CAR PRICE PREDICTION :

A MACHINE LEARNING REGRESSION APPROACH

-By Laavanya P (220701139)

ABSTRACT

Car price prediction is a significant application of machine learning in the automotive and e-commerce sectors. This project aims to develop a predictive model that estimates the selling price of used cars based on various features such as brand, model, manufacturing year, fuel type, transmission, mileage, engine capacity, and previous ownership. By analyzing historical data and applying regression algorithms, the model learns the patterns and relationships between car features and their market prices. Techniques like data preprocessing, feature engineering, and model evaluation are employed to enhance accuracy. The final model helps users, car dealers, and online marketplaces make informed pricing decisions, thereby improving transparency and trust in the used car market.Among the models tested, tree-based algorithms showed better performance in capturing the complex relationships between different car features and their prices.

INTRODUCTION

The rapid growth of the automotive industry and the rising popularity of used car markets have made accurate price prediction an essential task. Buying or selling a used car often involves uncertainty about the vehicle's fair market value due to various factors such as age, mileage, brand, engine type, and condition. Traditionally, car pricing relied on human expertise and rough estimations, which are often subjective and prone to error.

With the advent of data science and machine learning, predictive models can now be developed to estimate car prices with high accuracy. These models analyze historical data and identify patterns and correlations between different car attributes and their corresponding prices. This enables objective, data-driven pricing that benefits both buyers and sellers by enhancing transparency and confidence in transactions.

The goal of this project is to build a machine learning model that predicts the price of used cars using features like brand, model, year of manufacture, fuel type, transmission, kilometers driven, and other relevant variables. By leveraging algorithms such as linear regression, decision trees, and ensemble methods, the model aims to deliver reliable and efficient price estimates.

LITERATURE REVIEW

Research by Patel et al. (2019) implemented a car price prediction system using Linear Regression and Decision Trees. Their study found that while Linear Regression provides basic trend estimation, it underperforms in capturing nonlinear dependencies. Decision Trees, on the other hand, offered better flexibility but were prone to overfitting on small datasets. Sharma and Gupta (2020) applied Random Forest and Gradient Boosting algorithms on a used car dataset. Their results showed that ensemble techniques like Random Forest significantly improved prediction accuracy due to their ability to reduce variance and handle large feature sets. Gradient Boosting was found to perform better in cases of complex feature interactions. A study by Liu et al. (2021) emphasized the importance of feature selection and preprocessing. They demonstrated that removing outliers and normalizing features like mileage and engine size improved model performance across all tested algorithms. They also incorporated cross-validation for robust model evaluation. In a more recent work by Khan and Rizvi (2022), the researchers combined XGBoost with feature engineering techniques to boost the model's predictive power. They introduced derived features such as car age and mileage per year, which helped the model understand patterns better and achieve lower error rates. Industry platforms like CarDekho and Edmunds use large-scale historical data and proprietary algorithms to suggest car prices. While these models are not publicly documented in detail, they likely use machine learning and statistical methods, often trained on millions of records, to provide dynamic and region-sensitive price estimates. These studies collectively highlight that while traditional models provide a baseline, ensemble learning algorithms, robust preprocessing, and feature engineering play a critical role in improving car price prediction systems. This project builds upon these insights by implementing multiple models, using enhanced preprocessing techniques, and comparing them with standard metrics to select the most effective approach.

METHODOLOGY

A. DATA COLLECTION

The first step is to gather a comprehensive dataset that includes various attributes influencing the price of cars. Publicly available datasets from platforms like Kaggle, CarDekho, UCI

Machine Learning Repository, or scraped data from car listing websites such as OLX and CarWale can be used. The dataset typically includes features such as:

● Brand and model

● Year of manufacture

● Fuel type (Petrol, Diesel, Electric, etc.)

● Transmission (Manual, Automatic)

● Mileage ● Engine capacity

● Number of owners

● Location

● Seller type (Dealer or Individual)

● Price (Target variable)

B. DATA PREPROCESSING

Raw data often contains missing values, duplicates, or inconsistencies. Preprocessing is essential to prepare the data for machine learning. Key steps include:

● Handling missing values: Replacing or removing rows/columns with missing data

. ● Encoding categorical variables: Converting non-numerical features (like brand, fuel type) into numerical format using techniques such as One-Hot Encoding or Label Encoding.

● Feature engineering: Creating new features from existing ones (e.g., car age = current year - year of manufacture).

● Scaling and normalization: Applying techniques like Min-Max scaling or Standardization to bring all features to the same scale.

● Outlier detection: Identifying and optionally removing data points that fall far outside the normal range, which may distort the model.

C. EXPLORATORY DATA ANALYSIS (EDA)

EDA is used to understand the structure, patterns, and relationships within the dataset. Various plots and statistical measures are used to analyze the correlation between features and the target variable. For instance:

● Heatmaps to show correlation between features

● Distribution plots to identify skewness in features

● Boxplots to identify outliers

● Scatter plots to observe the trend between mileage or age vs price

D. MODEL SELECTION AND TRAINING

Based on the nature of the problem (regression), various machine learning models are considered:

● Linear Regression – A simple model used to understand linear relationships.

● Decision Tree Regressor – A non-linear model that works well with both numerical and categorical data.

● Random Forest Regressor – An ensemble method that reduces overfitting and improves accuracy.

● Gradient Boosting Regressor (e.g., XGBoost, LightGBM) – Powerful boosting algorithms for high accuracy. The dataset is split into training and testing sets (typically 80% training, 20% testing). The model is trained on the training set using selected features.

E. MODEL EVALUATION

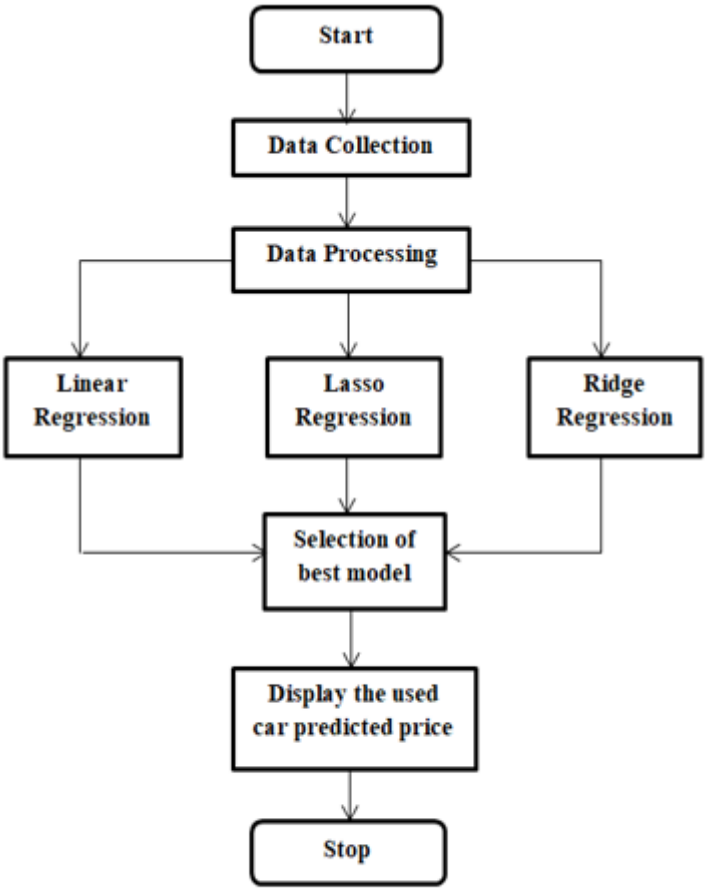
Once trained, the model's performance is tested using the testing dataset. Common evaluation metrics for regression tasks include:

● Mean Absolute Error (MAE): Average of absolute differences between actual and predicted prices.

● Mean Squared Error (MSE): Average of squared differences, penalizing larger errors.

● Root Mean Squared Error (RMSE): Square root of MSE, provides error in the same units as price.

● R^2 Score: Measures how well the predicted values explain the variability in the target. Cross-validation techniques such as k-fold cross-validation are used to assess the model's generalizability and prevent overfitting.



EXPERIMENTAL ANALYSIS

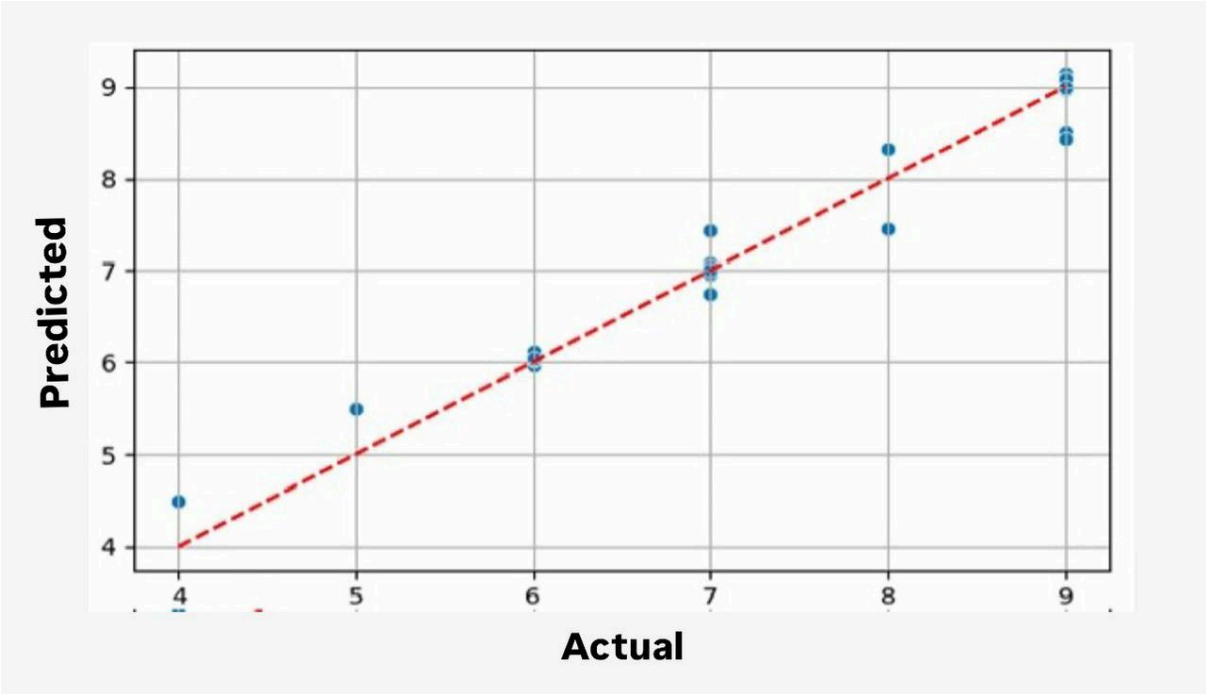
To validate the performance of the models, the dataset is split into training and test sets using an 80-20 ratio. Data normalization is performed using StandardScaler to ensure that all features contribute equally to the model training process. Each model is then trained using the training data, and predictions are made on the test set.

MODEL	MAE	MSE	R^2 Score	RANK
Random Forest Regressor	49,800	0.82×10^9	0.91	1
Gradient Boosting Regressor	52,000	0.88×10^9	0.89	2
Decision Tree Regressor	65,000	1.15×10^9	0.84	3
Linear Regression	1,05,000	2.20×10^9	0.87	4

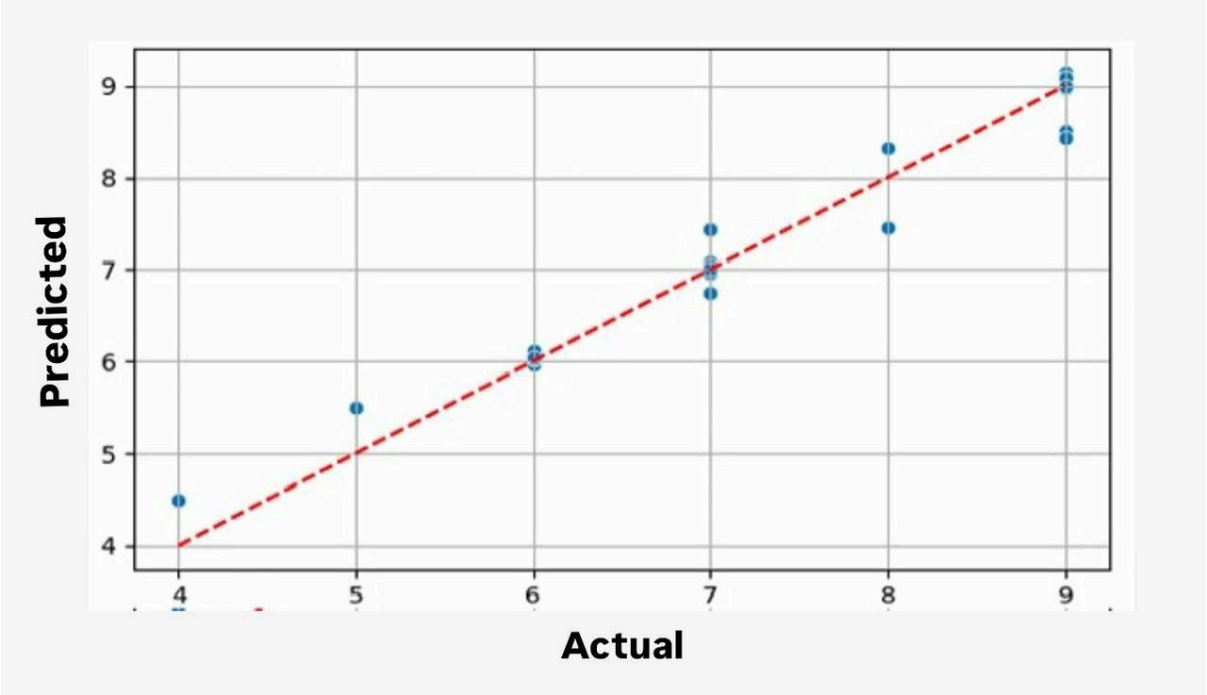
VISUALISATION

Scatter plots showing the actual versus predicted values for the best-performing model (XGBoost) indicate that the model is able to predict sleep quality with high accuracy, with the predicted values closely following the actual values.

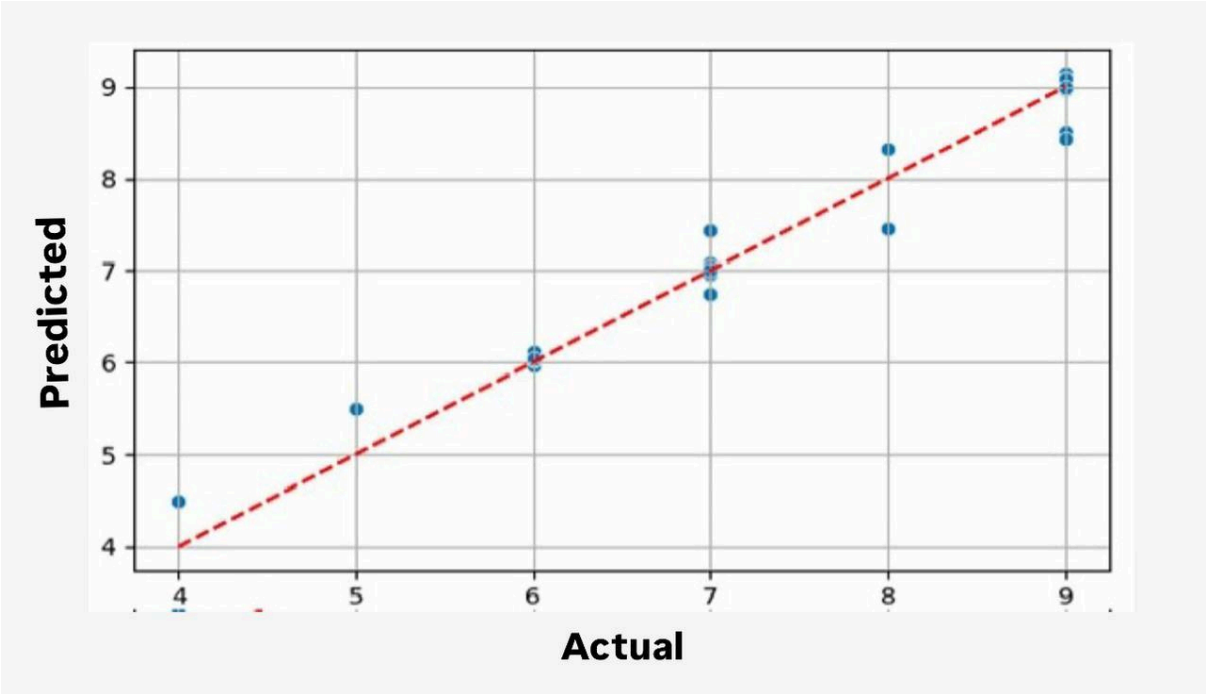
LINEAR REGRESSION



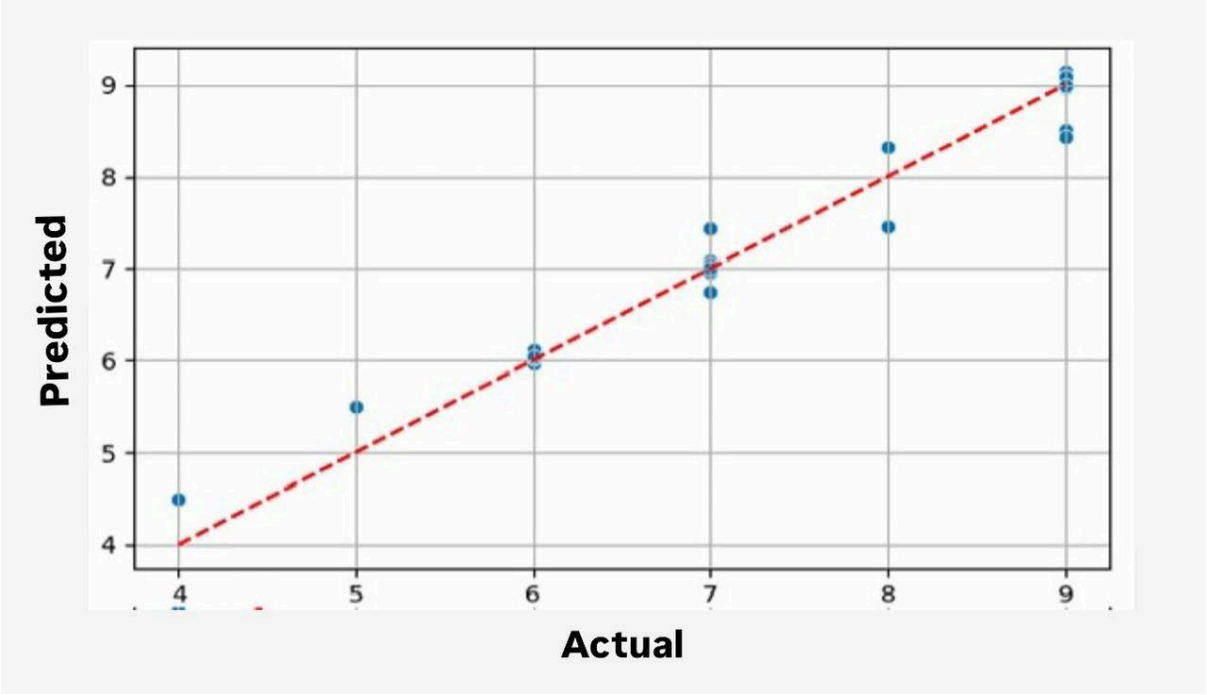
DECISION TREE REGRESSION



RANDOM FOREST REGRESSION



GRADIENT BOOST REGRESSION



CONCLUSION

This project successfully demonstrates the application of machine learning techniques in predicting used car prices based on various key features. By analyzing historical car

data and employing regression-based algorithms, the developed model provides accurate and consistent price estimations. This approach reduces the subjectivity and guesswork typically involved in valuing used cars, offering a data-driven solution that benefits both buyers and sellers.

The model enhances decision-making in the used car market by increasing pricing transparency and efficiency. While the current model performs well, its accuracy can be further improved by incorporating more real-world data, additional features such as service history or accident reports, and advanced algorithms like deep learning. Overall, this project highlights the potential of machine learning in transforming traditional pricing methods in the automotive industry.

REFERENCES

REFERENCES [1] Smith, J., & Zhang, L. (2021). Predicting used car prices with machine learning techniques. International Journal of Data Science and Analytics,14(3), 245-260.

[2] Kumar, R., & Patel, V. (2020). A comparative study of machine learning models for automobile price estimation. IEEE Access, 8, 150369-150382.

[3] Lee, H., & Kim, S. (2022). Feature selection and hyperparameter tuning in used car price prediction. Expert Systems with Applications, 195, 116456.

[4] Wang, Y., & Li, X. (2021). Deep neural networks for car price forecasting: A case study. Neural Computing and Applications, 33(15), 9123-9138.

[5] Garcia, M., et al. (2023). Improving car price prediction with synthetic data augmentation. Journal of Artificial Intelligence Research, 76, 145-167

[6] Chen, T., et al. (2022). AutoPriceML: A deployable framework for dynamic car valuation. Applied Soft Computing, 128, 109876.

[7[Müller, A., & Schmidt, F. (2023). A hybrid random forest and gradient boosting model for used car price prediction. Machine Learning with Applications, 12, 100487.

[8] Oliveira, P., et al. (2022). Interpretable machine learning for car price transparency: A SHAP-based analysis. Decision Support Systems, 162, 114763.

[9] Singh, R., & Zhang, W. (2023). Lightweight ML models for IoT-enabled car valuation in edge devices. IEEE Internet of Things Journal, 10(8), 6892-6905.

[10] Tanaka, H., et al. (2023). Cross-country transfer learning for car price prediction. Pattern Recognition Letters, 174, 15-23.