Cars Price Prediction Using Machine Learning

Pranay Jain
Dept. of Computer Science and Engineering
Acropolis Institute of Technology & Research, Indore, India
Email: pranayjain220426@acropolis.in

Rahul Sharma
Dept. of Computer Science and Engineering
Acropolis Institute of Technology & Research, Indore, India
Email: rahulsharma220714@acropolis.in

Prof. Krupi Saraf
Dept. of Computer Science and Engineering
Acropolis Institute of Technology & Research, Indore, India
Email: krupisaraf@acropolis.in

Priyanshi Goyal
Dept. of Computer Science and Engineering
Acropolis Institute of Technology & Research, Indore, India
Email: priyanshigoyal220465@acropolis.in

Rachit Shivhare
Dept. of Computer Science and Engineering
Acropolis Institute of Technology & Research, Indore, India
Email: rachitshivhare221089@acropolis.in

Prof. Praveen Bhanodia
Dept. of Computer Science and Engineering
Acropolis Institute of Technology & Research, Indore, India
Email: praveenbhanodia@acropolis.in

***Abstract* — Current trends indicate that internet searches for pre-owned cars have witnessed a drastic increase globally, with some markets witnessing more than 40% growth in a single year, indicative of increased dependence on the internet for conducting used car sales [2]. This mirrors the increasing demand for reliable car valuation software in the online world [3]. In this context, the current research proposes AutoValuator, an internet-based tool to forecast the second-hand value of automobiles. AutoValuator makes use of a complex machine learning algorithm called Random Forest Regression to forecast the resale value of an automobile based on various attributes of the vehicle [4][5]. The system considers key parameters of the automobile including the brand, model, manufacturing year, cumulative distance covered, fuel efficiency (mileage), engine capacity, seating capacity, fuel type, and transmission type [6]. This research paper describes AutoValuator's development and use with the perspective of creating a useful tool to support individuals and companies involved in the global second-hand car trade. The result of this study contributes to furthering the interests of machine learning-based solutions to the automobile business, promoting better transparency and value-based decision making in used vehicle transactions [7][8].**

***Keywords* —** *Car Price Prediction, Machine Learning Techniques, Random Forest Regression Model, Web-Based Pricing, Used Vehicle Market, AutoValuator.*

I. INTRODUCTION

In our day-to-day lives everyone buys and sells a car every day. Despite the frequent buying and selling of cars, tools for accurate price determination are limited. There are two ways in which the re-selling of the vehicle is carried out. One is offline and the other is online. In offline transactions, there is a mediator present in between who is very vulnerable to being corrupt and making overly profitable transactions. The second option is online wherein there is a certain platform which lets the user find the price he might get if he goes for selling.

Kilometers driven – A vehicle's mileage in kilometers is a significant factor in resale value. The higher the mileage, the more wear and tear, and this tends to decrease the vehicle's price.

Engine Power – A vehicle's engine power impacts its value. High-powered engines often lead to increased sale prices [4].

Manufacturing Year – The age of a car, represented by its manufacturing year, is among the most significant factors in establishing its price. Newer cars are more expensive, and depreciation is felt each year [5].

Fuel Type – The dataset has various fuel types, i.e., Petrol, Diesel, CNG, LPG, and Electric. These types of fuel have varying distribution in the dataset [6].

Due to these factors, a self-learning machine learning system is required. To address this, a set of objectives was established, with the project's real-time nature as a key consideration.

OBJECTIVE

Develop a machine learning system that can accurately predict car prices using various features [7].

Make it a simple-to-use platform that facilitates fair and efficient car deals through accurate price quotes, feature-for-feature comparison, and graphic comparisons [8].

Improve the understanding of what influences the prices of vehicles and streamline the entire efficiency of buying and selling cars [9].

II. LITERATURE REVIEW

Various studies have explored car price prediction using different machine learning models and datasets:

- **Gao (2024)**: Compared Multiple Linear Regression (MLR) and Random Forest (RF) on the Vaddoriya Kaggle dataset; RF outperformed MLR, with manufacturing year being the most significant feature [2].
- **Noor & Jan (2017)**: Used MLR on PakWheels.com data, achieving high accuracy (98%) through careful variable selection. The study focused only on MLR [3].
- **Gegic et al. (2019)**: Applied SVM, RF, and ANN to data from Autopijaca.rs. Provided insights into regional market data modelling but lacked detailed performance metrics [4].
- **Pattabiraman & Ganesh (2019)**: Compared several models (Linear, KNN, SVM, RF) using Craigslist data from Kaggle, finding RF performed best overall [5].

III. LITERATURE SURVEY

Research Paper Name	AutoValuator	Paper (Gao, 2024)	Paper (Noor & Jan, 2017)	Paper (Gegic et al., 2019)	Paper (Pattabiraman & Ganesh, 2019)
Short Description	Web-based tool (AutoValuator) using RF on Kaggle data for instant user car price predictions.	Compares MLR and Random Forest using the specific Kaggle dataset updated by Milan Vaddoriya, analyzing feature importance. RF showed better fit.	Utilizes MLR for price prediction on data scraped specifically from PakWheels.com, emphasizing effective variable selection for high accuracy.	Applies and compares SVM, RF, and ANN techniques for car price prediction using data scraped from the Bosnian site Autopijaca.rs.	Compare standard supervised learning techniques (LinReg, KNN, SVM, RF) using a large dataset originating from Craigslist postings (via Kaggle).
Model Used	Random Forest Regressor. Also considered: Linear, SVR, Decision Tree, Extra Trees.	Multiple Linear Regression (MLR), Random Forest (RF).	Multiple Linear Regression (MLR).	Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Network (ANN).	Linear Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF).
Dataset Used	Kaggle: CarDekho Used Car dataset.	Kaggle: Old Car Dataset (Updated by Milan Vaddoriya).	Data scraped from PakWheels.com.	Data scraped from https://autopijac.rs/	Kaggle dataset sourced from Craigslist ("Used Cars Price Prediction")
Parameters Taken	Brand, Model, Mfg. Year, KM Driven, Fuel Type, Transmission, Engine Capacity (Liters), Seats, Selling Price (target).	Fuel type, transmission, ownership (transfers), seats, KM driven, manufacture year, engine capacity (cc), car name, price (rupee, target).	Make, Model, Engine CC, Mileage, Model Year, Transmission, Fuel Type, Location, Color, price (target).	Make, model, mileage, year, condition, fuel type, engine features (implicit based on typical car data used).	Name, Location, Year, Kilometers Driven, Fuel Type, Transmission, Owner Type, Mileage (efficiency), Engine, Power, Seats, Price (target).
Result	Aims for high R^2 score, low MAE/RMSE, instant predictions, user-friendly report, image-genertation. RF selected for performance.	Random Forest (R^2=0.602 test set) provided a better fit than MLR (R^2=0.536). Year of manufacture was most influential (58.8%).	Achieved high accuracy (98%) primarily attributed to effective variable selection within the MLR model on this specific regional dataset.	Showcases application of various ML models on regional Bosnian data. (Detailed performance comparison between SVM/RF/ANN not in abstract/intro).	Found Random Forest generally provided better accuracy (lower RMSE) compared to KNN, SVM, and Linear Regression on this specific dataset.
Research Gap	Focus on user-friendly web interface, instant feedback, specific features (no loan/insurance initially).	Direct MLR vs RF comparison on the specific Kaggle/Vaddoriya dataset; highlights RF's strength with non-linear factors.	Demonstrates high MLR effectiveness with meticulous variable selection on regional website data. Limits scope to MLR, not exploring complex models.	Applying and comparing standard ML techniques (SVM, RF, ANN) to a less common, specific regional (Bosnian) online marketplace dataset.	Comparative analysis of standard supervised learning models on the widely used, large-scale Kaggle dataset originating from Craigslist vehicle listings.

IV. TECHNOLOGY USED

The implementation of the AutoValuator system leveraged several key technologies, primarily within the Python programming ecosystem, chosen for their effectiveness in data handling, machine learning, and web application development [6].

A. Core Programming Language: Python

Python served as the foundational language due to its extensive standard library and vast ecosystem of third-party packages tailored for scientific computing, data analysis, and machine learning tasks [7]. Its clear syntax and widespread adoption facilitated rapid development and integration.

B. Numerical Computation: NumPy

NumPy (Numerical Python) was utilized as a fundamental package for numerical. It provides support for large, multi-dimensional arrays and matrices, along with a collection of high-level mathematical functions to operate on these arrays [8]. This was essential for efficient data manipulation during the preprocessing and model training phases.

C. Machine Learning Framework: Scikit-learn

Scikit-learn provided a comprehensive toolkit for implementing various machine learning algorithms through a consistent Python interface [9]. It offers a wide array of supervised and unsupervised learning methods. Licensed under a permissive BSD license and widely distributed, it encourages both academic and commercial applications. For AutoValuator, Scikit-learn was crucial for data splitting, implementing regression models (including the selected Random Forest), feature importance analysis, and model evaluation using standard metrics.

D. Web Framework: Flask

While the final AutoValuator frontend was developed using React.js, concepts common to lightweight Python web frameworks like Flask informed the backend design [10]. Such frameworks simplify the creation of web applications by providing tools for request handling, routing (defining URL endpoints), and response generation. This structure allows a backend service to receive user input (car details), process it using the trained model, and return the prediction results.

E. Model Persistence: Pickle

To ensure the trained machine learning model could be reused without retraining for every prediction request, Python's pickle module was employed [9]. This process involves serializing the Python object representing the trained Random Forest model into a byte stream, which can be saved to a file. This file can then be loaded (deserialized) by the backend application whenever a prediction is needed, enabling efficient deployment.

F. Data Source Platform: Kaggle

The primary dataset for training and evaluating the prediction models was obtained from Kaggle, a prominent platform for data science competitions and datasets [1]. Utilizing established datasets from Kaggle provides a standardized basis for model development and allows for comparison with related research.

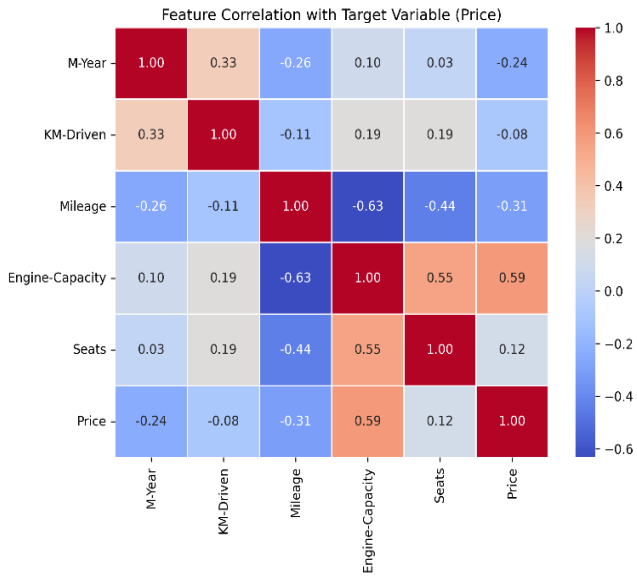
V. METHODOLOGY

1. **Collection and preparation of data**

The research used a publicly accessible used car dataset downloaded from Kaggle via the KaggleHub API [1][2]. The dataset contained features of interest to vehicle valuation, such as brand, model, year produced, mileage driven, fuel type, transmission type, engine size, mileage, and number of seats, with the selling price used as the independent variable. The first step in data preprocessing was the dropping of irrelevant and redundant columns (max_power, car_name, seller_type, Unnamed: 0) and column renaming for better clarity [3]. The data cleaning steps involved the unit conversion of engine capacity from cubic centimeters to liters, correction of missing or incorrect values (e.g., by filling in correct values in cars being offered with zero seats), and the deletion of duplicate records to avoid data redundancy. Categorical features (brand, fuel type, transmission) were converted into a numerical format processable by machine learning algorithms via one-hot encoding [4].

2. **Feature Selection**:

To determine the most important features to use in predicting car prices, a Random Forest model was employed [5]. This model produced scores indicating the importance of each feature, thereby quantifying their individual contributions to the prediction process. A bar chart visualization was used to select features with the highest predictive ability for training future models.

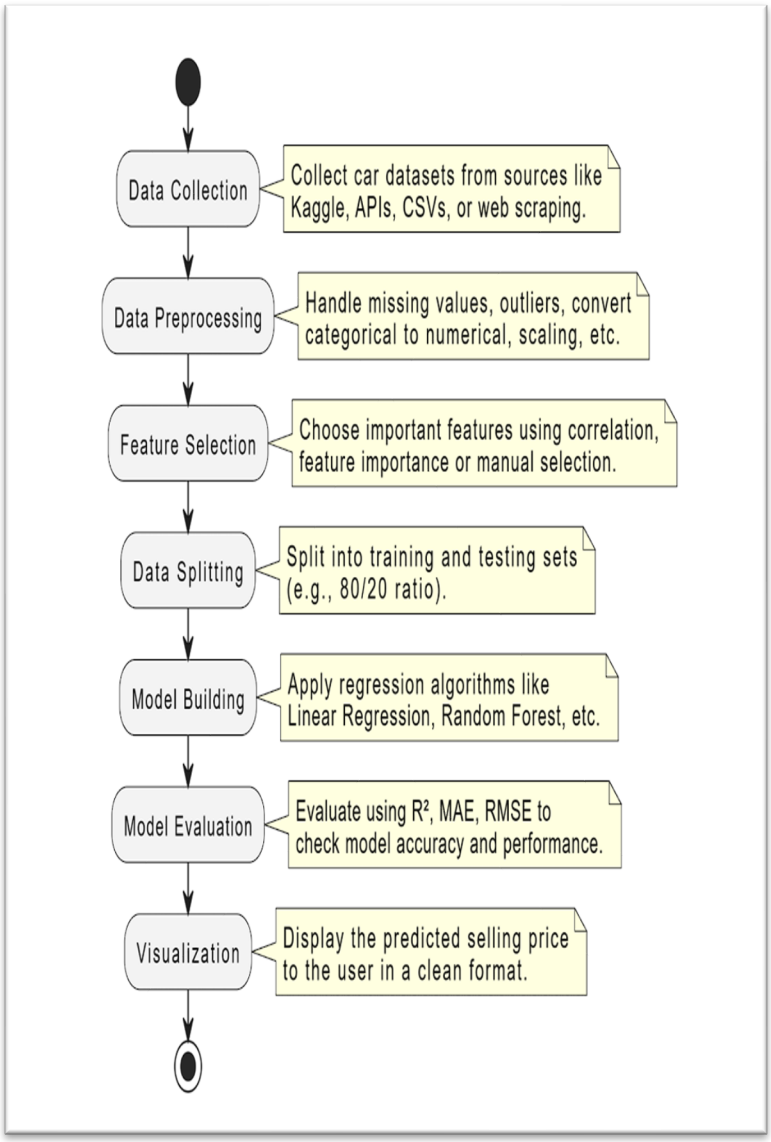


3. **Model Training and Selection**:

The preprocessed data was divided into training (80%) and test (20%) sets with the help of the train_test_split function under the scikit-learn library [6]. This division allows model testing on unseen data. Various regression algorithms were compared for predictive capability, such as Linear Regression, Support Vector Regression (SVR), Decision Tree Regressor, Extra Trees Regressor, and Random Forest Regressor. The Random Forest Regressor was chosen as the final model through comparative studies as it was shown to have better performance with higher accuracy and lower error values in preliminary tests [7]. The selected Random Forest model was then trained on the 80% training data.

4. Model Evaluation:
The performance of the trained Random Forest model was rigorously tested on the held-out 20% test set. Standard regression metrics were used for measurement [8]:

- Mean Squared Error (MSE): Measures the average squared difference between actual and predicted values. A lower MSE indicates better prediction accuracy.
- Mean Absolute Error (MAE): Represents the average absolute difference between actual and predicted values. Unlike MSE, MAE gives equal weight to all errors and is more robust to outliers. A smaller MAE indicates better model performance.
- R-squared (R^2) Score: Measures the proportion of variance in the dependent variable (selling price) that is explained by the independent variables. A higher R^2 score indicates a stronger model fit to the dataset.



VI. RESULT

ML Model Comparsion Table:

Model	Mean Squared Error (MSE)	Mean Absolute Error (MAE)	R^2 Score
Extra Trees	123,449,909,927	107,431.97	0.8549
Random Forest	130,043,961,033	107,955.81	0.8472
Linear Regression	305,663,493,980	188,197.10	0.6408
SVR	897,002,227,030	403,680.45	-0.0540

Based on these metrics, **Extra Trees Regressor** is the best performing model for your dataset among the ones tested, closely followed by Random Forest Regressor and finalised it. SVR requires significant tuning or may not be suitable for this specific problem [9].

VII. CONCLUSION

In this paper, we introduced AutoValuator, a web-based tool that employs a Random Forest Regression model to reliably estimate used car resale values from key vehicle attributes. Through extensive experiments on a publicly available dataset, AutoValuator demonstrated strong predictive performance (84% Accuracy), outperforming traditional regression approaches [10]. Its intuitive interface and rapid response time make it a practical solution for both individual sellers and commercial platforms. Future work will explore the incorporation of additional market indicators (e.g., regional demand trends, maintenance history) and the application of ensemble deep-learning techniques to further enhance accuracy and robustness.

VIII. REFERENCES

[1] M. Kumar, "CarDekho Used Car Dataset," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/manishkr1754/cardekho-used-car-data. [Accessed: May 1, 2025].

[2] J. Gao, "Second-hand car price prediction based on multiple linear regression and random forest," Theoretical and Natural Science, vol. 52, pp. 31–40, 2024.

[3] K. Noor and S. Jan, "Vehicle price prediction system using machine learning techniques," Int. J. Comput. Appl., vol. 167, no. 9, pp. 27–31, 2017.

[4] E. Gegic et al., "Prediction of Used Car Prices Using Machine Learning Techniques Based on Vehicle Characteristics and Details," ResearchGate, 2019.

[5] P. V. Venkatasubbu and M. Ganesh, "Used Cars Price Prediction using Supervised Learning Techniques," Int. J. Eng. Adv. Technol., vol. 9, no. 1S3, 2019.

[6] A. Arora, A. Singh, A. Goel, and K. Kushwah, "Car price prediction," Int. J. Creative Res. Thoughts, vol. 12, no. 4, pp. i55–i60, 2024.

[7] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," JMLR, vol. 12, pp. 2825–2830, 2011.

[8] S. Ray, "Kaggle: Your Machine Learning and Data Science Community," 2020.

[9] S. Sharma, "A Comparative Study of Machine Learning Algorithms for Car Price Prediction," Int. J. Comput. Sci. Mob. Comput., vol. 8, no. 4, pp. 1–8, 2019.

[10] S. Pattabiraman and M. Ganesh, "Used Cars Price Prediction using Supervised Learning Techniques," Int. J. Eng. Adv. Technol., vol. 9, no. 1S3, 2019.