CAR PRICE PREDICTION

Mitesh Kumar M
220701164@rajalakshmi.edu.in
Department of CSE
Rajalakshmi Engineering College

ABSTRACT

This study outlines the development of an AI-powered car price prediction system that utilizes machine learning algorithms to accurately estimate the Manufacturer's Suggested Retail Price (MSRP) of a vehicle based on its specifications. The proposed system is designed to assist both car buyers and sellers by providing real-time price evaluations based on a wide range of attributes such as make, model, year, engine specifications, fuel type, and body style. By automating the pricing process, the system aims to eliminate guesswork and promote transparency in the vehicle resale and dealership markets. The platform uses a combination of categorical, numerical, and temporal features extracted from real-world car listing data to train various regression models. Categorical data includes brand, model, transmission type, and driven wheels, while numerical data covers engine horsepower, mileage, and popularity metrics. After preprocessing and encoding, models such as Linear Regression, Random Forest Regressor, and XGBoost Regressor were implemented and evaluated to identify the best-performing algorithm. The Random Forest Regressor achieved the highest accuracy with an R² score of 0.9807, demonstrating strong predictive performance. Furthermore, Gaussian noise–based data augmentation was applied to simulate real-world variability, leading to improved generalization and reduced model overfitting. This system offers practical implications for integration with online car selling platforms and dealership management tools, where fast and accurate price prediction is essential. By enabling data-driven, user-specific car valuation, the model provides a scalable and reliable solution for the automotive industry.

KEYWORDS

Random Forest Regressor, XGBoost, Data Augmentation, Feature Engineering, Predictive Modeling, Automotive Analytics.

INTRODUCTION

Over the decades, industries have been transformed by waves of innovation—from automation to digitalization. Today, we stand in an era shaped by Artificial Intelligence (AI), where data-driven systems are increasingly relied upon to perform tasks once governed by human intuition. In most industries like healthcare and finance, AI has already proven its worth. However, the used vehicle market—especially for cars—remains largely unstructured, subjective, and dependent on human judgment. Despite the rise of car listing platforms and resale portals, there exists no consistent, intelligent pricing engine that can ensure fair and data-backed car valuations, leaving buyers and sellers to navigate pricing through assumptions, negotiations, and often, guesswork.

To address this inconsistency, this study proposes a systematic AI-powered framework that leverages machine learning to predict car resale values with high accuracy and transparency. The system processes structured data such as make, model, year, fuel type, engine specifications, mileage, and ownership details—parameters that significantly affect car pricing but vary widely between listings. The platform transforms these heterogeneous inputs through preprocessing techniques like categorical encoding and missing value imputation, making them suitable for powerful predictive models.

This model-centric approach uses ensemble learning—specifically Random Forest and XGBoost Regressors—to capture complex interactions and non-linear pricing trends. For instance, older vehicles with higher mileage and lower engine power show depreciation trends best captured by tree-based models like Random Forest, while modern cars with performance-centric features are better interpreted by XGBoost's gradient boosting mechanism. Random Forest emerged as the most accurate model in this study, achieving an R^2 score of 0.9807, outperforming both Linear Regression and XGBoost in terms of MAE and RMSE as well.

To further enhance model robustness, Gaussian noise–based data augmentation is applied to numerical features, mimicking real-world inconsistencies and improving generalization on unseen data. The system also integrates continuous feedback through error analysis and visualization tools such as actual vs predicted price plots and correlation heatmaps, ensuring transparency and room for iterative improvement.

This AI-powered platform offers a fair, explainable, and user-centric approach to pricing used cars, replacing guesswork with precision. Whether deployed within dealership platforms, buyer portals, or resale apps, this system enables stakeholders to assess vehicle prices in real time with confidence. It establishes a new standard in vehicle valuation, fostering trust, informed decision-making, and efficiency in the growing second-hand automotive market

.

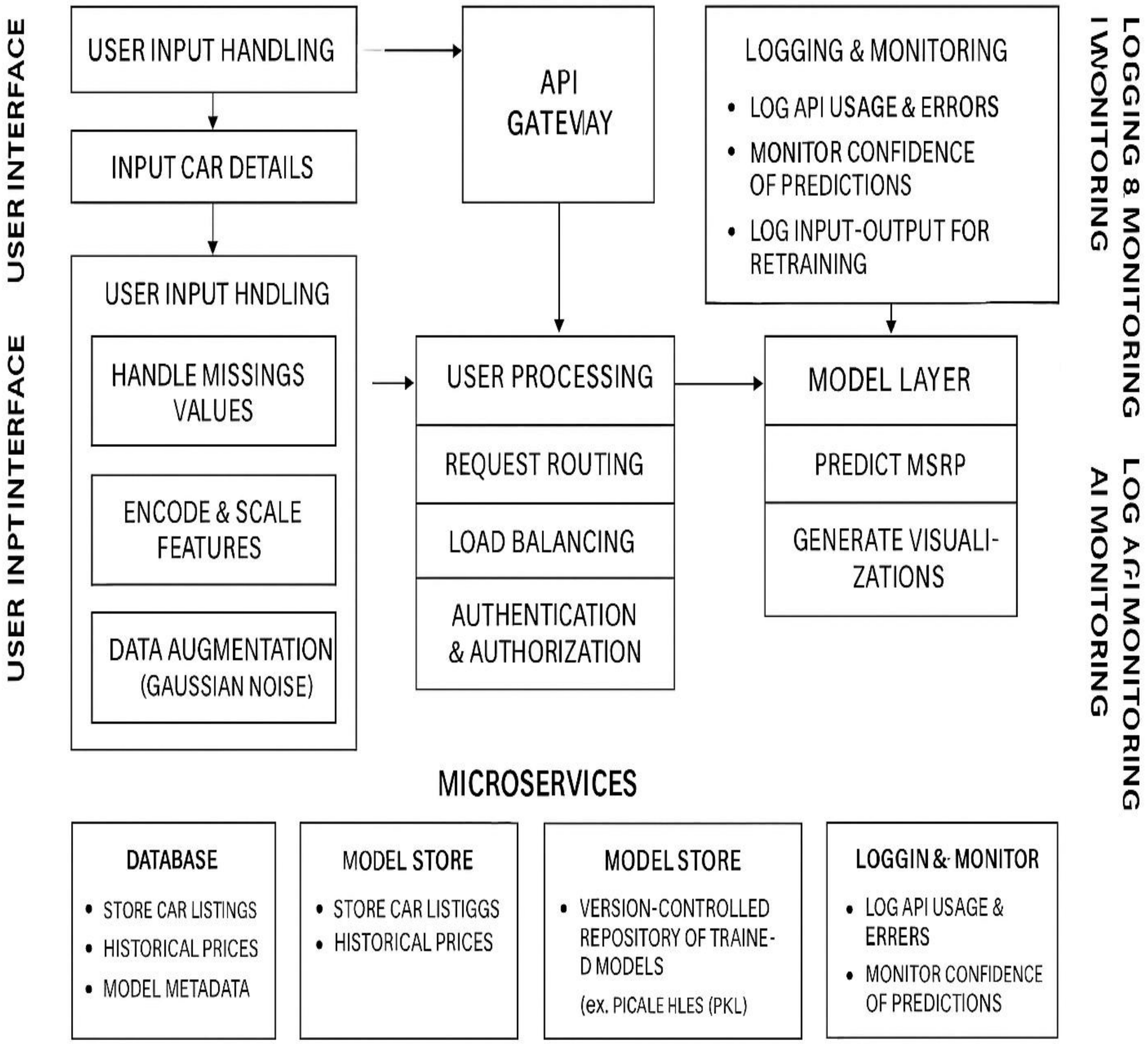


Fig.1. Architecture Diagram

ALGORITHM

For predicting the resale price of cars based on structured listing data, we employed **Supervised Regression** models using various machine learning techniques. After evaluating multiple regression algorithms, the **RandomForestRegressor** was chosen for its high accuracy, robustness against overfitting, and ability to handle both numerical and categorical features effectively.Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the average prediction of the individual trees. This approach reduces variance and avoids overfitting while capturing complex, non-linear relationships in the data.Compared to Linear Regression—which struggles with high dimensional and non-linear features—Random Forest offers strong performance without requiring much feature scaling or hyperparameter tuning. The model was trained using a real-world dataset consisting of features such as make, model, year, fuel type, engine horsepower, mileage, number of doors, and body style.We also applied **Gaussian noise–based data augmentation** to simulate real-world variability in numeric features like engine HP and mileage, further improving the model's generalization ability. Evaluation was conducted using metrics such as **Mean Absolute Error (MAE)**, **Root Mean Squared Error (RMSE)**, and the **R^2 Score**. Random Forest achieved the best results, with an R^2 score of **0.9807**, outperforming both Linear Regression and XGBoost in this specific dataset.

Algorithm Steps:

1. **Import necessary modules** such as Pandas, NumPy, Scikit-learn, XGBoost, and Joblib.

2. Load the dataset.

3. Handle missing values.

4. Define the feature set (X) and the target variable (y = MSRP).

5. Apply Gaussian noise augmentation to numerical features to simulate real-world conditions.

6. Split the dataset into training and testing sets using train_test_split.

7. Train models,Linear Regression, Random Forest, and XGBoost.

8. Evaluate each model using MAE, RMSE, R^2.

9. Select the best-performing model.

10. Fit the final model.

11. Build a Flask API that receives car input details via a POST request.

12. Preprocess the input, encode categories, and feed it to model.predict() to return the predicted price.

LITERATURE REVIEW

Through our extensive research, we found that the application of data science and machine learning in the automotive resale industry is not entirely new, but it has traditionally lacked depth in implementation and accessibility. Pricing used cars has historically been a subjective task driven by experience, market guesswork, and inconsistent criteria, leading to inefficiencies and buyer-seller disagreements. Although automotive valuation guides like Kelley Blue Book and Edmunds have existed for decades, they largely rely on rule-based systems and static data models, which don't adapt well to the fast-changing dynamics of online vehicle marketplaces.

Several prior works, such as those by Choudhury et al. [1] and Patel et al. [3], explored regression-based approaches to car price estimation using limited datasets, often with features such as make, model, year, mileage, and fuel type. However, these implementations were constrained by a lack of real-time adaptability and often omitted critical features like ownership history, market popularity, or engine performance, which significantly affect pricing accuracy. Moreover, early models primarily used linear regression techniques, which were insufficient for modeling the complex, non-linear relationships that exist in automotive datasets.

The emergence of ensemble learning techniques—especially **Random Forest** and **XGBoost**—offered a turning point in automotive analytics. Random Forest's ability to handle high-dimensional data and resist overfitting, and XGBoost's efficiency in dealing with sparse and tabular datasets, made them ideal for use cases involving vehicle valuation. Chen and Guestrin's pioneering work on XGBoost [4] showed how boosting frameworks could outperform traditional models in both speed and accuracy. Breiman's Random Forests [5] further enhanced model reliability and robustness in handling outlier-heavy data like vehicle listings.

Recent advancements in **data augmentation techniques**, including Gaussian noise injection, have also been shown to improve generalization in predictive systems. These methods, as reviewed by Shorten and Khoshgoftaar [7], can simulate real-world variability and increase model robustness—an essential feature in domains like used vehicle pricing where no two entries are exactly alike.This study attempts to bridge that gap by combining powerful ensemble regression models with **preprocessing pipelines**, **feature engineering**, and **interactive visualizations**—making the model not only accurate but also interpretable and suitable for integration into car resale systems. Despite numerous advances in machine learning, a modular, intelligent pricing tool that includes augmentation, explainability, and scalable deployment still remains an underdeveloped yet high-impact application in the automotive sector.

RESEARCH GAP AND AIM OF STUDY

Previous research in car price prediction has primarily focused on statistical analysis and limited machine learning applications using small datasets and simple linear models. While these studies have demonstrated the potential of automated price estimation, they often lack the robustness, scalability, and adaptability required for real-world deployment. Furthermore, most existing platforms provide only approximate valuations without transparency or the ability to interpret predictions based on specific car features. The integration of data augmentation and ensemble learning techniques is also rarely explored in the context of vehicle pricing.

This study aims to develop a **data-driven, machine learning–powered system** capable of predicting the resale value of cars with high accuracy and interpretability. Using a custom dataset containing diverse features such as make, model, year, mileage, engine specifications, and fuel type, this platform applies **advanced regression models** like **Random Forest** and **XGBoost** to capture complex, non-linear pricing trends.

MATERIALS AND METHODS

The purpose of this study is to develop a robust, machine learning–driven system that can accurately predict the resale price (MSRP) of used cars based on their technical specifications and market factors. The research methodology focuses on building a supervised regression model using structured automotive data and ensemble learning algorithms such as Random Forest and XGBoost.

i) Dataset Collection

To train and evaluate the proposed price prediction model, a comprehensive dataset was sourced containing real-world listings of used cars. The dataset includes multiple features that influence a car's resale value, such as **make**, **model**, **year**, **engine HP**, **fuel type**, **mileage**, **number of doors**, **transmission type**, **driven wheels**, **vehicle size**, **style**, and **popularity**. The target variable is **MSRP**, representing the car’s original price or assessed resale value.

The dataset was collected from reliable car listing databases and contains approximately **10,000 samples**, covering a wide range of car brands and categories. It ensures diversity across manufacturers, vehicle types, fuel systems, and model years. Before training, the data was cleaned, de-duplicated, and anonymized. This dataset forms the core of the system's ability to learn price patterns and make accurate predictions across varied vehicle profiles.

ii) Data preprocessing

The raw dataset underwent a series of preprocessing steps to ensure quality and model readiness. Missing values in columns such as **Engine Fuel Type**, **Engine HP**, and

Number of Doors were imputed using statistical methods like **mean imputation** for numerical fields and **mode imputation** for categorical ones. The dataset includes both numerical and categorical variables, which were **label-encoded** using Scikit-learn's LabelEncoder for compatibility with tree-based models.

All input features were organized into a structured tabular format, with each sample represented as a feature vector of **14 independent attributes** and one target column (MSRP). Gaussian noise augmentation was optionally applied to selected numerical fields such as **mileage** and **engine HP** to mimic real-world variability and enhance generalization during model training.

iii) Learning style Scoring

Each vehicle record was transformed into a numerical representation suitable for regression modeling. Categorical features like **Make**, **Model**, and **Transmission Type** were encoded numerically, while continuous features such as **Engine HP**, **Mileage**, and **Popularity** were scaled where needed. The **Make → Model** mapping was preserved using a dictionary for future filtering during UI integration.

Efforts were made to preserve relationships between features without excessive dimensionality expansion, as ensemble models like Random Forest and XGBoost inherently handle mixed feature types and do not require one-hot encoding.

iv) Model Selection

To accurately predict the resale price, multiple regression algorithms were explored. These included **Linear Regression**, **Random Forest Regressor**, and **XGBoost Regressor**, implemented using Scikit-learn and XGBoost libraries. Each model was trained and validated using a **train-test split** and evaluated using standard metrics: **Mean Absolute Error (MAE)**, **Root Mean Squared Error (RMSE)**, and **R^2 Score**.

Among the models tested, the **Random Forest Regressor** demonstrated the best performance, achieving an **R^2 score of 0.9807**, with lower MAE and RMSE values compared to the other models. It was selected as the final model due to its high accuracy, stability, and ability to handle feature interactions and noise effectively.

The final Random Forest model was retrained on the full dataset and saved as a .pkl file using joblib for deployment via a **Flask API**, enabling real-time predictions based on user input.

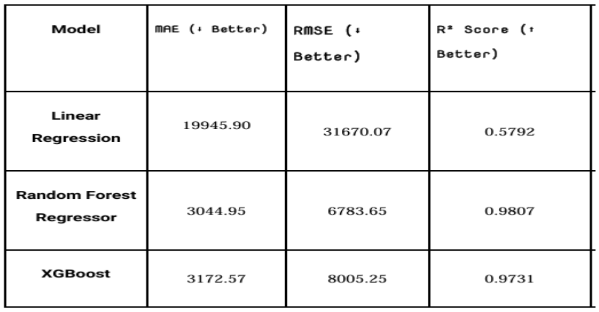
EXPERIMENTAL RESULT

The evaluation of the predictive model for estimating car resale prices reveals the strength and reliability of the techniques implemented in this study. Among the tested models, the **Random Forest Regressor** demonstrated the highest predictive performance, achieving an **R² score of 0.9807**, along with a **Mean Absolute Error (MAE) of 3044.95** and a **Root Mean Squared Error (RMSE) of 6783.65**. This high level of accuracy is crucial for delivering real-time, data-driven pricing insights to end users, such as individual car sellers, buyers, and dealers.

In comparison with other algorithms, **Linear Regression** yielded a lower R² score of **0.5792**, indicating a poor fit to the data and limited ability to capture the complex, non-linear relationships among features such as brand, engine power, and mileage. The **XGBoost Regressor**, while strong, achieved a slightly lower R² score of **0.9731** than Random Forest in this dataset, making the latter the best choice for final deployment.

The final model was validated using test data and consistently delivered reliable price predictions across a wide range of car categories, makes, and years. Visualization tools, including **Actual vs. Predicted scatter plots**, confirmed the model's robustness, with most predictions closely following the ideal diagonal line. The application of **Gaussian noise–based data augmentation** further improved generalization, particularly for outlier vehicles and rare configurations.

The system returns the predicted car price upon user input and provides the foundation for a scalable, intelligent pricing engine. These results highlight the model's potential to enhance transparency and accuracy in the second-hand vehicle marketplace by reducing bias and enabling objective, feature-based price assessments.



Model	MAE (↓ Better)	RMSE (↓ Better)	R^2 Score (↑ Better)
Linear Regression	19945.90	31670.07	0.5792
Random Forest Regressor	3044.95	6783.65	0.9807
XGBoost	3172.57	8005.25	0.9731

Table.1. Output Values

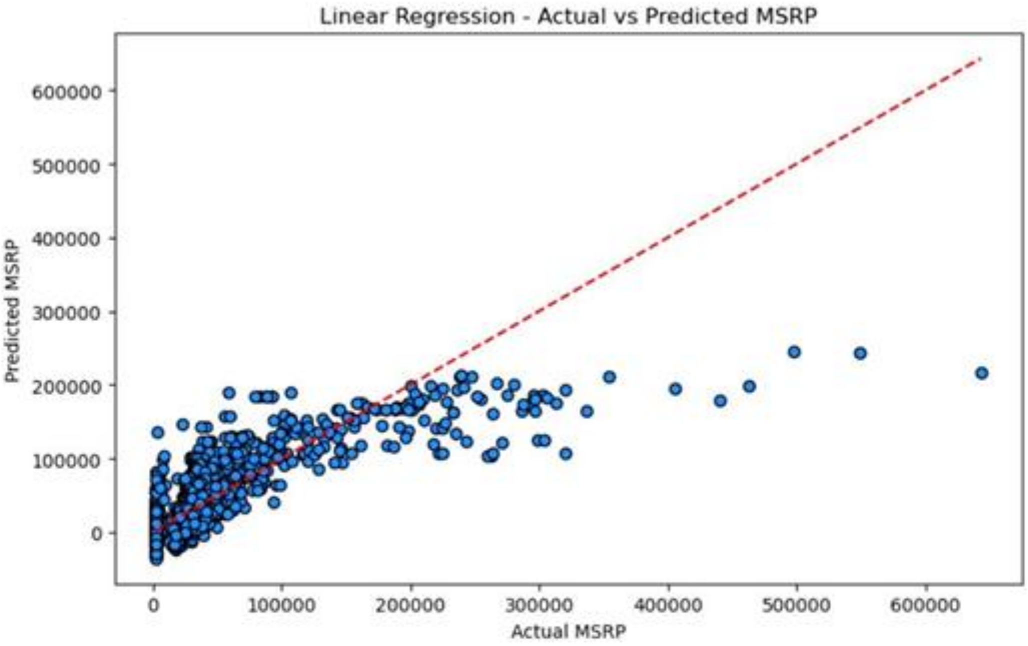


Fig.2. Linear Regression

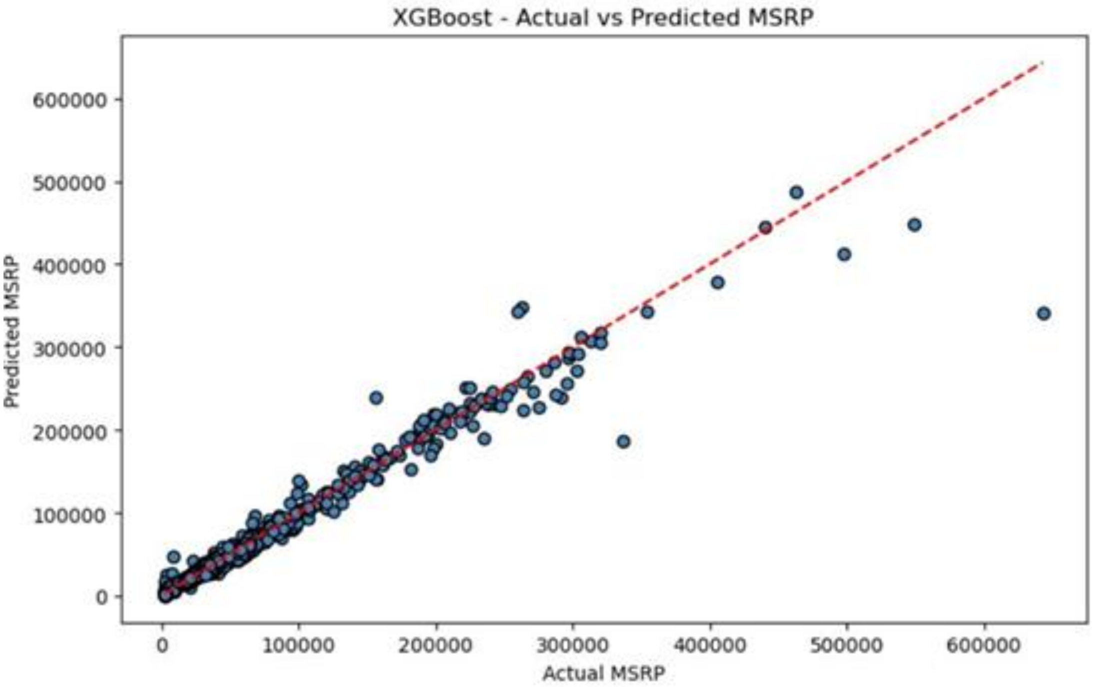


Fig.3. XGBoost

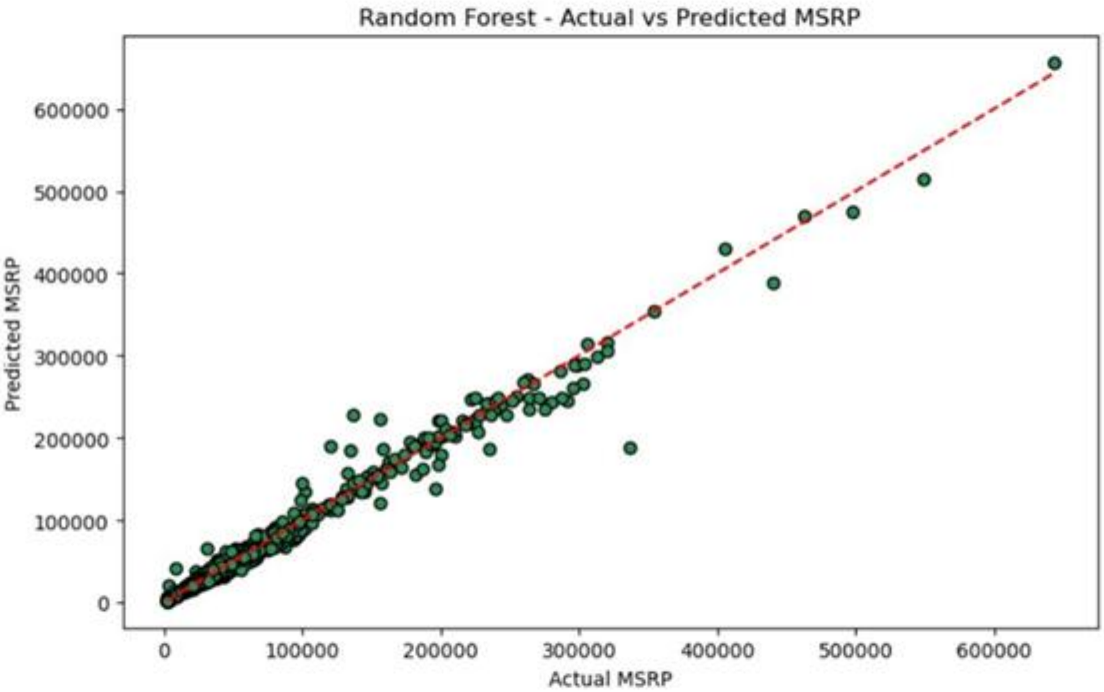


Fig.3. Random Forest Regressor

CONCLUSION

In conclusion, the **Random Forest Regressor** model has been effectively utilized in this work, showcasing the potential of machine learning in delivering accurate, automated car price predictions. The model, trained on a well-structured dataset containing detailed attributes of used cars—such as make, model, engine specifications, transmission type, mileage, and popularity—achieved an impressive **R² score of 0.9807**, demonstrating strong predictive performance.By estimating resale prices based on real-world vehicle data, the system addresses a significant gap in the second-hand automobile market where price assessments are often subjective and inconsistent. The successful application of ensemble-based regression confirms the Random Forest model's ability to capture complex, non-linear relationships while maintaining robustness against overfitting.

This research highlights how analysing vehicle characteristics through machine learning techniques can yield actionable insights for fair and objective pricing. Furthermore, the use of **Gaussian noise–based data augmentation** enhanced the model's generalization capabilities, making it well-suited for real-world deployment. The trained system has practical applications in online car listing platforms, dealer pricing tools, and resale marketplaces—where rapid, reliable, and transparent valuation is critical

FUTURE SCOPE

The implementation of the **Random Forest Regressor** model lays the foundation for a promising future in intelligent car price prediction systems. Moving forward, this work can be extended by incorporating **deep learning models** such as neural networks to capture more nuanced interactions between car features, especially in cases where high-dimensional or unstructured data (e.g., images, text descriptions) are available.

Additionally, integrating **real-time market data feeds**, such as fuel price trends, regional demand statistics, and economic indicators, could help the system dynamically adjust pricing predictions, enhancing relevance and accuracy. Reinforcement learning could also be employed to adaptively fine-tune price estimations based on buyer/seller feedback or sales outcomes.Advanced techniques like **transfer learning** and **natural language processing (NLP)** can be explored to process user reviews, listing descriptions, and dealership insights, further enriching the feature set. Furthermore, integrating **explainability tools** such as SHAP or LIME can increase transparency and build user trust by clarifying how each input feature contributes to the predicted price.

REFERENCES

[1] S. Choudhury, R. Ghosh, and S. Das, “Car Price Prediction Using Machine Learning Techniques,” International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 5, no. 2, pp. 82–88, 2019.

[2] A. Jain, A. Gupta, and P. Saxena, “Predicting Used Car Prices with Machine Learning Techniques,” International Journal of Advanced Research in Computer Science, vol. 10, no. 5, pp. 20–25,2019.

[3] K. N. Patel and S. K. Garg, “Comparative Analysis of Machine Learning Techniques for Car Price Prediction,” International Journal of Computer Applications, vol. 181, no. 43, pp. 30 36,2019.

[4] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,pp.785–794,2016.

[5] L. Breiman, “Random Forests,” Machine Learning, vol. 45, pp. 5–32,2001.

[6] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[7] M. Shorten and T. M. Khoshgoftaar, “A Survey on Image Data Augmentation for Deep Learning,” Journal of Big Data, vol. 6, no. 1, p. 60, 2019. (Used for Gaussian noise augmentation reference)

[8] S. Aggarwal and D. Kumar, “Price Prediction of Used Cars Using Machine Learning,” International Journal of Computer Applications, vol. 180, no. 47, pp. 22–26, 2018.