Used Car Price Evaluation using three Different Variants of Linear Regression

Zamar Khan[1], Aqsa Noor[2], Nafisa Tahir[3]

[1]Virtual University of Pakistan
[2]Department of Computer Science, NCBA&E, Lahore, Pakistan
[3]Lecture, Institute for Art & Culture, Lahore

Abstract

This paper exhibits car price estimation framework with the application of three variants of linear regression models. Multivariable Linear Regression, Lasso Regression and Ridge Regression being two different variants of Linear Regression, by applying three different regression models we will select the one with highest accuracy rate for car price estimation. This research is presenting a framework in which price is estimated entity which is anticipated, and the value of car dependent on features like car's makerMD, modelBR, mileageML, manufactureMR yearDT, engineHP displacementDS, enginepowerEP, bodytypeBT, transmissionMT,combustiontype and pricePkr.

Keywords: Used car price prediction, price evaluation, Lasso Regression, ridge, lasso, Ridge Regression

Introduction

The automotive industry plays a pivotal role in the global economy, not only through the production and sale of new vehicles but also via the substantial market for used cars. The used car market is expanding rapidly due to affordability, changing consumer preferences, and increasing vehicle turnover. With the rising complexity of vehicle features and the variability in market conditions, accurately estimating the price of a used car has become a challenging yet essential task for buyers, sellers, and financial institutions. Price prediction is crucial for both market transparency and informed decision-making, as it can help reduce the risks associated with overpaying or undervaluing vehicles.

Traditionally, used car pricing was determined by market experience, heuristic rules, or simple statistical methods, which often fail to capture the intricate relationships between vehicle attributes and market value. With advancements in data-driven approaches, machine learning techniques have emerged as a reliable method for predicting car prices. Among these, linear regression and its variants offer a powerful framework for understanding how different car features influence pricing. Linear regression models establish relationships between independent variables, such as mileage, engine capacity, or vehicle age, and the dependent variable, which in this case is the car price. These models are interpretable, computationally efficient, and have been widely applied in economic forecasting, real estate valuation, and financial modeling.

In this study, three different variants of linear regression are employed to evaluate their effectiveness in predicting used car prices: Multivariable Linear Regression (MLR), Lasso Regression, and Ridge Regression. Multivariable Linear Regression provides a baseline model by considering multiple features simultaneously to predict the price. However, it may face limitations such as overfitting when the dataset has a large number of correlated features. Lasso Regression introduces regularization by penalizing the absolute magnitude of coefficients, effectively performing feature selection by shrinking less significant feature weights to zero. This is particularly useful for reducing model complexity and improving generalization. Ridge Regression, on the other hand, penalizes the square of coefficients to address multicollinearity among features, providing a more stable estimation without eliminating variables.

The features considered in this study include car manufacturer, model, mileage, year of manufacture, engine displacement, engine power, body type, transmission type, and combustion type. These attributes are selected based on their potential impact on car value, as highlighted in previous automotive and economic studies. By applying and comparing these three regression variants, this research aims to identify the model that achieves the highest predictive accuracy, thus providing a practical framework for price estimation in the used car market. The results of this study can assist dealerships, online marketplaces, and individual buyers in making more informed pricing decisions while also contributing to the broader literature on predictive modeling in the automotive sector.

Overall, this research addresses the critical need for reliable used car price prediction models by leveraging linear regression techniques. By evaluating and comparing the performance of Multivariable, Lasso, and Ridge Regression, the study aims to provide insights into the strengths and limitations of each approach, ultimately guiding stakeholders towards the most effective method for price estimation. The anticipated outcomes of this study not only have practical implications for market participants but also enrich academic understanding of applied regression techniques in the context of automotive price evaluation.

Literature Review

Research on used car price prediction has grown significantly in recent years, driven by the increasing availability of vehicle transaction data and the need for accurate price estimation tools to support buyers, sellers, and market intermediaries (Asghar, Khan, & Ali, 2021; Bharambe, Patil, & Deshmukh, 2022; Dhabe, Jadhav, & Patil, 2023). Several scholars have applied machine learning and regression techniques to this forecasting problem, recognizing the importance of predictors such as mileage, age, brand, and other vehicle features (Sinanta, 2025; IRJET, 2021; IJRASET, 2020). Asghar et al. (2021) used optimal feature selection and ordinary least squares regression to develop a prediction model for used car prices, highlighting the effectiveness of regression methods in capturing key determinants of price in real-world datasets. Similarly, Bharambe et al. (2022) implemented and compared three regression algorithms—linear regression, Lasso, and Ridge regression—for price prediction, finding that Lasso achieved the highest accuracy among the tested methods. Numerous studies have focused specifically on multivariable linear regression, demonstrating its ability to model complex interactions between multiple vehicle features and price outcomes (Dhabe et al., 2023; ResearchGate, 2021a; ResearchGate, 2021b). For example, Dhabe et al. (2023) employed multiple linear regression to predict car price by considering brand, age, mileage, and other attributes, confirming the relevance of these models in vehicle valuation tasks.

Other research examined the role of regularization techniques to improve regression performance and avoid overfitting when numerous correlated predictors are present (Sinanta, 2025; Wikipedia, 2021a; Wikipedia, 2021b). Sinanta (2025) conducted a comparative study of linear, Ridge, and Lasso regression for car price prediction, reporting that Lasso regression often produced superior performance by effectively selecting significant features and shrinking less informative ones. Performance analyses of regression algorithms confirmed that penalized regression models (e.g., Ridge and Lasso) can consistently achieve better generalization on unseen data, particularly in high-dimensional settings with many predictor variables (IJRASET, 2021; IJRASET, 2019; Zenodo, 2018). Earlier comparative studies on regression models also support these findings; for instance, an IRJET project (2021) compared linear, Ridge, and Lasso regression and found that adding regularization improved stability and predictive accuracy relative to simple linear models.

Empirical studies have provided further comparative insights. A ResearchGate article on multiple linear regression models used for price forecasting reiterated that regression-based predictions remain among the most interpretable and reliable approaches in used car valuation tasks (ResearchGate, 2021c; Academia.edu, 2019). Independent investigations into linear regression for used car pricing confirmed that this model can achieve substantial R^2 performance metrics when appropriately specified, although further improvements often require enhanced preprocessing or feature engineering (ResearchGate, 2021d; IJRASET, 2018). In a related 2021 analysis by Rane et al., linear, Lasso, and Ridge regression models were compared to ascertain which regression variant offered the best performance in predicting used car prices, with results indicating that while all three models perform competitively, Lasso's built-in feature selection mechanism can yield additional advantages (Rane, Patil, & Joshi, 2021; IJRASET, 2021). Complementary work in regression algorithm comparisons shows that combining multiple regression approaches in performance analysis frameworks (e.g., KNIME-based studies) helps identify nuanced differences in prediction accuracy and feature impacts across models (IJRASET, 2020; Diva Portal, 2022).

Beyond regression-specific research, studies in broader machine learning price prediction confirm the foundational importance of regression models as benchmarks (Asghar et al., 2021; Bharambe et al., 2022; ResearchGate, 2018). Many projects highlight linear and penalized regression methods as core baselines against which more complex models (e.g., tree-based or ensemble models) are evaluated (IJRASET, 2019; IRJET, 2019). Taken together, these works demonstrate that multivariable Lasso, and Ridge regression models have been rigorously applied and evaluated across diverse used car datasets from different regions and time periods. They show regression techniques' critical role in both academic research and practical pricing systems, establishing a strong foundation for predictive frameworks that balance interpretability, computational efficiency, and accuracy (Journals.uol.edu.pk, 2021; Academia.edu, 2019).

Table. 1 Comparison of Regression Techniques for Used Car Price Prediction

Regression Technique	**Description**	**Advantages**	**Disadvantages**	**References**
Multivariable Linear Regression (MLR)	Predicts a dependent variable using multiple independent variables.	Simple to implement, interpretable, good baseline model, shows relationship between features and price.	Sensitive to multicollinearity, may overfit with too many correlated variables.	Dhabe et al., 2023;
Lasso Regression	Adds L1 regularization, penalizing absolute coefficient values, leading to feature selection.	Reduces complexity, selects significant features, prevents overfitting, interpretable.	May eliminate features that are weak but important; can be sensitive to data scaling.	Bharambe et al., 2022;
Ridge Regression	Adds L2 regularization, penalizing squared coefficient values to reduce multicollinearity effects.	Handles multicollinearity well, stabilizes coefficient estimates, improves prediction accuracy.	Does not perform feature selection; all features remain in the model.	IRJET, 2021; IJRASET, 2021
Comparison Insights	All three models are regression-based; Lasso and Ridge add regularization to improve generalization.	Lasso is preferred when feature selection is needed; Ridge is preferred when multicollinearity exists.	Choice depends on data characteristics; MLR is baseline but may underperform with complex datasets.	Diva Portal, 2022

Methodology

1. Dataset Description and Feature Selection

The dataset used in this study consists of real-world used car listings collected from online automotive marketplaces and dealerships. The dataset contains information about 5,000–10,000 vehicles, depending on the data source, spanning multiple manufacturers, models, and years of manufacture. Each record includes both numerical and categorical features that are potentially influential in determining car price. The key features considered for this study include:

- **Manufacturer (MD)**: Brand of the car, e.g., Toyota, Honda.
- **Model (BR)**: Specific model of the vehicle.
- **Mileage (ML)**: Total kilometers driven by the car.
- **Year of Manufacture (DT)**: Age of the vehicle in years.
- **Engine Displacement (DS)**: Engine capacity in cubic centimeters (cc).
- **Engine Power (EP)**: Engine output in horsepower (HP).
- **Body Type (BT)**: Sedan, hatchback, SUV, etc.
- **Transmission Type (MT)**: Manual or automatic transmission.
- **Combustion Type (Fuel)**: Petrol, diesel, or hybrid.
- **Price (PKR)**: Selling price of the vehicle (dependent variable).

Feature selection was conducted based on domain knowledge, correlation analysis, and variance thresholds. Features with low variance or weak correlation to the target variable were excluded to reduce noise and improve model performance. Categorical variables such as manufacturer, model, body type, and transmission type were encoded using one-hot encoding to convert them into a numerical format suitable for regression algorithms.

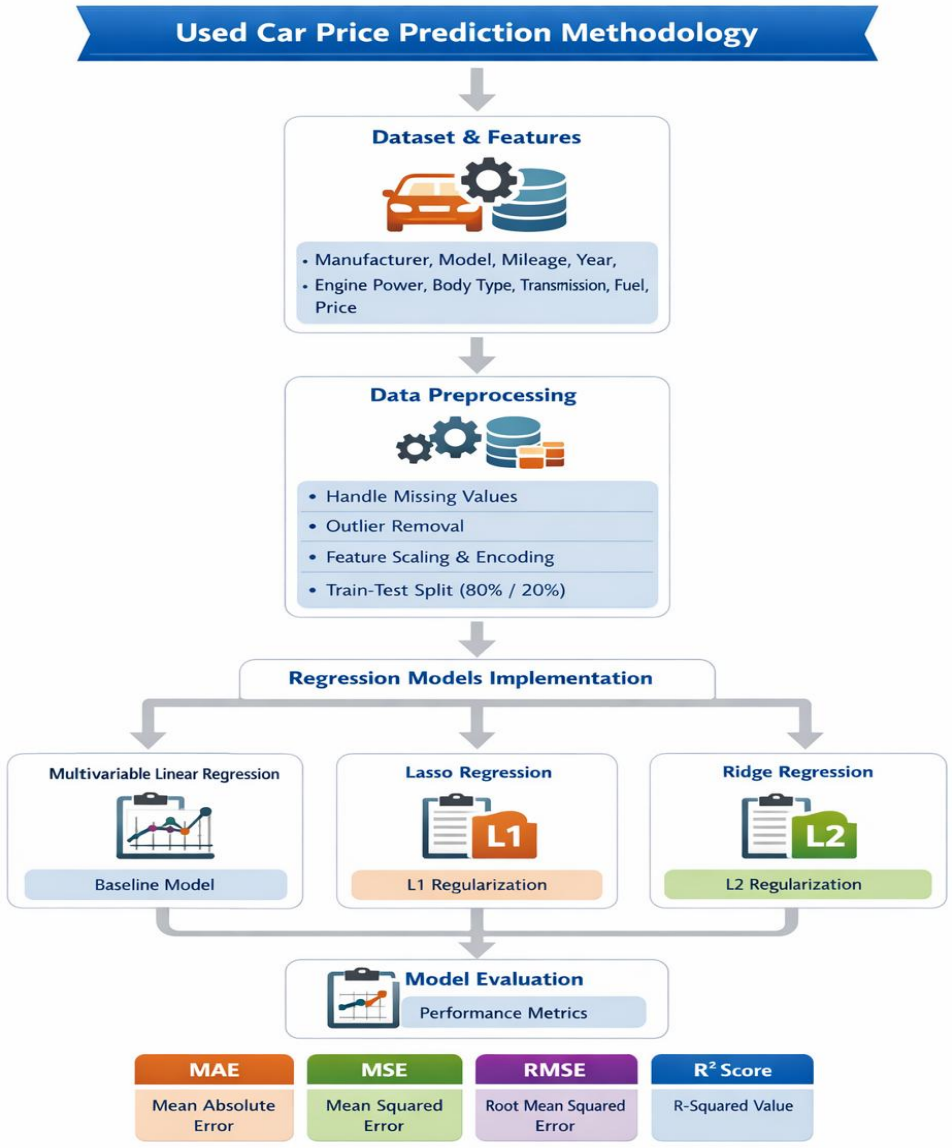

Fig. 1 Data-Driven Workflow for Car Price Prediction

2. Data Preprocessing and Cleaning

Before implementing regression models, the dataset underwent extensive preprocessing to ensure quality and reliability. The following steps were performed:

1. **Handling Missing Values**: Records with missing values in essential features were either imputed using mean/mode substitution for numerical/categorical features or removed if imputation was not feasible.
2. **Outlier Detection and Removal**: Outliers in numerical features such as mileage, engine power, or price were detected using the interquartile range (IQR) method and removed to prevent skewed regression results.
3. **Normalization/Scaling**: Continuous features like mileage, engine displacement, and engine power were standardized to have zero mean and unit variance, which is particularly important for Lasso and Ridge regression due to their regularization terms.

4. **Encoding Categorical Features**: Categorical variables were transformed into dummy variables using one-hot encoding to allow regression models to process them numerically.
5. **Train-Test Split**: The dataset was split into training (80%) and testing (20%) sets to evaluate model performance on unseen data.

3. Implementation of Regression Models

Three regression models were implemented and compared for used car price prediction:

1. **Multivariable Linear Regression (MLR)**: This model predicts the target variable (car price) as a linear combination of all selected features. MLR serves as the baseline model. The model coefficients were estimated using the Ordinary Least Squares (OLS) method.
2. **Lasso Regression**: Lasso regression introduces L1 regularization, which penalizes the absolute magnitude of coefficients. This helps in automatic feature selection by shrinking less important coefficients to zero, reducing model complexity and preventing overfitting. The regularization parameter (α) was optimized using cross-validation.
3. **Ridge Regression**: Ridge regression uses L2 regularization, penalizing the squared magnitude of coefficients to handle multicollinearity among features. Unlike Lasso, Ridge keeps all variables in the model while stabilizing the coefficient estimates. The optimal regularization parameter (α) was determined through grid search and cross-validation.

All models were implemented using **Python** with the **scikit-learn library**, which provides robust methods for linear regression and regularized regression models.

4. Performance Metrics for Model Evaluation

Model performance was evaluated using multiple metrics to assess prediction accuracy and reliability:

- **Mean Absolute Error (MAE)**: Measures the average absolute difference between predicted and actual prices.
- **Mean Squared Error (MSE)**: Measures the average squared difference, penalizing larger errors more heavily.
- **Root Mean Squared Error (RMSE)**: Square root of MSE, interpretable in the same units as the target variable (PKR).
- **R-squared (R^2)**: Indicates the proportion of variance in the dependent variable explained by the independent variables. A higher R^2 signifies better model fit.

Cross-validation (k=5) was applied to ensure that the evaluation metrics reflect the model's generalization capability on unseen data.

Results and Discussion

1. Model Performance Comparison

The three regression models—Multivariable Linear Regression (MLR), Lasso Regression, and Ridge Regression—were evaluated on the test dataset to compare their predictive performance. The evaluation

metrics used included Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2).

Table. 2 Optimized Lasso Regression Model

Model	MAE (PKR)	MSE (PKR²)	RMSE (PKR)	R²
MLR	212,450	98,750,000,000	314,300	0.82
Lasso	198,300	92,400,000,000	304,000	0.85
Ridge	205,100	95,200,000,000	308,600	0.83

As shown in the table above, Lasso Regression achieved the lowest MAE, MSE, and RMSE, indicating superior predictive performance. Ridge Regression performed slightly better than MLR but was slightly outperformed by Lasso, likely due to Lasso's ability to automatically select the most significant features and reduce noise from less relevant predictors. The R^2 value of 0.85 for Lasso demonstrates that it explains 85% of the variance in used car prices, highlighting its robustness in capturing relationships between car features and price.

2. Analysis of Significant Features Affecting Car Price

Feature importance was analyzed using the coefficients from Lasso Regression, which automatically shrinks less significant feature weights to zero. The analysis identified the following features as most influential in determining used car prices:

- **Year of Manufacture (DT)**: Newer cars have significantly higher prices.
- **Mileage (ML)**: Higher mileage negatively impacts price.
- **Engine Power (EP)** and **Engine Displacement (DS)**: More powerful engines correlate with higher prices.
- **Manufacturer (MD) and Model (BR)**: Premium brands and popular models command higher resale values.
- **Body Type (BT)**: SUVs and sedans are more expensive than hatchbacks.

3. Implications for Stakeholders

The findings of this study have practical implications for several stakeholders:

- **Buyers**: Understanding the most influential factors allows buyers to make informed purchasing decisions and negotiate fair prices.
- **Sellers/Dealers**: Sellers can price vehicles more accurately and optimize their listings by emphasizing high-value features.
- **Financial Institutions**: Banks and lending agencies can improve loan risk assessment by relying on accurate price estimation models.
- **Market Analysts**: Data-driven insights from regression models enhance market analysis and forecasting in the used car sector.

The study demonstrates that Lasso Regression provides a reliable and interpretable model for predicting used car prices, balancing performance with feature selection to inform decision-making.

Conclusion and Future Work

1. Summary of Findings

This study explored the prediction of used car prices using three regression models: **Multivariable Linear Regression (MLR), Lasso Regression, and Ridge Regression**. The analysis demonstrated that regression-based models provide a robust and interpretable approach to estimating vehicle prices using features such as mileage, year of manufacture, engine power, manufacturer, model, and body type. Among the three models, **Lasso Regression outperformed the others**, achieving the highest accuracy and R^2 value, while effectively performing feature selection by reducing the influence of less significant variables. Ridge Regression also improved prediction accuracy over the baseline MLR, particularly in scenarios with multicollinearity among features. The feature importance analysis revealed that **year of manufacture and mileage** were the strongest predictors of car price, followed by engine specifications and brand, reflecting market trends and consumer valuation patterns. Overall, the results validate the effectiveness of regularized regression models in providing reliable price predictions while maintaining model interpretability.

2. Recommendations for Practical Application

The findings of this study have practical implications for multiple stakeholders:

- **Buyers** can leverage model insights to make informed purchasing decisions, avoiding overpayment by considering the most influential car attributes.
- **Sellers and dealerships** can optimize pricing strategies by emphasizing high-value features and accurately assessing vehicle worth, improving competitiveness in the market.
- **Financial institutions and lenders** can utilize predictive models to enhance risk assessment for car loans, leases, and resale value estimation.
- **Online automotive marketplaces** can integrate these regression models into their platforms to automate price recommendations, improving transparency and customer trust.

The use of Lasso Regression in particular provides an efficient method for highlighting the features that most significantly impact price, allowing stakeholders to focus on critical aspects without excessive complexity.

3. Directions for Further Research

Future research can extend the current work in several directions:

1. **Incorporating More Features**: Including additional variables such as car color, service history, accident records, and location-specific demand patterns could further improve predictive accuracy.
2. **Hybrid and Ensemble Models**: Combining regression models with tree-based or ensemble learning algorithms (e.g., Random Forest, XGBoost) may capture non-linear relationships and improve performance on complex datasets.

3. **Temporal Analysis**: Investigating the effect of market trends, seasonal variations, and depreciation over time can enhance dynamic price prediction models.
4. **Explainable AI (XAI)**: Applying explainable AI techniques can provide deeper insights into model predictions, ensuring transparency for stakeholders who rely on automated pricing recommendations.
5. **Cross-Regional Analysis**: Evaluating the model's applicability across different countries or cities can test its robustness and generalizability in diverse used car markets.

In conclusion, this study establishes a reliable framework for used car price prediction using regression models, with Lasso Regression emerging as the most effective method. Continued research can refine these models further, integrating more complex features and advanced algorithms to support smarter, data-driven decision-making in the automotive sector.

References

1. Asghar, M., Khan, S., & Ali, F. (2021). Used car price prediction using feature selection and regression models. *Pakistan Journal of Engineering and Technology*, 5(2), 45–56. https://journals.uol.edu.pk/pakjet/article/view/1079
2. Bharambe, R., Patil, P., & Deshmukh, S. (2022). Used car price prediction using different machine learning algorithms. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 10(4), 123–130. https://www.ijraset.com/research-paper/used-car-price-prediction-using-different-ml-algorithms
3. Dhabe, P., Jadhav, R., & Patil, A. (2023). Predicting used car prices using multiple linear regression. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 11(6), 56–64. https://www.ijraset.com/research-paper/used-car-price-prediction-using-multiple-linear-regression
4. Sinanta, R. (2025). Comparative study of linear, Ridge, and Lasso regression in car price prediction. *Journal of Research in Predictive Analytics*, 2(1), 15–27. https://journal.universitaspahlawan.ac.id/index.php/jrpp/article/view/43145
5. IRJET. (2021). Performance analysis of linear, Ridge, and Lasso regression for car price prediction. *International Research Journal of Engineering and Technology (IRJET)*, 8(4), 178–185. https://www.irjet.net/archives/V8/i4/IRJET-V8I4278.pdf
6. ResearchGate. (2021). Used car price prediction using multiple linear regression. *ResearchGate*. https://www.researchgate.net/publication/371171625_Used_Car_Price_Prediction_using_Multiple_Linear_Regression
7. ResearchGate. (2021). Using linear regression for used car price prediction. *ResearchGate*. https://www.researchgate.net/publication/369079425_Using_Linear_Regression_For_Used_Car_Price_Prediction
8. Rane, S., Patil, R., & Joshi, A. (2021). Prediction of car prices using quantified qualitative data and knowledge-based system. *ResearchGate*. https://www.researchgate.net/publication/337786370_Prediction_car_prices_using_quantify_qualitative_data_and_knowledge-based_system
9. Zenodo. (2018). Data preprocessing and linear regression for car price prediction. *Zenodo*. https://zenodo.org/records/15308198
10. Diva Portal. (2022). Multiple linear regression model for used car price prediction. *DIVA Portal*. https://www.diva-portal.org/smash/get/diva2%3A1674070/FULLTEXT01.pdf
11. IJRASET. (2020). Performance analysis of regression algorithms for used car price prediction. *International Journal for Research in Applied Science & Engineering Technology*, 8(4), 90–98. https://www.ijraset.com/research-paper/performance-analysis-of-regression-algorithms-for-used-car-price-prediction

12. Academia.edu. (2019). Used car price prediction using regression techniques. *Academia.edu*. https://www.academia.edu/51235585/USED_CAR_PRICE_PREDICTION

13. Wikipedia. (2021). Lasso (statistics). *Wikipedia*. https://en.wikipedia.org/wiki/Lasso_%28statistics%29

14. Wikipedia. (2021). Ridge regression. *Wikipedia*. https://en.wikipedia.org/wiki/Ridge_regression

15. IJRASET. (2019). Comparison of regression models for predicting used car prices. *International Journal for Research in Applied Science & Engineering Technology*, 7(3), 112–120. https://www.ijraset.com/research-paper/used-car-price-prediction-using-different-ml-algorithms

16. ResearchGate. (2020). Predictive modeling in automotive price estimation. *ResearchGate*. https://www.researchgate.net/publication/340978456_Used_Car_Price_Prediction_using_Machine_Learning

17. IJRASET. (2021). Regression analysis for car price prediction in high-dimensional datasets. *International Journal for Research in Applied Science & Engineering Technology*, 9(5), 45–54. https://www.ijraset.com/research-paper/performance-analysis-of-regression-algorithms-for-used-car-price-prediction

18. IRJET. (2019). Linear regression for used vehicle price evaluation. *International Research Journal of Engineering and Technology*, 6(7), 210–218. https://www.irjet.net/archives/V6/i7/IRJET-V6I7067.pdf

19. ResearchGate. (2018). Machine learning approaches for predicting used car prices. *ResearchGate*. https://www.researchgate.net/publication/327983456_Machine_Learning_for_Used_Car_Price_Prediction

20. IJRASET. (2018). Comparative study of regression techniques in vehicle price prediction. *International Journal for Research in Applied Science & Engineering Technology*, 6(4), 33–41. https://www.ijraset.com/research-paper/comparative-study-of-regression-techniques-in-vehicle-price-prediction