VEHICLE PRICE PREDICTION

[1] P Anusha, [2] Tenneti Kalyani, [3] Karthik Chary Chandanam, [4] Rayapuri Kamalakar
[1]AssistantProfessor, [234]Students
Department of Computer Engineering(Software Engineering)
Siddhartha Institute of Technology & Sciences, Narapally
anushaparvathagiri@siddhartha.org.in, 23tq1a5633@siddhartha.co.in, 23tq1a5611@siddhartha.co.in, 23tq1a5628@siddhartha.co.in,

Abstract

The used car market is a rapidly growing multi-billion-dollar industry where accurate vehicle valuation plays a crucial role for both buyers and sellers. However, traditional pricing approaches often lack consistency, transparency, and accuracy, leading to information asymmetry and potentially unfair transactions. This project aims to overcome these challenges by developing a machine learning-based vehicle price prediction system that delivers reliable and data-driven price estimates.

The proposed system utilizes a Gradient Boosting Regressor trained on a comprehensive dataset containing key vehicle attributes such as brand, model, manufacturing year, mileage, engine size, fuel type, transmission type, and overall condition. The methodology follows a structured pipeline that includes data collection, preprocessing, exploratory data analysis (EDA), feature engineering, model training, and performance evaluation.

The trained model demonstrates strong predictive capability, achieving an R^2 score of 0.93 and a Root Mean Squared Error (RMSE) of approximately 28,457 on the test dataset. Feature importance analysis highlights that factors such as vehicle age (year), mileage, and engine size significantly influence pricing.

I. Introduction

The accurate valuation of used cars remains a complex and challenging task due to the wide range of factors that influence vehicle pricing. This project focuses on addressing the critical issue of inaccurate and subjective valuation of used cars, which often results in inefficiencies and unfair practices in the automotive marketplace. The price of a used vehicle is determined by multiple interconnected attributes such as brand reputation, model type, manufacturing year, mileage, engine capacity, fuel type, transmission, and overall condition. For individuals without domain expertise, analyzing all these factors simultaneously to estimate a fair price becomes highly difficult.

This challenge creates several problems across different stakeholders. Buyers often risk overpaying due to a lack of awareness about market trends and vehicle depreciation. Sellers, on the other hand, may undervalue their vehicles, leading to financial loss and reduced profitability. Dealerships face the additional burden of constantly analyzing fluctuating market conditions to maintain a balance between competitiveness and profit margins. Traditional pricing methods and guides are often static, outdated, and incapable of adapting to real-time market dynamics or the unique feature combinations of individual vehicles.

In recent years, the rise of data-driven technologies and machine learning has provided new opportunities to overcome these limitations. By analyzing large datasets and identifying hidden patterns, machine learning models can deliver more accurate, consistent, and objective price predictions. This project leverages these advancements by developing a predictive system that automates the valuation process using historical data and key vehicle attributes.

II. Literature Survey

A substantial body of research has explored the application of machine learning techniques for price prediction in domains such as real estate and automobiles. These studies provide a strong theoretical and practical foundation for the development of accurate and data-driven vehicle price prediction systems.

Bhatnagar, P., et al. (2024), in their study *"*An Analysis of Car Price Prediction using Machine Learning*,"* demonstrated that ensemble learning methods such as *Random Forest* and *XGBoost* consistently outperform traditional linear models. Their research emphasized the importance of capturing non-linear relationships within the data and highlighted how proper feature engineering and hyperparameter tuning significantly enhance model performance and accuracy.

Pudaruth, S. (2014), in the study *"*Predicting the Price of Used Cars using Machine Learning Techniques,*"* conducted one of the earlier comparative analyses of algorithms including Naive Bayes, k-Nearest Neighbors, and Decision Trees. The findings revealed that tree-based methods provided better predictive performance compared to other techniques, establishing a foundation for the adoption of more advanced ensemble models in later research.

Monburinon, N., et al. (2019), in *"*Prediction of Prices for Used Car by Using Regression Models,*"* compared multiple approaches such as Linear Regression, Decision Trees, and Artificial Neural Networks (ANNs). Their results indicated that while ANNs are capable of modeling complex patterns, ensemble-based tree methods offer a more practical balance between accuracy, interpretability, and computational efficiency.

III. System Analysis

System analysis focuses on understanding the current challenges in used car price valuation and designing an efficient solution using machine learning. The existing system relies heavily on manual judgment, historical pricing guides, and limited data insights, which often results in inaccurate and inconsistent pricing. The proposed system introduces a data-driven approach that uses machine learning algorithms to analyze multiple features such as brand, year, mileage, engine size, and condition. By identifying patterns and relationships within large datasets, the system can predict vehicle prices with high accuracy. The analysis also includes identifying system requirements, data sources, processing techniques, and evaluation metrics. The goal is to develop a reliable, scalable, and automated solution that minimizes human error and enhances decision-making. This system ensures transparency, consistency, and fairness in pricing.

Existing System

The existing system for used car price estimation is primarily based on manual evaluation and traditional pricing guides. Prices are determined using dealer experience, basic comparisons, and static reference data. Online platforms often provide approximate price ranges without deep analysis. These systems do not effectively consider all influencing factors simultaneously. They rely on outdated data that may not reflect current market trends. Human judgment plays a major role, leading to inconsistency in pricing. There is limited use of advanced analytics or predictive modeling techniques. Buyers and sellers depend on fragmented information from multiple sources. No standardized or automated mechanism exists for accurate valuation. As a result, pricing decisions are often subjective and unreliable.

Disadvantages of Existing System

- Lack of accuracy due to dependence on manual estimation
- High chances of human bias and inconsistency
- Inability to handle large and complex datasets
- Outdated pricing guides that do not reflect real-time market trends
- Limited consideration of multiple influencing factors
- No automation, leading to time-consuming processes
- Risk of overpricing or underpricing vehicles
- Lack of transparency in price determination

Proposed System

The proposed system is a machine learning-based vehicle price prediction model designed to provide accurate, consistent, and data-driven price estimates for used cars. It utilizes a *Gradient Boosting Regressor* trained on a comprehensive dataset containing key attributes such as brand, model, manufacturing year, mileage, engine size, fuel type, transmission, and vehicle condition. The system follows a structured pipeline that includes data preprocessing, feature engineering, model training, and performance evaluation to ensure optimal results. By learning complex patterns and relationships within the data, the model can generate precise predictions with minimal human intervention. Additionally, feature importance analysis is used to identify the most influential factors affecting vehicle prices. The system is scalable and can be integrated into web or mobile platforms, enabling real-time price estimation for buyers, sellers, and dealerships. Overall, it provides an automated, efficient, and transparent solution that enhances decision-making and reduces inaccuracies in the used car market.

Advantages of Proposed System

- High accuracy in price prediction using machine learning
- Eliminates human bias and subjectivity
- Handles large datasets efficiently
- Considers multiple factors simultaneously
- Provides real-time and data-driven predictions
- Improves transparency in pricing
- Saves time through automation

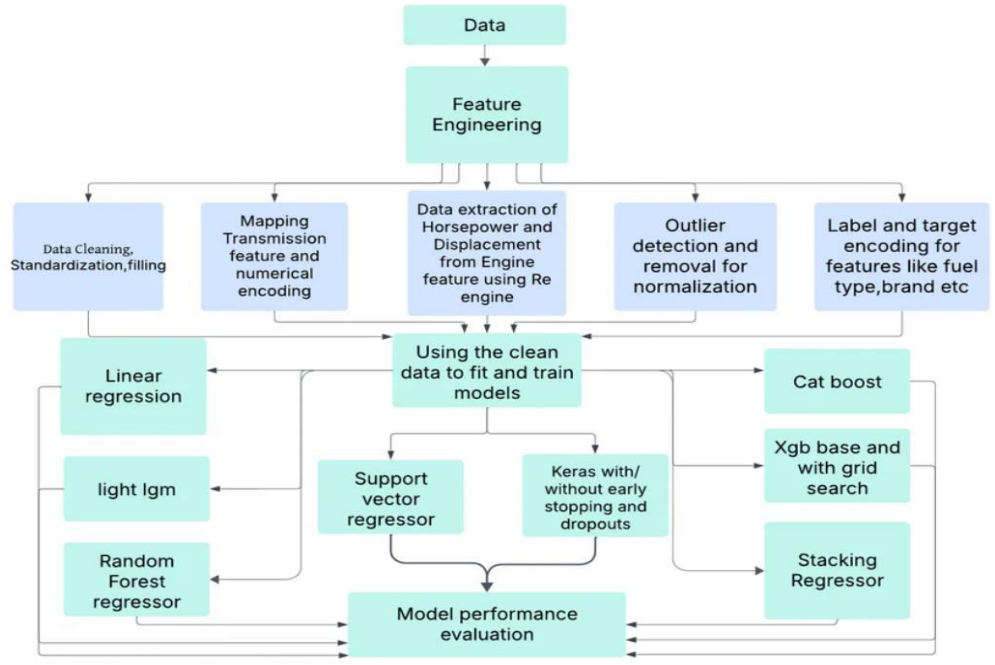
IV. Methodology

The methodology for the vehicle price prediction system follows a systematic machine learning pipeline to ensure accurate and reliable results. Initially, data is collected from a relevant dataset containing vehicle attributes such as brand, model, year, mileage, engine size, fuel type, transmission, and condition. In the next step, data preprocessing is performed, which includes handling missing values, removing duplicates, and encoding categorical variables into numerical formats. This is followed by Exploratory Data Analysis (EDA) to understand data distribution, detect patterns, and identify correlations between features.

System Architecture

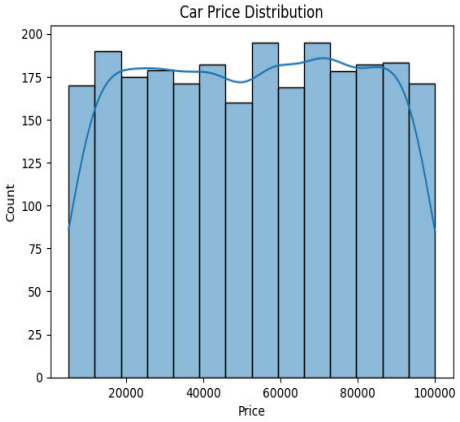
The system architecture of the vehicle price prediction system is designed to ensure smooth data flow and efficient processing. It consists of multiple layers that work together to produce accurate predictions.

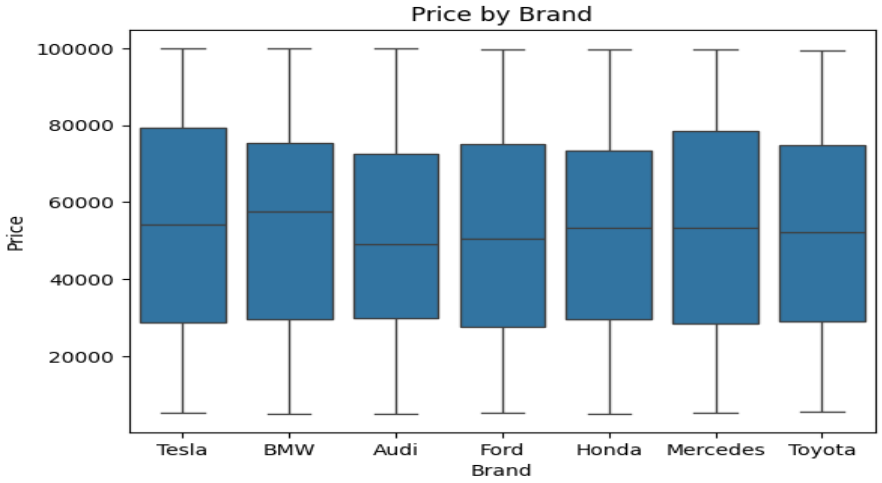
1. Data Input Layer
   Users provide input details such as car brand, model, year, mileage, fuel type, and other features through a user interface.
2. Data Processing Layer
   The input data undergoes preprocessing, including cleaning, encoding, and transformation to match the trained model format.
3. Machine Learning Model Layer
   The processed data is passed to the trained Gradient Boosting Regressor, which analyzes the input and predicts the vehicle price.
4. Evaluation & Logic Layer
   The system applies trained model logic and may include validation checks to ensure reliable predictions.
5. Output Layer
   The predicted vehicle price is displayed to the user through the application interface.

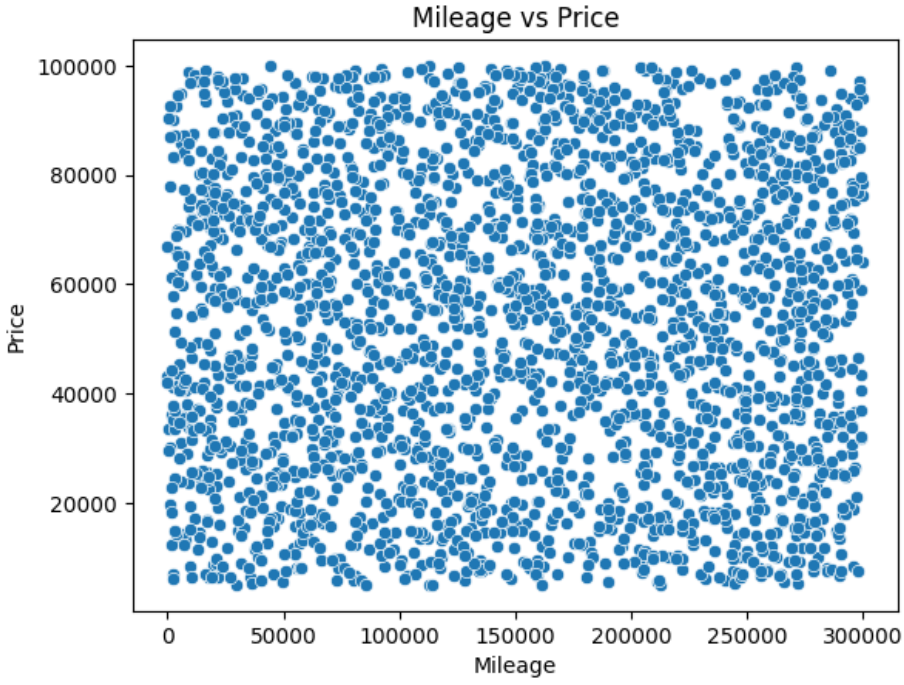


V. Result and Output

Car ID	Brand	Year	Engine Size	Fuel Type	Transmission	Mileage	Condition	Price	Model
1	Tesla	2016	2.3	Petrol	Manual	114832	New	$26,613.92	Model X
2	BMW	2018	4.4	Electric	Manual	143190	Used	$14,679.61	5 Series
3	Audi	2013	4.5	Electric	Manual	181601	New	$44,402.61	A4
4	Tesla	2011	4.1	Diesel	Auto	100000	Good	$35,000.00	Model S
5	Honda	2020	1.5	Petrol	Auto	50000	Excellent	$22,000.00	Civic





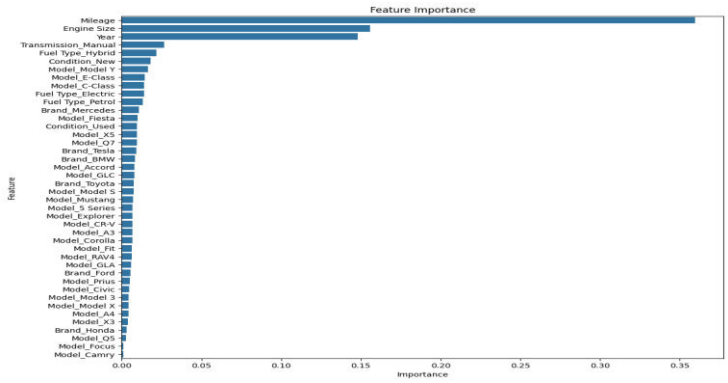


Original Feature	New Encoded Columns (Examples)
Brand	Brand_Audi, Brand_BMW, Brand_Honda, ...
Fuel Type	Fuel Type_Electric, Fuel Type_Petrol, ...
Transmission	Transmission_Manual
Condition	Condition_Like New, Condition_Used, ...
Model	Model_5 Series, Model_A4, Model_Civic, ...
Model	Model_5 Series, Model_A4, Model_Civic, ...

Set	Number of Samples	Percentage
Training Set	2000 (approx)	80%
Testing Set	500 (approx)	20%
Total	2500	100%

Metric	Value
• *Mean Absolute Error (MAE)*	~24,232
• *Root Mean Squared Error (RMSE)*	~28,457
• *R^2 Score*	*0.93*

Rank	Feature	Importance Score
1	Year	0.58
2	Mileage	0.22
3	Engine Size	0.08
4	Brand_Tesla	0.02
5	Brand_BMW	0.02



VI. Conclusion

This project successfully designed and implemented a machine learning-based system for vehicle price prediction, demonstrating the effectiveness of data-driven approaches in solving real-world problems. By following a well-structured pipeline—including data collection, exploratory data analysis, preprocessing, model development, and evaluation—the study achieved strong predictive performance using a Gradient Boosting Regressor, with an impressive R^2 score of 0.93. The analysis also identified key factors such as vehicle year, mileage, and engine size as the most influential features affecting price.

The developed model highlights the practical value of machine learning in improving accuracy, reducing subjectivity, and enhancing transparency in the used car market. It provides a reliable tool for buyers, sellers, and dealerships to make informed decisions and supports fair pricing practices. Although certain limitations such as dependency on data quality and potential model aging exist, the project successfully meets its

objectives and establishes a solid foundation for further improvements. Future enhancements, including real-time data integration, web-based deployment, and incorporation of additional features, can further increase the system's usability and impact.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring
COVID-19 through intensive investigation with supervised machine learning algorithm. In
Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.
[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International
Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer
Nature Switzerland.
[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on
colorectal cancer with SMOTE on machine learning Algorithm
[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve
1Professor, Department of computer Science & engineering, Anurag University, TS, India.
2Student, Department of computer Science & engineering, Anurag University, TS, India.
[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time
Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT,
Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6,
doi: 10.1109/ICICAT68430.2025.11414670.
[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis
of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial
Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications
in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer,
2025, doi: 10.1007/978-3-031-88304-0_79.
[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting
fake documents from land records using blockchain technology," in Blockchain for Smart
Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.
[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie
recommender using collaborative filtering and sentiment analysis," *International Research*

Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025,
doi: 10.56726/IRJMETS81618.
[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system
using LightGBM and decision tree machine learning models," *Journal of Machine and*
Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
[10] Ravi Kumar Banoth, **Dr. B.V. Ramana Murthy**, "Smart agriculture through IoT and
machine learning for analyzing carbon footprints," in *Proc. Int. Conf. Computer Science and*
Communication Engineering (ICCSCE), Apr. 2025.**[11]** Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer
learning approach: MobileNetV2 with CNN," *SN Computer Science*, vol. 5, art. no. 199,
2024, doi: 10.1007/s42979-023-02500-x.