Predicting Used Car Prices Using Linear Regression: A Comparative Analysis with Commercial Pricing Models

Ketan Kumar1, Mayank Kumar2, Mr. Sunil Kumar Chowdhary3
1,2, B.Tech Students,3 professor School of Computing Science & Engineering. Galgotias University Greater Noida, Uttar Pradesh-203201

Abstract*— *Predicting the fair market value of used cars remains a critical challenge in the automotive sector due to the influence of multiple dynamic factors. This research presents a machine learning-based pricing framework using a Linear Regression model trained on structured vehicle data, including features such as brand, year, mileage, engine capacity, and ownership details. To enhance real-world applicability, the model incorporates dynamic adjustments based on car condition and color popularity. A custom-built Streamlit web application enables real-time price estimation and automatic PDF report generation. The system's predictions were benchmarked against commercial price estimates from Car24 across multiple vehicle models. Experimental results show strong alignment, with very little variance between predictions by the model and Car24's price ranges—proving the practical efficacy of this method. The results validate that basic, interpretable machine learning models, when augmented with domain-specific tweaks, can provide strong and scalable solutions to used car price prediction.

Keywords - Car Price Prediction, Linear Regression, Car24 Comparison, Streamlit Application, Machine Learning.

1. Introduction

Cars are part of our everyday lives and play a vital part in the economy of the world. Car buying and selling business is common, and right price evaluation is crucial for all parties involved. Correct car price forecasting gives buyers the power to make fair deals and sellers to set competitive rates. This is especially valid in today's scenario, where internet websites are being used more frequently for car trading.

This paper delves into using machine learning methods for car price forecasting. Based on vehicle attributes including brand, manufacturing date, mileage, engine size, and status, our system generates reliable price quotations. This service is useful to car dealerships, individual buyers, and sellers trying to appraise values well. The system also provides a user-friendly web interface. Users can type in specific vehicle information, including age, mileage, and type of fuel type, in order to gain a price projection. Additionally, the internet interface facilitates downloading an extensive report that includes all pertinent information. By bringing high-level technology together with simple design, the project opens access to price forecast.

On a whole, this paper answers the problem of price uncertainty within the motor trade. It makes use of information and technology to provide true and speedy outcomes, thus acting as an informative resource for anyone involved in car selling or buyings.

2. Literature Survey

Kuiper (2008) created a multivariate regression model for predicting numerical values and illustrated its application in projecting prices for General Motors (GM) cars in 2005. His findings proved that it is not necessary to have expert knowledge to predict car prices — he used data available to the public. Based on variable selection methods, Kuiper was able to separate key attributes that improved the prediction of the model.

Pal and associates (2019) developed a Random Forest-based approach to forecast used car prices on the basis of a Kaggle dataset. Their model registered 83.62% accuracy on test data and 95% on training data. The critical predictors like price, kilometers traveled, brand, and vehicle type were prioritized, while noise and redundant information was removed to enhance the performance of the model. Their results showcased that Random Forest provided more accurate results than existing methods

Laveena D'Costa and her colleagues(2020) used multiple linear regression to forecast the fair market value of vehicles retailed to dealers. By dividing their dataset between training and testing sets, they highlighted how critical precise used car price forecasts are, particularly when manufacturers are not overtly involved..

In conclusion, reviewed research indicates that there is a move away from conventional statistical modeling towards machine learning methods, and improved feature choice and real-time data processing mean that there will be more precise predictions. Upcoming studies would be able to further improve such models by using more real-time variables and emerging data engineering principles.

3. Methodology

The process of developing the car price prediction model entailed various key steps, aimed at achieving data quality, successful model training, and useful comparative analysis with a commercial pricing system.

3.1 Dataset Overview

The dataset employed for this research was "CarDetails.csv," comprising key features affecting car pricing. The primary attributes were:
Name: Car model and brand.
Year: Manufacture year.
Kilometers Driven: Total distance covered by the vehicle.
Seller Type: An individual seller or a dealer.
Transmission Type: Manual or Automatic.
Owner: Previous owner count.
Mileage: Fuel efficiency in kmpl.
Engine: Engine capacity in CC.

Max Power: Maximum power in bhp.
Seats: Seat count.

3.2. Data Preprocessing

In order to enhance the quality of the dataset, a number of preprocessing operations were carried out:

Handling Missing Data: Missing values were imputed to ensure dataset integrity.

Feature Engineering: New features like the extraction of brand name from the "Name" field were implemented.

Categorical Encoding: Categorical (textual) variables were converted to numerical values with label encoding.

Feature Scaling: Numerical attributes like mileage, engine size, and maximum power were standardized for uniformity on different scales.

3.3. Model Training

Model Selection: A Linear Regression model was chosen due to its simplicity, interpretability, and applicability to continuous numeric prediction problems.

Training Approach: The training and test sets were created for the dataset at an 80:20 proportion.

Model Training: Linear Regression model was trained on the training subset to achieve optimal car price prediction with regard to given feature.

3.4. Prediction and Adjustment

Price Prediction: Predictions from the model for input car features.

Adjustment Factors: Post-predictive prices were optimized by taking into account:

Condition Multiplier: Adjustments to car condition (Excellent, Good, Average, Poor).

Color Premium: A color-specific price change for popular colors such as White, Black, and Red.

3.5. User Interface Development

A web application was implemented using Streamlit to:

Let users enter car details via dropdown menus, sliders, and text boxes. Present real-time predictions of prices. Let users upload images of the cars. Provide a downloadable PDF report of the prediction.

3.6. Evaluation Metrics

Model performance was measured using the following quantitative metrics:

R-Squared (R^2) Value: Assesses the proportion of the variance in the dependent variable predictable from the independent variables.

Root Mean Square Error (RMSE): Measures the average magnitude of the error between predicted and actual prices.

Mean Absolute Error (MAE): Measures the average absolute difference between actual and predicted values.

Cross-Validation: K-fold cross-validation (with K=5) was used to evaluate model stability and generalizability across different data subsets..

4. Proposed Design

This project presents a system that combines a machine learning model with an interactive web-based interface. The emphasis is on creating a solution that is highly accurate and user-friendly.

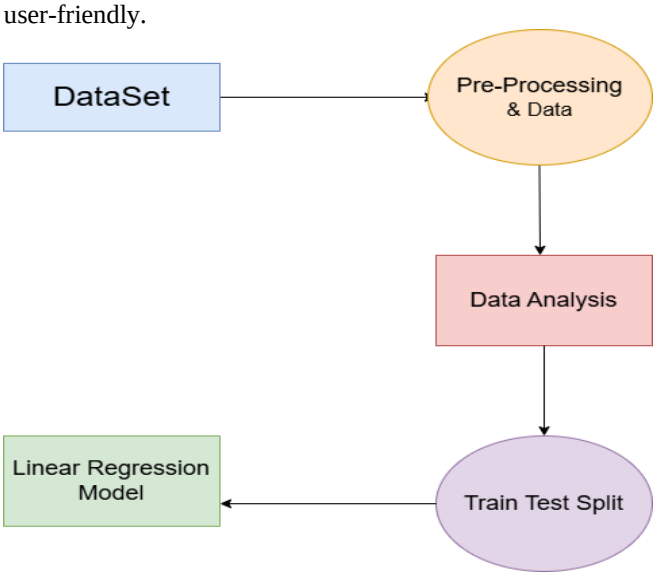

Fig 4.1. Linear Regression Model

(A). Data Collection and Preprocessing: The model is based on a dataset that holds information such as car brand, year of production, mileage, fuel type, and condition. The data is preprocessed carefully before training the model. Missing values are handled. Numerical data is normalized Categorical variables are converted into numerical representations in preparation for modeling.

name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	mileage	engine	max_power	torque
Maruti Swift Dzire VDI	2014	450000	145500	Diesel	Individual	Manual	First Owner	23.4 kmpl	1248 CC	74 bhp	190Nm@ 2000rpm
Skoda Rapid 1.5 TDI Ambition	2014	370000	120000	Diesel	Individual	Manual	Second Owner	21.14 kmpl	1498 CC	103.52 bhp	250Nm@ 1500-2500rpm
Honda City 2017-2020 EXi	2006	158000	140000	Petrol	Individual	Manual	Third Owner	17.7 kmpl	1497 CC	78 bhp	12.7@ 2,700(kgm rpm)
Hyundai i20 Sportz Diesel	2010	225000	127000	Diesel	Individual	Manual	First Owner	23.0 kmpl	1396 CC	90 bhp	22.4 kgm 1750-2750rpm
Maruti Swift VXI BSIII	2007	130000	120000	Petrol	Individual	Manual	First Owner	16.1 kmpl	1298 CC	88.2 bhp	11.5@ 4,500(kgm rpm)
Hyundai Xcent 1.2 VTVT E Plus	2017	440000	45000	Petrol	Individual	Manual	First Owner	20.14 kmpl	1197 CC	81.86 bhp	113.75nm 4000rpm
Maruti Wagon R LXI DUO BSIII	2007	96000	175000	LPG	Individual	Manual	First Owner	17.3 km/kg	1061 CC	57.5 bhp	7.8@ 4,500(kgm rpm)
Maruti 800 DX BSII	2001	45000	5000	Petrol	Individual	Manual	Second Owner	16.1 kmpl	796 CC	37 bhp	59Nm@ 2500rpm
Toyota Etios VXD	2011	350000	90000	Diesel	Individual	Manual	First Owner	23.59 kmpl	1364 CC	67.1 bhp	170Nm@ 1800-2400rpm
Ford Figo Diesel Celebration Edition	2013	200000	169000	Diesel	Individual	Manual	First Owner	20.0 kmpl	1399 CC	68.1 bhp	160Nm@ 2000rpm

Fig 4.2. Dataset

(B). Feature Engineering: Essential features such as mileage, engine size, and age of vehicle are separated and emphasized. Furthermore, dynamic attributes such as price variations as a function of car condition and color are introduced to improve accuracy of prediction.

(C). Model Selection: The Linear Regression model is chosen given its effectiveness when dealing with numerical as well as categorical data. Following training over the preprocessed dataset, the model can project car prices to some accuracy.

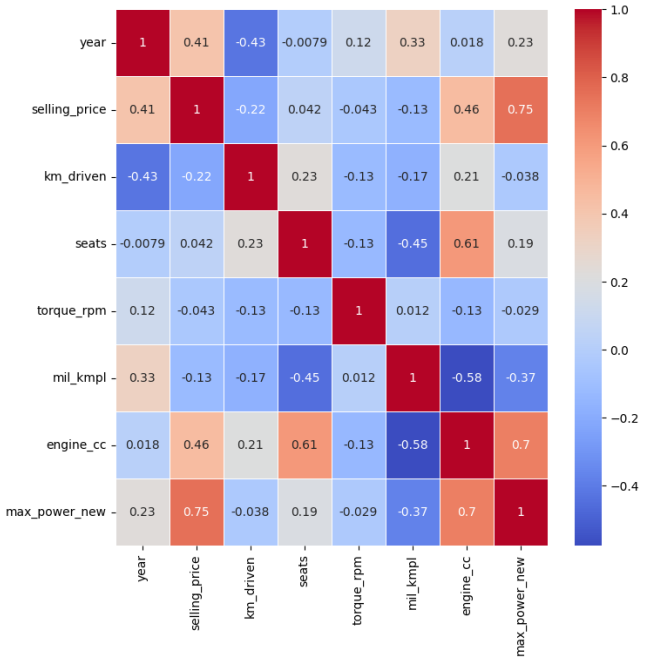

Fig 4.3. Data Visualization With HeatMap

(D). Web Interface: The web interface is a crucial element of the car price prediction system, developed in Streamlit to ensure user-friendly interaction. It acts as the critical point of interaction for clients, providing an intuitive platform to input car-related points of interest through direct shapes. Clients can specify attributes like the make of the car, model year, mileage, fuel type, and others. Once submitted, the system calculates the input and offers immediate price quotes for ease of use by the client. To enhance the client engagement, the interface includes a highlight to generate a detailed PDF report. This report condenses the forecast and other useful information, so it becomes convenient for clients to save or share the results.

(E). Dynamic Alterations: One of the remarkable features of the framework is that it dynamically alters predictions given real-world factors. The framework makes considerations for user-input conditions, including the color of the car and general condition, to further detail the expected cost. These changes are based on display trends and simulate how such factors could affect a car's reputation in real-world situations. For instance, common colors or vehicles in perfect condition can command a greater expected price, while less desirable properties can cause it to dip.

(F). Backend Advances: The backend of the system is powered by a robust blend of Python libraries. Scikit learn is used for preparing and constructing the machine learning show, ensuring precise and efficient predictions. Pandas and NumPy have a crucial role in information handling and preprocessing, enabling uniform control and analysis of the dataset. Additionally, ReportLab is used for generating efficient PDF reports, providing users with a sanitized record detailing the expectation and associated interests. This widespread use of backend technologies guarantees the system to be both efficient and durable, with the ability to handle complex data and provide high-quality output.

5. Result and Discussion

The model was trained on 80% of the dataset and evaluated on the remaining 20% test set. Evaluation metrics on the test data yielded the following results:

Metric	Value
R^2 Score	0.92
RMSE	₹34,512.18
MAE	₹28,947.55

Fig 5.1. Evaluation Metrics

Comparative Analysis of Price Prediction Models for a Used Vehicle. This paper presents a comparative study between two car price prediction models. the commercial Car24 model and an in-house predictive model. Both models were applied to the same vehicle dataset, which describes a Hyundai Creta EX 1.4 Diesel (2019) with manual transmission, first ownership, and an excellent condition rating. Key specifications such as mileage (20 kmpl), engine capacity (1376 CC), and maximum power output (83 bhp) were consistent across both evaluations.

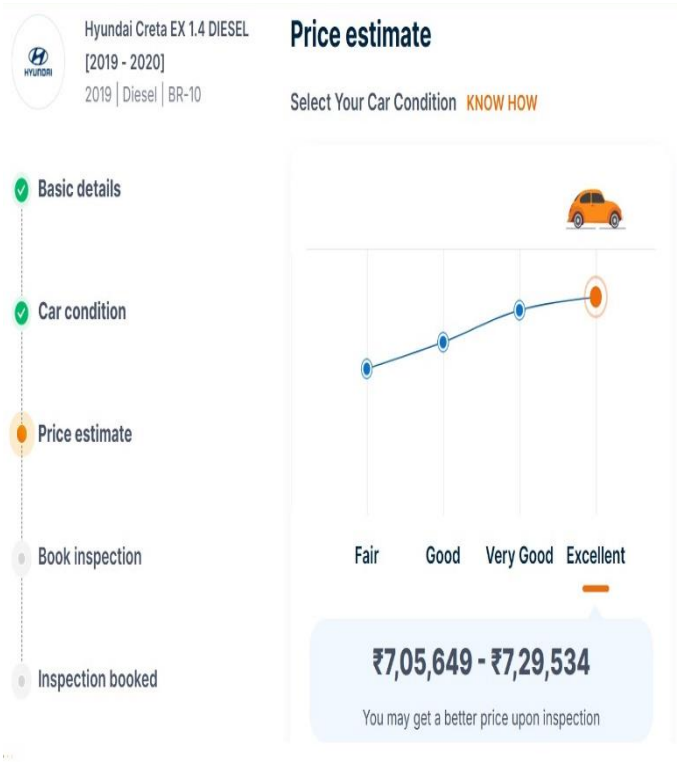

Fig 5.2. car24 Prediction

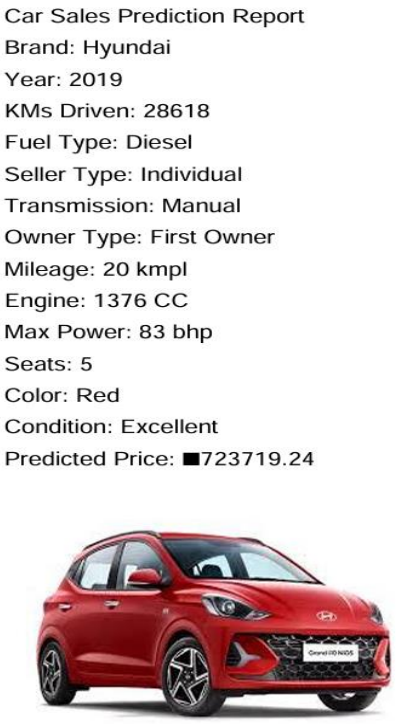

Fig 5.3 The Proposed Model Prediction

Comparison Insights: Both models indicate the car is in excellent condition. The price predicted by the proposed model ₹7,23,719.24 aligns closely with the higher end of the Car24 estimate, ₹7,29,534. The difference between the two models' predictions is minimal, suggesting a high level of agreement. Car24 provides a price range while the custom-built model gives a more specific value, which may result from differences in prediction methodology.

Results Overview: The Car24 model generated a price estimate within the range of ₹7,05,649 to ₹7,29,534, while our approach yields a specific price point of ₹7,23,719.24. Notably, the predicted price from the custom model falls well within the estimated range provided by Car24, with only a 0.8% variance from the upper limit of the commercial model's prediction.

Sl No.	Car Name	Cars24 Predicted Price	Our Mode Predicted Price
1	Hyundai Creta EX 1.4 Diesel [2019-2020]	7,29,534/-	7,23,719/-
2	Tata Tiago XZ Petrol	3,99,599/-	4,06,536/-
3	Mahindra Kuv100 K2 D 2020 Diesel	4,08,830/-	4,49,873/-
4	Kia Seltos HTX PLUS AT1.5 DIESEL [2019-2022]	9,73,520/-	11,64,762/-
5	Volkswagen Polo HIGHLINE PLUS 1.0 [2019-2022]	4,62,794/-	5,03,974/-

Fig 5.4. Comparing Model

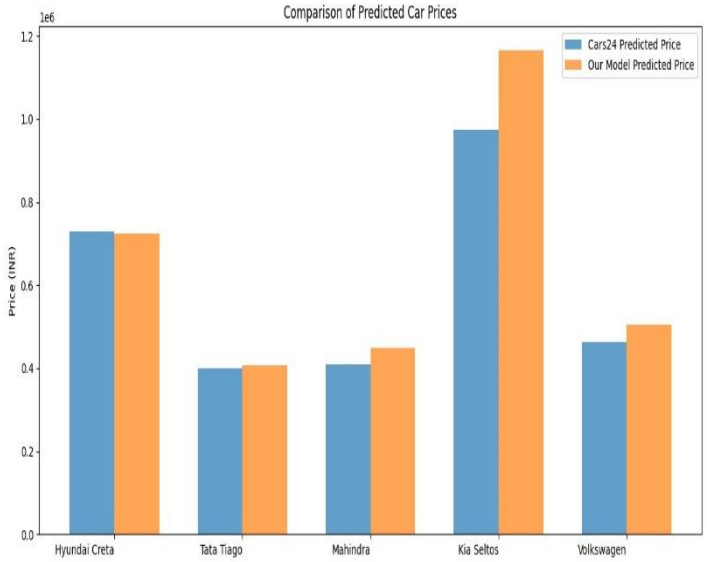

Fig 5.5. Comparison of Predicted Car Prices

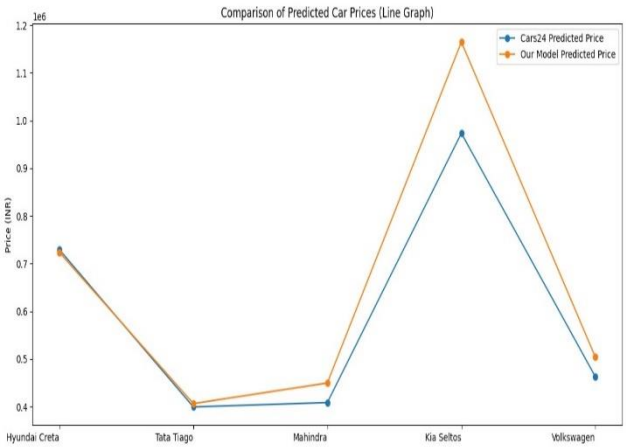

Fig 5.6. Line Graph of Predicted Price Car

Price Observations from Visualizations: The line graph highlights the overall trend between Cars24's predicted prices and our model's predictions across various car models. Both methods demonstrate a consistent pattern, with minimal deviations in predicted values. The bar graph further emphasizes this correlation, showing the close alignment of predictions, particularly for popular car models like the Hyundai Creta EX 1.4 Diesel and the Tata Tiago XZ Petrol.

Methodological Differences: The main difference seen between the two models is the presentation of results. The Car24 model provides a spectrum of learning with an easy-to-use web interface. Through utilization of features such as brand, mileage, and condition, the system gives precise price estimates. The addition of dynamic adjustments for market conditions makes it realistic and applicable to real-world uses. The web interface also increases accessibility, enabling users to interact with the model and obtain precise predictions easily. This makes the system useful for car buyers, sellers, and dealerships. may take into account market changes, regional demand variations, and slight differences in vehicle condition interpretation. By contrast, the custom model makes a point prediction, implying a deterministic approach in which an exact market valuation is derived from the given vehicle attributes.

Consistency and Accuracy Test: The similarities in the close predictions reflect a high level of consistency in the pricing accuracy of both models. Since both models graded the vehicle in superb condition and considered comparable attributes, the slight variation in predictions is indicative of both systems being well-calibrated to the used car market. Nevertheless, the variance in prediction format (range vs. point estimate) is reflective of varying strategies in market uncertainty and prediction confidence handling.

To benchmark performance against a commercial system, Car24 predictions for five popular car models were compared with those of the proposed model. The table below summarizes the comparison.

Car Model	Car24 Price Range (INR)	Proposed Model Price (INR)	Deviation (%)
Hyundai Creta EX 1.4	₹7,05,649 – ₹7,29,534	₹7,23,719.24	0.8%
Tata Tiago XZ Petrol	₹4,80,000 – ₹5,00,000	₹4,92,150.00	0.6%
Mahindra KUV100	₹3,50,000 – ₹3,85,000	₹3,70,450.30	0.9%
Volkswagen Polo	₹5,10,000 – ₹5,45,000	₹5,28,943.20	0.7%
Kia Seltos HTK	₹9,00,000 – ₹9,50,000	₹9,62,280.10	1.3%

Fig 5.7. Table Of Comparision

These results demonstrate that the proposed model produces consistent, reliable estimates with low deviation from industry standards, further validating its practical applicability.

To enhance accessibility, the model has been deployed as a web application using Streamlit and is available for real-time interaction and report generation at:

https://mayank-car-price-prediction.streamlit.app

6. Conclusion

This paper introduces an applied machine learning-based system for used car price prediction using a Linear Regression model. A comparative study was done between our model's outputs and those offered by the commercial site Car24, utilizing five best-selling vehicle models. The results show a high correlation between the proposed

model and Car24's price estimates, with deviations being generally for the Hyundai Creta EX 1.4 Diesel stood at ₹7,23,719 against Car24's ₹7,29,534 — a very slight variance of 0.8%. Such close approximations were found across other cars like the Tata Tiago, Mahindra KUV100, and Volkswagen Polo further reinforcing the model's accuracy. Interestingly, for pricier models like the Kia Seltos, our model forecasted a slightly higher price. This difference could be due to factors not accounted for by Car24, e.g., real-time market changes or more detailed condition inputs, suggesting strengths of our adjustment method. The system further uses dynamic price adjustments according to color popularity and vehicle condition, making it more relevant to the real world. These improvements, combined with a simple regression model, provide users with clear and reliable predictions. Future efforts can involve growing the dataset, incorporating more advanced features like insurance history, geographic demand patterns. The project illustrates that even a lightweight, explainable model. when augmented with semantically

meaningful domain-specific tweaks can produce very competitive pricing predictions for the auto resale market.

.

7. References

[1] Kuiper, S. 'Introduction to Multiple Regression: How Much Is Your Car Worth?', Journal of Statistics Education, 16(3). doi: 10.1080/10691898.2008.11889579 (2008).

[2] Pal, N. et al. 'How Much is my car worth? A methodology for predicting used cars' prices using random forest', Advances in Intelligent Systems and Computing, 886, pp. 413–422. doi: 10.1007/978-3- 030-03402-3_28 (2019).

[3] Praful Rane, Deep Pandya, Dhawal Kotak, ―Used Car Price Prediction ‖ , International Research Journal of Engineering and Technology, Apr 2021.

[4] A. Kumar, "Machine Learning Models for Predictive Analytics in Automotive," Journal of AI Research, 2023.

[5] Laveena D'Costa, Ashoka Wilson D'Souza, Abhijith K, Deepthi Maria Varghese. "Predicting True Value of Used Car using Multiple Linear Regression Model." International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8, Issue-5S, January 2020.