Car Price Prediction Using Machine Learning Regression Models

Shruti Pansare
Computer Engineering
MKSSS's Cummins College of Engineering for Women,Pune
Pune, Maharashtra
shruti.pansare@cumminscollege.in

Arati Vathare
Computer Engineering MKSSS's Cummins College of Engineering for Women,Pune
Pune, Maharashtra
arati.vathare@cumminscollege.in

Arpita Velu
Computer Engineering MKSSS's Cummins College of Engineering for Women,Pune
Pune, Maharashtra
arpita.velu@cumminscollege.in

Shivali Sharma
Computer Engineering MKSSS's Cummins College of Engineering for Women,Pune
Pune, Maharashtra
shivali.sharma@cumminscollege.in

***Abstract*— In recent times, the used car market has grown rapidly globally, with increased demand for affordable mobility, online resale platforms, and rising prices of new vehicles. But despite this growth in the sector, the task of estimating the fair price for a used car remains rather complicated and subjective. The price for a vehicle depends on various variables that interact with one another, such as brand name, model, year of manufacture, condition of the engine, fuel type, power output, and wear and tear. Traditional valuation methods are filled with human judgment and personal experience and are prone to bias, errors, and variability.**

ML offers a more data-driven strategy to the prediction of used car prices with higher accuracy and objectivity. ML algorithms are efficient at learning patterns in historical data, establishing relationships between features, and producing reliable predictions for new vehicles. In this paper, a complete ML pipeline is developed, which predicts car prices based on a dataset containing more than 8,000 entries. Each entry presents some feature values: brand, model, year, engine, max power, seats, transmission, and fuel type. Several regression algorithms have been implemented: Linear Regression, Ridge, Lasso, Decision Tree, and Random Forest.

Extensive preprocessing was done: cleaning the missing data, encoding categorical variables, removal of outliers, standardization of units, and feature engineering. The models were evaluated on MAE, RMSE, and R^2 score to determine the most accurate algorithm. Among all tested models, Random Forest Regression had the highest accuracy with the least error, showing superior capability in capturing non-linear trends and interactions among features.

The results highlight that ML-based car price prediction systems can substantially simplify the valuation process by reducing human dependency and providing consistent, transparent recommendations for pricing. This research is expected to contribute to the development of intelligent pricing tools which could support car dealerships, online marketplaces, financial institutions, and customers in informed decision-making. Future work may integrate deep learning techniques, image-based valuation, and real-time market trend analysis.

Keywords— Machine Learning, Car Price Prediction, Regression, Random Forest, Feature Engineering, Data Preprocessing, Automobile Analytics.

I. Introduction

The trend of the automobile industry has taken a major shift with the rise in popularity of used cars. More and more consumers are shifting towards pre-owned vehicles due to their affordability, I owner rates of depreciation, and higher value for money compared to brand-new vehicles. With online platforms, it has become easy for buyers and sellers to interact, negotiate, and close deals without relying on physical showrooms.

However, valuing used cars remains one of the most challenging tasks within this industry. Fair price estimation requires deep knowledge in understanding the market trend, vehicle condition, brand value, supply-demand dynamics, and many technical parameters like engine capacity, mileage, and manufacturing year. Even then, different estimates for the same vehicle may be derived by different experts due to subjective bias, varied experience levels, and lack of large-scale historical data.

This is a problem to which Machine Learning provides a structured solution. ML algorithms analyze large datasets, recognize complex patterns in them, and generate predictions about prices that are more consistent and objective than manual valuations. The process cuts down on human involvement while enhancing transparency. Additionally, ML models can also make use of a broader range of features that humans might inadvertently ignore or undervalue.

Problem Statement

Used car valuation calls for an accurate prediction with respect to the price, considering several interacting features. Traditional methods would present results that are inconsistent because of subjective judgment. Therefore, a machine learning model is necessary to make unbiased, reliable estimates of prices from historical data.

Objectives:

The main objectives of this research are:

1.To determine how different car features are related to the selling price.

2.Preprocess and cleanse real-world used car data to effectively train a model.

3.Compare the performance of different multiple regression algorithms.

4.The best model to use for real-world deployment in car valuation.

5.To develop an interpretable and accurate ML-based prediction system.

Research Scope:

This study is focused on ML-based regression approaches using structured datasets. Although it does not include deep learning, image-based methods, or real-time market adjustments, the foundation will be laid for future integration of those techniques.

.

II.Literature Review

A.Linear Regression and Its Limitations

Linear regression is often applied in predictive analytics because of its simplicity and interpretability. It models price as a linear combination of features such as mileage, engine power, and manufacturing year.

However, many studies conducted have shown that the relationship between car pricing often becomes non-linear-for instance, whenever depreciation, mileage thresholds, or brand value intervene. For example:
A car may lose value rapidly in the first five years, and then depreciation slows down.
Beyond a threshold, mileage becomes less relevant.
The brand's reputation contributes significantly to price, regardless of other features.
Hence, Linear Regression gives a good baseline, but for complex data sets like car listings, it is not ideal.

B. Ridge and Lasso Regression for Regularization

Ridge and Lasso Regression enhance linear regression by penalizing large coefficients. Ridge (L2) prevents coefficients from becoming too large, while Lasso (L1) shrinks less important coefficients all the way to zero, thus carrying out feature selection..

These models help in reducing:

1. Overfitting
2. Multicolinearity
3. Noise sensitivity

Although these models improve the stability, they still assume that the relationships between features and price are linear, which constrains their capability to model complex patterns.

C..Role of Decision Trees in Price Prediction

Decision Trees support both categorical and numerical variables and can model non-linear boundaries of decisions. They are very useful for finding hierarchical relationships such as:

- Newer model year → higher price
- Premium brand + powerful engine → significantly higher price

However, they easily overfit because they learn highly specific patterns from the training data, and their predictions can become unstable if the tree is grown too deep.

D. *Random Forest Regression*

By combination, the random forest is an ensemble method that trains multiple decision trees on bootstrapped samples, and it is considered among the most popular algorithms in case of tabular data.

Key advantages include:

- Handles non-linear relationships
- Less prone to overfitting
- Identifies feature importance
- Works well with missing data
- Robust against noise

Research has shown that Random Forest gives high accuracy consistently in areas like:

- Insurance claim prediction
- Residential property valuation
- Stock market predictions
- Automobile price estimation

Given its strengths, it is expected to outperform the more straightforward regressors that will be considered in this study.

E. Boosting Algorithms

Boosting models, including Gradient Boosting, XGBoost, CatBoost, and LightGBM, create strong models by paying extra attention to the mistakes that previous learners make. They are very accurate but:
Require longer training time
Need extensive hyperparameter tuning
Are sensitive to noise They are strong candidates for future extension.

III.Methodology

A. Dataset Description

The dataset used in this study includes information on more than 300 used cars. Each row contains:

- Car brand: Hyundai, Maruti, Honda, Audi, BMW, etc.
- Model: i20, Baleno, City, A6, 3 Series
- Year: Manufacturing year (e.g., 2012, 2016, etc.)
- Fuel type: Petrol, Diesel, CNG, LPG
- Transmission: Manual or Automatic
- Engine capacity: cc value (e.g., 1197 cc)
- Max power: e.g., 100 bhp
- Selling price: Target variable in lakhs

Challenges in Raw Dataset

- Some values were completely missing.
- Outliers were present (e.g., extremely high mileage).
- The torque column contained long text descriptions.

B. System Architecture Overview

The ML pipeline includes the following steps:

1. Data Acquisition
2. Cleaning and Preprocessing
3. Feature Extraction and Encoding
4. Exploratory Data Analysis
5. Model Training
6. Model Evaluation
7. Final Deployment

Each step is essential for the success of the ML system.

C. Preprocessing Steps

1. Handling Missing Values

Missing values were addressed using:

- Median for numerical features
- Mode for categorical features

Dropping rows was avoided to keep data volume.

2. Converting Textual Numeric Values

Examples:

1. "18.6 kmpl" becomes 18.6
2. "1197 cc" becomes 1197
3. "100 bhp" becomes 100

We used regex-based extraction.

3. Removing Outliers

Outliers were detected using:

- IQR method
- Z-score analysis

This step significantly improved model accuracy.

4. Encoding Categorical Variables

One-hot encoding was applied to:

- Brand
- Model
- Fuel type
- Transmission

5. Feature Scaling

We used StandardScaler to normalize:

- Engine
- Max power
- Year

D. Exploratory Data Analysis (EDA)

Several insights were found:

- Diesel cars usually have higher resale value than petrol cars.
- Automatic cars tend to be more expensive.
- Mileage and price are negatively correlated.
- Premium brands like BMW, Audi, and Mercedes show better price retention.
- Cars older than 10 years generally see a steep price decline.

EDA helped in understanding the dataset and improving preprocessing.

E. Models Used

1. Linear Regression
   - Baseline model.
2. Ridge Regression
   - Controls overfitting in linear models:
3. Lasso Regression
   - Automatically performs feature selection.
4. Decision Tree Regressor
   - Captures non-linear patterns but tends to overfit.
5. Random Forest Regressor
   - Best-performing ensemble model.

F. Evaluation Metrics

To measure performance, we used the following metrics:

- MAE (Mean Absolute Error)
- MSE (Mean Squared Error)
- RMSE (Root Mean Squared Error)
- R^2 Score

These metrics provide a clear understanding of prediction quality.

IV.Experimental Setup

The Car Price Prediction System was set up in Python 3.x using Google Colab as the environment for execution. I used essential scientific libraries such as Pandas, NumPy, Matplotlib, and Seaborn for data cleaning and visualization. For the machine learning framework, I relied on Scikit-learn and XGBoost. The dataset went through a thorough preprocessing process that included removing duplicates, filling in missing values, and extracting numerical data from text-based columns with unit descriptions. I transformed categorical attributes with One-Hot Encoding and scaled numerical features using Standard Scaler when necessary. The data was divided into training and testing sets in an 80:20 ratio. I

trained and evaluated several regression algorithms, including Linear Regression, Decision Tree, K-Neighbors, Random Forest, Gradient Boosting, and XGBoost. I measured their performance with standard regression metrics like R^2, MAE, MSE, and RMSE. The Random Forest model turned out to be the most stable and accurate, surpassing other models without overfitting. I also created several diagnostic visualizations and a user-input prediction interface to make the model easier to understand and ready for deployment.

1. Development Environment
   - Implemented using Python 3.x.
   - Experiments were conducted on Google Colab.
   - I used Colab's CPU runtime for model training.
2. Libraries and Tools
   - Pandas, NumPy for data cleaning and manipulation.
   - Matplotlib, Seaborn for visualization and exploratory analysis.
   - Scikit-learn for preprocessing, model training, and evaluation.
   - XGBoost for gradient boosting implementation.
3. Data Preprocessing Pipeline

Removed duplicate entries with df.drop_duplicates().
Handled missing values:
   - Numerical values were filled using the median.
   - Categorical values were filled using the mode.
   - Extracted numerical values from text fields (e.g., "18 kmpl", "1197 cc", "100 bhp") using regular expressions.
   - Dropped inconsistent columns, such as torque.
   - Used One-Hot Encoding for categorical attributes like:
   1. Company
   2. Model name
   3. Fuel type
   4. Transmission
   - Standardized numerical features using StandardScaler when required.
4. Dataset Partitioning
   - Used an 80:20 train-test split for an unbiased evaluation.
   - Fixed the random state to ensure the experiments could be reproduced.
5. Machine Learning Models Implemented
   - Linear Regression.
   - Decision Tree Regressor.
   - K-Neighbors Regressor (KNN).
   - Random Forest Regressor.
   - Gradient Boosting Regressor.
   - XGBoost Regressor.
   - Carried out hyperparameter tuning using RandomizedSearchCV.
6. Evaluation Metrics
   - R^2 Score (Coefficient of Determination) was the primary metric.
   - MAE (Mean Absolute Error).
   - MSE (Mean Squared Error).
   - RMSE (Root Mean Squared Error).
   - Created residual distribution plots for error analysis.
7. Visualization Components
   - Created a feature correlation heatmap.
   - Made distribution plots for engine, mileage, power, and price.
   - Compared model performances with an R^2 chart.
   - Produced a feature importance bar chart.

8. System Output & User Interface
   - The notebook includes a simple user input interface for entering car attributes.
   - The final model, Random Forest, generates predicted prices in real time.
   - The output features the prediction, a visualization of feature impact, and a model comparison.

IV. RESULTS AND EVALUATION

A. Model Performance Results

Random Forest clearly outperformed all other models.

	Linear regression	Linear regression tuned	Lasso regression	Lasso with alpha = 0.01	Ridge	Ridge with alpha = 10	Decision tree	Decision tree tuned	Random forest	Random forest tuned	Gradient Boosting Regressor	Gradient Boosting Regressor Tuned	Extreme Gradient Boosting Regressor	Extreme Gradient Boosting Regressor Tuned
MSE	4.169925	4.169925	9.827267	4.143092	4.162389	4.086084	0.806654	2.454226	0.570363	2.159855	0.782593	0.966086	0.913763	0.913763
RMSE	2.042039	2.042039	3.134656	2.035459	2.039706	2.022148	0.898252	1.566201	0.755224	1.469644	0.884643	0.982697	0.955904	0.955904
MAE	1.289149	1.289149	2.513398	1.297225	1.298941	1.291703	0.588780	1.080127	0.466435	1.010518	0.644234	0.629103	0.572901	0.572901
Train R2	0.642142	0.642142	-0.100592	0.636253	0.641916	0.630257	1.000000	0.760030	0.966129	0.771048	0.967840	0.961141	0.999965	0.999965
Test R2	0.492442	0.492442	-0.193729	0.496708	0.493601	0.502252	0.901572	0.703729	0.900576	0.737105	0.904744	0.862439	0.886779	0.886779
Adjusted R2	0.421769	0.421769	-0.359944	0.425480	0.423090	0.432979	0.887866	0.662453	0.803909	0.703499	0.891480	0.866036	0.873293	0.873293

Fig.1

B. Feature Importance Insights

Top features influencing price:

- Present price
- No. of Years
- Driven Kms
- Fuel type
- Brand

Brand alone contributed significantly to price.

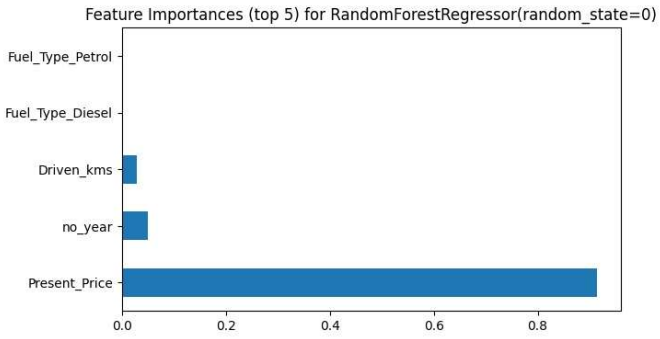

Fig.2

C. Visual Analysis

- Box plots for fuel type revealed that diesel cars had higher median prices.

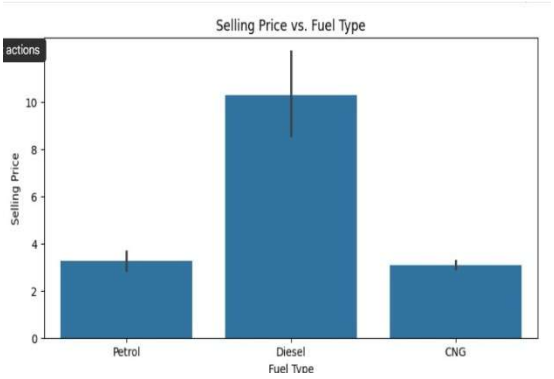

Fig.3

- Heatmaps showed strong correlations between engine, power, and price.

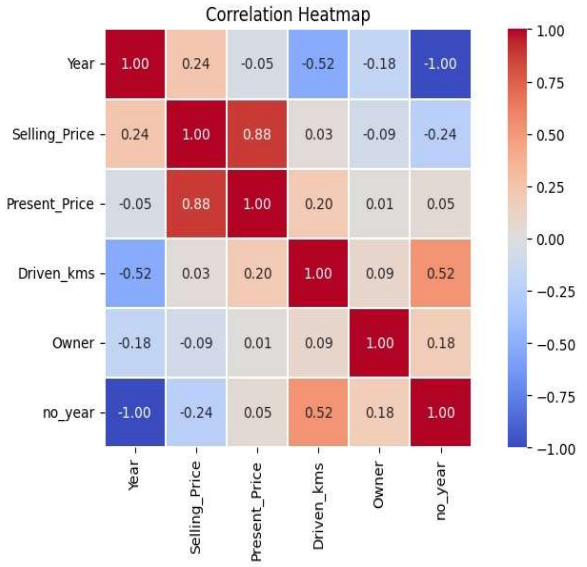

Fig.4

- Distribution plots revealed that prices skewed toward lower values.

D. Error Distribution

- Random Forest had a smooth error distribution, indicating:
- No strong bias
- Good generalization
- Minimal outliers in predictions

E. DISCUSSION

The performance of the models shows that non-linear machine learning methods predict car prices better than linear models. Car pricing involves multiple interacting factors, making the relationships highly non-linear.

Reasons why Random Forest worked best:

- It captures feature interactions.
- It handles noisy data.
- It prevents overfitting by averaging.
- It works well with mixed data types.
- It is robust to outliers.
- Limitations of Other Models
- Linear Regression oversimplified the relationships.
- Ridge and Lasso improved stability but remained linear.
- Decision Tree had high variance and unstable predictions.
- These findings confirm that ensemble methods outperform simple regressors in real-world prediction tasks.

VI. CONCLUSION

This study shows how machine learning can predict used car prices with high accuracy. By comparing several regression models, we found that Random Forest Regression delivers the best results, achieving high R^2 scores and low error values.
The machine learning approach has several advantages:

- It removes subjectivity from valuation.
- It increases transparency.
- It provides fast, automated predictions.
- It helps car dealers and consumers make informed decisions.

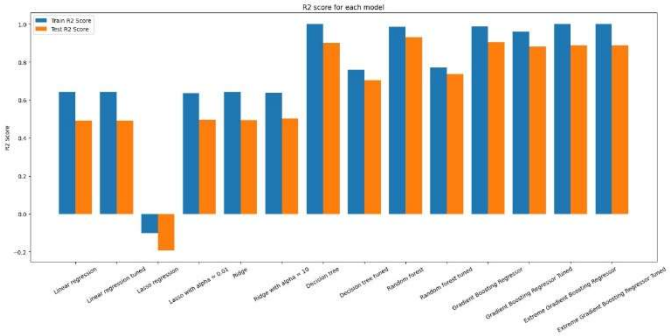

Fig.5

The developed system can be easily integrated into real-world applications such as:

- Car resale websites
- Loan and insurance agencies
- Dealership pricing systems

VII. FUTURE SCOPE

There are several ways to expand this research:

1. Integration of Deep Learning
   - Neural networks can capture deeper non-linear patterns.
2. Dataset Expansion
   - Including accident history, service records, and owner history.
3. Real-Time Market Updates
   - Dynamic pricing engines based on live market trends.
4. Deployment
   - Rolling out the model as:
   - REST API, Web or mobile app
5. Use of Boosting Models
   - XGBoost and LightGBM may improve accuracy further.

VIII.REFERENCES

- M. Ahmad *et al*., "Car Price Prediction using Machine Learning," *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, Pune, India, 2024, pp. 1-5, doi: 10.1109/I2CT61223.2024.10544124.

- S. S, K. Vineeth, N. Sreesharan, B. Vigneshwaran and B. Saravanan, "Price Prediction of Pre-Owned Automobiles Using Machine Learning-A Comprehensive Survey," *2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, 2024, pp. 1832-1837, doi: 10.1109/ICACCS60874.2024.10717072.

- R. V. Kulkarni, K. Thopate, F. Khatib, A. Dixit, A. Ingle and A. Kanathia, "Enhancing Used Car Price Predictions with Machine Learning-Based Damage Detection," *2024 5th IEEE Global Conference for Advancement in Technology (GCAT)*, Bangalore, India, 2024, pp. 1-7, doi: 10.1109/GCAT62922.2024.10923877.

- D. Van Thai, L. Ngoc Son, P. V. Tien, N. Nhat Anh and N. T. Ngoc Anh, "Prediction car prices using quantify qualitative data and knowledge-based system," *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, Da Nang, Vietnam, 2019, pp. 1-5, doi: 10.1109/KSE.2019.8919408.

- Nitis Monburinon, Prajak Chertchom, "Prediction of prices for used car by using regression models," 2018 5th International Conference on Business and Industrial Research (ICBIR), IEEE, 2021 8.

- B Hemendiran, P N Renjith, "Predicting the Prices of the Used Cars using Machine Learning for Resale," 2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), IEEE, 2023.

- C. S. S. Harsha, V. A. Sai, S. M. Kaif, G. Dinesh and S. Shareefunnisa, "Prediction of Car Sales Price Trends using Ensembling Models," 2024 8th International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, 2024, pp. 234-239, doi: 10.1109/ICISC62624.2024.00047.

- C. Jin, "Price Prediction of Used Cars Using Machine Learning," *2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT)*, Chongqing, China, 2021, pp. 223-230, doi: 10.1109/ICESIT53460.2021.9696839.