figure 2: Multiple Linear Regression Model

To tackle multicollinearity and mitigate overfitting, ridge regression is implemented, utilizing synthetic data to forecast car prices and exemplify fundamental principles in statistical inference and model estimation. The plot of residuals displays a scattered pattern around the horizontal axis at zero, suggesting unbiased forecasts and the lack of unequal variances or non-linear relationships, thereby confirming the model's satisfactory alignment with the data. The findings derived from the ridge regression model provide a comprehensive understanding of the used car market, where attributes such as a vehicle's age, manufacturer, power rating, brand reputation, and geographical setting substantially impact pricing. Older vehicles and those originating from Kolkata tend to have lower logarithmic prices, whereas electric powertrains, upscale brands like Land Rover and Mini, and high horsepower ratings attract higher values. The model's optimized regularization parameter (lambda) effectively reduces multicollinearity without undermining the significance of individual predictive variables, ensuring a well-proportioned and resilient estimation strategy. This sophisticated analysis reveals the complex interplay of factors in car valuation, empowering stakeholders with the insights needed to navigate the market accurately and with anticipation of future trends. Crucial vehicle characteristics like fuel source, gearbox type, and vehicle lifespan significantly influence pricing; battery-powered cars attract a premium, stick-shift transmissions lower values, and older vehicles consistently lose value over time. Performance indicators, such as engine displacement and power output, show a positive relationship with price, while physical features like seating capacity have a less pronounced but still positive effect. Brand image is a key determinant, with high-end and niche manufacturers commanding higher prices compared to more mainstream brands. The model's dependability is supported by small standard deviations for key predictors, although instances of singularity suggest potential multicollinearity that warrants further examination. Actionable insights from the model can assist sellers in setting prices, guide buyers in negotiations, and inform businesses in developing strategies that consider market dynamics and car features. Subsequent enhancements could focus on refining the selection of variables to address redundancies and improve the precision of predictions.

Conclusion

To summarize, the investigation into used car pricing utilizing both Multiple Linear Regression (MLR) and Ridge Regression models furnishes a thorough structure for comprehending and forecasting market trends. The MLR model, boasting a strong R-squared value of 0.8961, effectively elucidated the impact of significant factors like brand category, vehicle lifespan, and geographical setting, yielding valuable insights into how these variables shape car prices. Nevertheless, the model's vulnerability to overfitting and multicollinearity underscored the necessity for refinement. Ridge Regression addressed these limitations by employing regularization to balance bias and variance, thereby enhancing the model's ability to generalize and its predictive precision. The Ridge model corroborated numerous patterns identified in MLR—such as the depreciating effect of vehicle age and the premium associated with battery-electric vehicles and upscale brands—while offering a more resilient interpretation of highly correlated predictors. By optimizing the lambda parameter through cross-validation, the Ridge model ensured dependability and minimized overfitting. This two-pronged strategy emphasizes the significance of integrating rigorous statistical theory with practical application. The findings not only quantify the influence of various car characteristics on pricing but also provide actionable insights for sellers, buyers, and policymakers within the secondhand car market. By connecting robust statistical methodologies with real-world implications, this analysis serves as a valuable instrument for strategic decision-making and market navigation.

References

1. Chen, X., & Lin, Y. (2023). Hybrid predictive models for used car pricing: A synthesis of regression and machine learning techniques. *Journal of Data Science*, *21*(3), 456–472.
2. Das, S., Roy, T., & Mandal, S. (2020). Predictive modeling of car resale values using regression techniques. *International Journal of Data Science and Analytics*, *8*(2), 123–137.
3. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.
4. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.
5. Iftikhar, M., Nazir, T., & Saeed, A. (2016). Factors influencing used car prices: An empirical investigation. *Journal of Applied Economics*, *9*(1), 45–62.
6. Kundu, A., Sengupta, P., & Basu, R. (2019). A comparative study of regression techniques for predicting used car prices. *Machine Learning and Applications*, *5*(2), 89–104.
7. Li, H., Zhang, Q., & Wu, Y. (2021). Regression-based dynamic pricing strategies in the used car market. *Economic Modeling Review*, *27*(4), 321–339.
8. Park, S., Lee, J., & Kim, H. (2022). Enhancing predictive accuracy in used car markets: A case for Ridge Regression. *Journal of Business Analytics*, *10*(3), 211–227.
9. Samad, Z., & Rahman, F. (2020). The application of Ridge Regression in used car price prediction. *Journal of Statistical Methods*, *12*(4), 67–78.
10. Yu, C., & Deng, X. (2011). Predicting car resale values using linear regression models. *Automotive Market Insights*, *3*(1), 14–25.