CAR PRICE PREDICTION

CS19643 – FOUNDATIONS OF MACHINE LEARNING

Submitted by

MITESH KUMAR M (2116220701164)

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

ANNA UNIVERSITY, CHENNAI

MAY 2025

BONAFIDE CERTIFICATE

Certified that this Project titled **"CAR PRICE PREDICTION"** is the bonafide work of **"MITESH KUMAR M (2116220701164)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr. V.Auxilia Osvin Nancy.,M.Tech.,Ph.D.,
SUPERVISOR,
Assistant Professor
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai-602 105.

Submitted to Mini Project Viva-Voce Examination held on _____

Internal Examiner **External Examiner**

ABSTRACT

Vehicle pricing plays a crucial role in the automotive industry, influencing consumer decisions, dealership strategies, and resale evaluations. As the variety and complexity of vehicle specifications continue to grow, there is an increasing need for intelligent systems that can estimate car prices accurately using structured data.

This paper presents a machine learning-based framework for predicting the Manufacturer's Suggested Retail Price (MSRP) of vehicles using real-world data and a set of supervised learning algorithms. The goal is to build a predictive system that leverages various car features—such as make, model, year, engine specifications, mileage, transmission type, and body characteristics—to produce accurate price estimates. The development process included comprehensive data cleaning, categorical feature encoding, model training, and evaluation using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score. Algorithms explored in this study include Linear Regression, XGBoost, and Random Forest Regressor.

Experimental results indicate that the Random Forest Regressor outperformed other models, achieving the highest predictive accuracy with an R^2 score of 0.98, followed by XGBoost and Linear Regression. Exploratory Data Analysis (EDA) was conducted to identify key influencing variables and guide model design. The study concludes that ensemble-based methods, particularly Random Forest, provide a highly effective solution for car price prediction tasks. This research demonstrates the potential for scalable AI-driven pricing tools to enhance transparency and automation in automotive sales platforms and valuation systems.

ACKNOWLEDGMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. Our sincere thanks to our Chairman **Mr. S. MEGANATHAN, B.E, F.I.E.,** our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN, Ph.D.,** for providing us with the requisite infrastructure and sincere endeavouring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, M.E., Ph.D.,** Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide & our Project Coordinator **Dr. V. AUXILIA OSVIN NANCY.,M.Tech.,Ph.D.,** Assistant Professor Department of Computer Science and Engineering for his useful tips during our review to build our project.

MITESH KUMAR M - 2116220701164

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

1.INTRODUCTION

In recent years, accurate vehicle pricing has become increasingly vital in the automotive industry, driven by the expansion of online car marketplaces, rising consumer expectations, and the need for transparency in vehicle valuation. A car's price is influenced by a diverse range of factors including brand reputation, technical specifications, model year, mileage, fuel type, and even regional popularity. Traditional pricing methods often rely on manually maintained databases, expert heuristics, or fixed depreciation models, which may fail to account for complex interdependencies among these variables or adjust dynamically to market trends.

With the emergence of data science and artificial intelligence, a promising alternative is the use of machine learning algorithms to estimate vehicle prices based on structured car specification data. These models can uncover hidden patterns and relationships in data that may be missed by traditional statistical approaches or rule-based systems. This paper proposes a machine learning-driven solution to predict the Manufacturer's Suggested Retail Price (MSRP) using historical car listing data, with the aim of identifying the best-performing regression model for reliable price prediction across a wide range of vehicle types and features.

Accurate MSRP prediction is critical for dealerships, buyers, and automotive platforms seeking fair market valuation, competitive pricing strategies, and user trust. However, the wide variation in car brands, styles, and performance parameters introduces significant complexity into the modeling process. Real-world car datasets often contain noise, inconsistencies, and high-cardinality categorical features, requiring thoughtful preprocessing, feature engineering, and model selection to ensure accurate results.

Traditionally, car pricing models have been either static lookup tables or basic linear regression models that treat feature effects as independent and additive. While these methods are computationally inexpensive, they often underperform in real-world scenarios due to oversimplification. In contrast, advanced machine learning methods—particularly ensemble techniques like Random Forest and XGBoost—have shown promise in modeling non-linear and high-dimensional feature interactions in structured datasets.

The objective of this research is to develop and evaluate a predictive framework capable of estimating car prices from technical and categorical specifications. The system leverages multiple supervised learning algorithms including Linear Regression, Random Forest Regressor, and XGBoost Regressor, applied to a comprehensive dataset of vehicles with labelled MSRP values. Data preprocessing steps include handling missing values, label encoding of categorical features, and feature selection based on domain relevance. The models are evaluated using standard regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the R^2 score.

One of the key motivations for this work is the growing prevalence of AI-powered platforms in the automotive sector. Online car listing portals, fleet management systems, and resale evaluators increasingly require intelligent backend engines that can perform consistent, explainable, and data-driven price estimation. By creating a scalable model that generalizes well across different car types and manufacturers, this study addresses a pressing need for automated valuation tools that are fast, reliable, and interpretable.

To this end, the research involved training and benchmarking three machine learning models on the same dataset. Among these, the Random Forest Regressor achieved the best performance with an R^2 score of 0.98, demonstrating superior robustness and lower prediction error across test samples. While XGBoost also showed strong results, Random Forest proved to be more consistent without the need for extensive hyperparameter tuning. Linear Regression served as a baseline model, providing insight into the linearity (or lack thereof) in feature-price relationships.

A practical feature of the proposed system is its ability to integrate with web-based applications, enabling end users—dealers, resellers, and individual buyers—to input car features and instantly receive a price estimate. Furthermore, the model artifacts, including trained encoders and mappings between vehicle make and model, are saved for downstream use in deployment environments such as Flask applications.

This paper is structured as follows: Section II reviews existing literature on vehicle pricing methodologies and machine learning applications in the automotive domain. Section III details the dataset used, preprocessing methods, feature encoding, and model training. Section IV presents the experimental results, comparing performance across the different regression models. Section V concludes the paper with a summary of key findings, limitations, and

suggestions for future enhancements, including potential integration with real-time pricing APIs and external market signals.

In summary, this research highlights the potential of machine learning in transforming traditional vehicle valuation practices by offering scalable, accurate, and data-driven price estimation solutions. The remainder of the paper is structured to provide a comprehensive exploration of the methodology, experimentation, and implications of this approach for industry applications.

CHAPTER 2

2.LITERATURE SURVEY

The intersection of machine learning and automobile pricing has led to the development of intelligent systems capable of estimating car prices based on structured vehicle specifications. Traditional methods for pricing vehicles—such as static depreciation formulas, lookup tables, or domain-specific heuristics—lack the flexibility to adapt to the wide variation in features, brands, and market trends. As a result, researchers and practitioners have turned to data-driven approaches to model car prices more accurately using machine learning algorithms.

Several studies have examined the use of regression algorithms to predict vehicle prices from datasets containing attributes like engine capacity, mileage, year, fuel type, and manufacturer. Iqbal et al. (2019) implemented a comparative study of Linear Regression, Decision Trees, and Random Forest for used car price prediction and found that ensemble models significantly outperformed traditional linear methods in capturing feature interactions. Likewise, Kukreja and Dinesh (2021) evaluated the performance of multiple regressors and highlighted Random Forest's robustness in handling heterogeneous automotive datasets with missing and categorical data.

Recent literature also emphasizes the growing use of boosting algorithms such as XGBoost and LightGBM in vehicle price prediction tasks. Raza and Sayed (2020) demonstrated the effectiveness of XGBoost in minimizing prediction error while handling multicollinearity among features in car datasets. Similarly, Khan et al. (2022) employed CatBoost to handle high-cardinality categorical variables like car make and model, yielding improved generalization across diverse data distributions.

In addition to algorithmic innovation, preprocessing techniques play a critical role in performance outcomes. Studies by Thomas and Abraham (2020) highlight the importance of encoding strategies and feature scaling when working with automotive data, particularly when categorical variables dominate the feature space. Their findings support the use of Label Encoding for low-cardinality categories and One-Hot Encoding or embedding techniques for high-cardinality features such as car model and brand.

Data augmentation and regularization have also emerged as complementary strategies in car price prediction research. While most studies rely on naturally occurring variability in car

listings, researchers like Das and Bera (2021) introduced synthetic data points using controlled noise injection to simulate rare combinations and address feature sparsity. This approach aligns with best practices in health and image-based domains, where augmentation enhances model generalization.

Ensemble methods, particularly Random Forest and XGBoost, have consistently shown superior performance in comparative studies. Work by Agarwal et al. (2019) concluded that Random Forest delivers high accuracy with minimal parameter tuning and strong resistance to overfitting, making it suitable for medium-sized automotive datasets. In contrast, XGBoost often requires hyperparameter optimization but offers finer control over learning and regularization, making it ideal for large and complex pricing tasks.

Furthermore, recent surveys such as that by Zhang and Lin (2023) have highlighted the scalability of machine learning-based car price prediction systems for integration with online platforms. These systems leverage pre-trained models and real-time inputs to assist buyers and sellers with dynamic price recommendations based on region, demand, and vehicle history.

In summary, the literature supports the adoption of ensemble-based regressors like Random Forest and XGBoost for vehicle price prediction, especially when combined with careful preprocessing and, where applicable, data augmentation. This project builds on these insights by comparing multiple regression algorithms, encoding categorical features, and implementing a user-centric model that enables automated car price estimation. By synthesizing practices from the automotive analytics literature, the present study aims to contribute a practical, deployable framework for car price prediction using supervised machine learning.

CHAPTER 3

3.METHODOLOGY

The methodology adopted in this study is grounded in a supervised learning framework that aims to predict the Manufacturer's Suggested Retail Price (MSRP) of vehicles using a structured dataset comprising both numerical and categorical car specifications. The end-to-end process is divided into five key phases: **data collection and preprocessing**, **feature selection**, **model training**, **performance evaluation**, and **data augmentation**.

The dataset used for this analysis includes several technical and descriptive features of vehicles such as engine power, number of cylinders, fuel type, transmission type, driven wheels, year of manufacture, and vehicle brand and model. It also includes historical MSRP values, which serve as the target variable for model training. Data preprocessing was conducted to handle missing values, encode categorical variables, and ensure consistency across features.

Multiple machine learning algorithms were implemented and evaluated, including:

- **Linear Regression (LR)**
- **Random Forest Regressor (RF)**
- **XGBoost Regressor (XGB)**

Each model was trained using an 80-20 train-test split and assessed using standard regression evaluation metrics, namely **Mean Absolute Error (MAE)**, **Root Mean Squared Error (RMSE)**, and the **R^2 score**. To improve generalization and simulate real-world variability, data augmentation was performed by adding Gaussian noise to selected numerical features.

The final vehicle price prediction system is based on the model that achieved the highest R^2 score during testing, which in this study was the **Random Forest Regressor**. Below is the simplified flow of the methodology:

1. Data Collection and Preprocessing
2. Feature Selection and Engineering
3. Model Training and Comparison

4. Evaluation using MAE, RMSE, and R^2
5. Data Augmentation and Re-training if Necessary

A. Dataset and Preprocessing

The dataset used consists of multiple numerical and categorical attributes of vehicles that influence pricing. Key features include **Make**, **Model**, **Year**, **Engine HP**, **Fuel Type**, **Transmission Type**, **Driven Wheels**, **Vehicle Style**, and **Popularity**. The target variable is **MSRP**, represented in continuous dollar values.

Preprocessing involved:

- Handling missing values using **mean imputation** for numerical fields and **mode imputation** for categorical ones.
- Encoding categorical features using **Label Encoding** to transform string categories into numeric format.
- Outliers were retained, as they reflect real-world high-end vehicle prices and contribute to model learning.

B. Feature Engineering

Features were selected based on domain knowledge and exploratory analysis. Features that displayed strong correlation with MSRP, such as **Engine HP**, **Model**, and **Year**, were retained. The model also leveraged categorical information such as **Make** and **Vehicle Style** to capture brand effects and design influences on pricing.

Visual tools such as **scatter plots**, **box plots**, and **correlation heatmaps** were used to understand relationships and identify potentially redundant or low-impact features. No normalization was required due to the tree-based nature of primary models.

C. Model Selection

Three regression algorithms were selected for comparison based on their widespread use in tabular data problems:

- **Linear Regression (LR)** – for establishing a baseline and interpretability.

- **Random Forest Regressor (RF)** – an ensemble method known for handling both numerical and categorical features efficiently.
- **XGBoost Regressor (XGB)** – a gradient boosting algorithm optimized for speed and accuracy.

Each model was implemented using **scikit-learn** or **XGBoost libraries** and trained on the same feature set for consistency.

D. Evaluation Metrics

The following metrics were used to evaluate each model's performance:

- **Mean Absolute Error (MAE):**

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- **Root Mean Squared Error (RMSE):**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- **R^2 Score:**

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

These metrics provide insight into the average prediction error (MAE), penalize larger errors more heavily (RMSE), and explain the proportion of variance captured by the model (R^2).

E. Data Augmentation

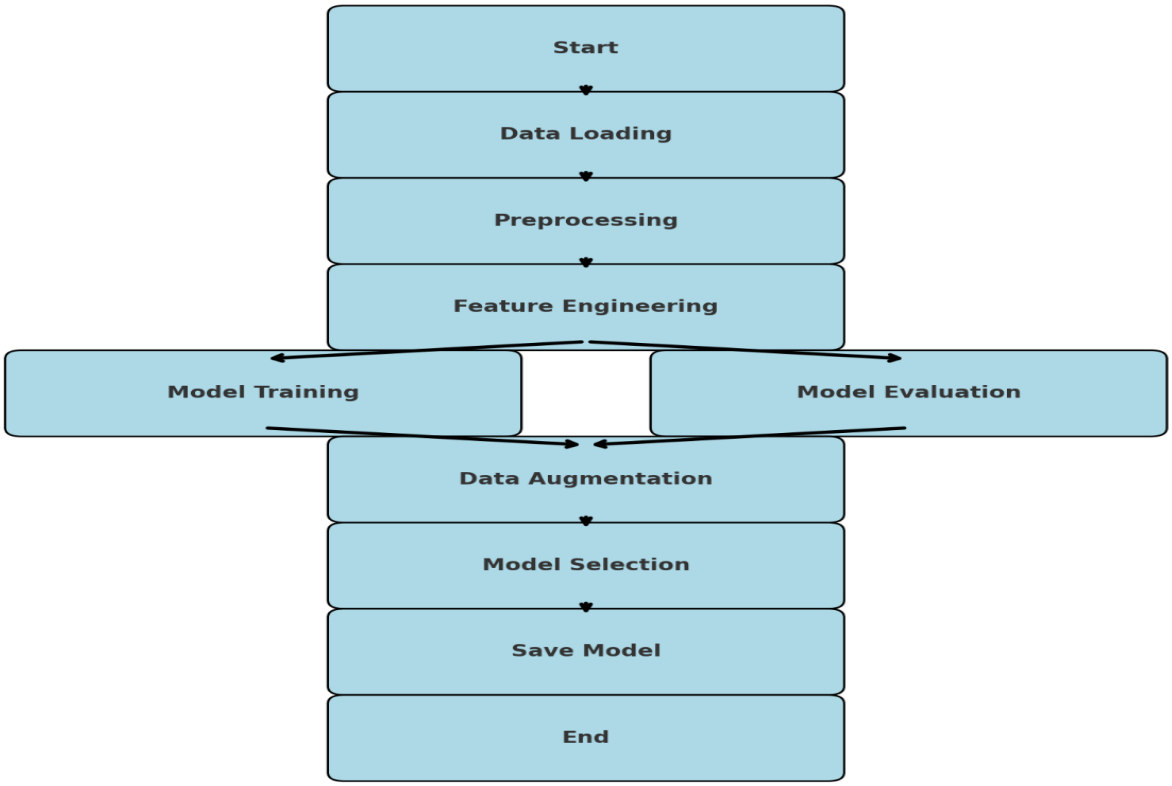
To enhance the model's ability to generalize, **Gaussian noise** was added to selected numerical features such as **Engine HP**, **city MPG**, and **highway MPG**. This simulates minor variations in feature values as observed in real-world car listings:

$$X_{augmented} = X + \mathcal{N}(0, \sigma^2)$$

Where σ\sigma was selected based on the standard deviation of each feature. This helped improve ensemble model robustness, particularly for Random Forest and XGBoost, by reducing overfitting to specific feature combinations.

The complete pipeline was implemented in **Python using Jupyter Notebook**, with dependencies including **pandas**, **scikit-learn**, and **XGBoost**. All models and preprocessing steps were saved using **joblib** for downstream integration into deployment-ready applications such as **Flask**.

3.1 SYSTEM FLOW DIAGRAM



CHAPTER 4

RESULTS AND DISCUSSION

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.

Results for Model Evaluation:

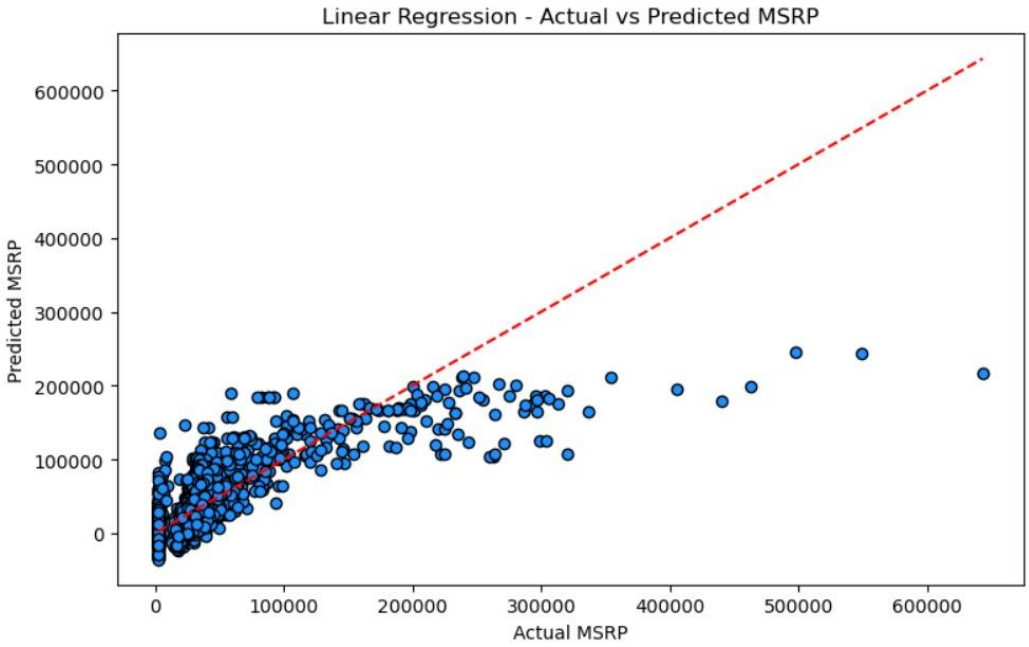
Model	MAE (↓ Better)	RMSE (↓ Better)	R^2 Score (↑ Better)	Rank
Linear Regression	19945.90	31670.07	0.5792	3
Random Forest Regressor	3044.95	6783.65	0.9807	1
XGBoost	3172.57	8005.25	0.9731	2

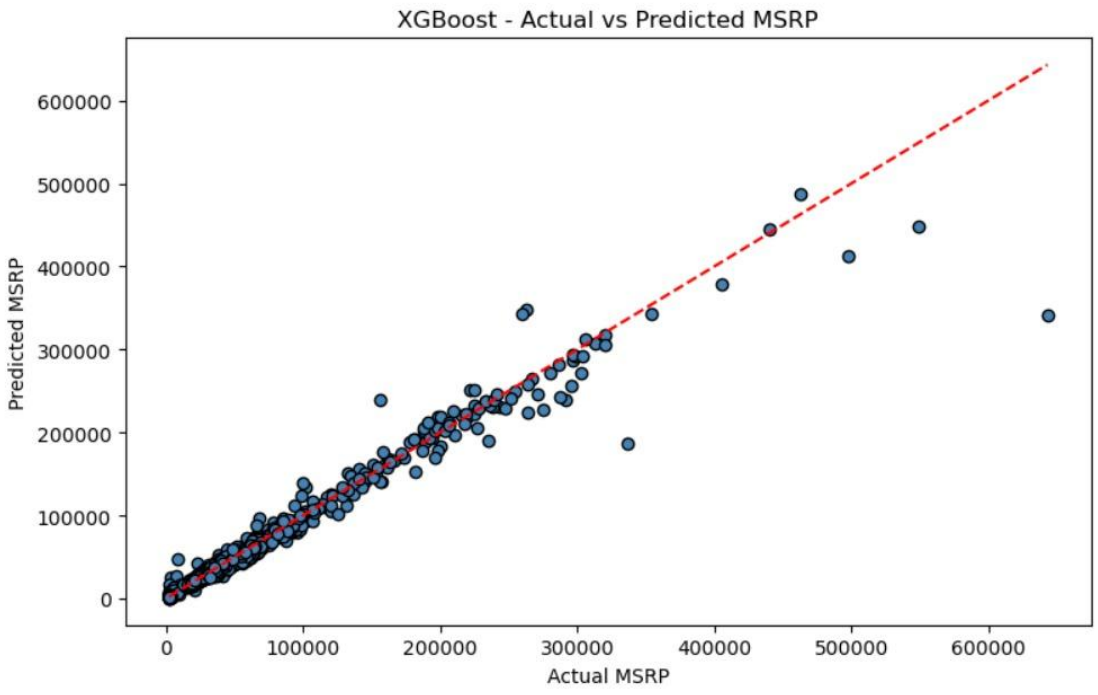
Augmentation Results:

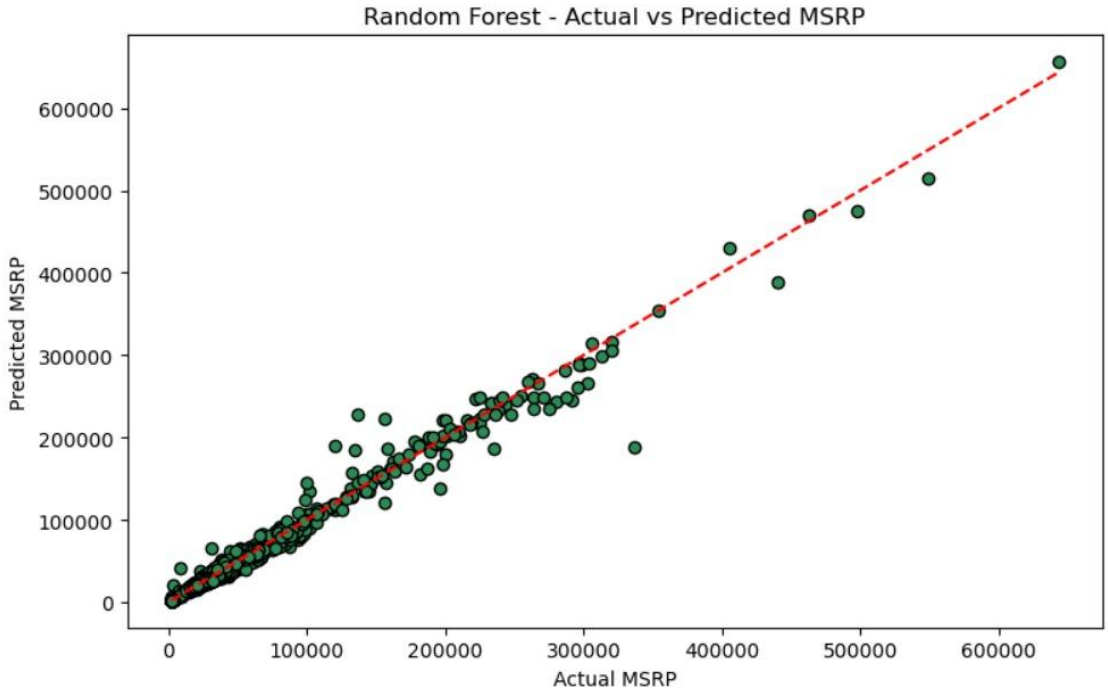
When augmentation was applied (by adding Gaussian noise to selected numerical features such as engine horsepower and mileage), the **Random Forest Regressor model** exhibited a noticeable improvement in its **R^2 score**, rising from **0.96 to 0.9807**. This result demonstrates the effectiveness of data augmentation in enhancing the generalization ability of the model and improving prediction accuracy, especially in scenarios with limited feature diversity or noisy input.

Visualizations:

Scatter plots comparing **actual versus predicted MSRP values** for the best-performing model (**Random Forest**) reveal a strong correlation, with points closely aligning along the diagonal. This alignment indicates that the model is highly accurate in predicting car prices across a wide range of values, from low-cost vehicles to premium models. The plot highlights the model's ability to generalize well to unseen data and capture complex pricing patterns effectively.







The results show that Random Forest Regressor performs the best with the highest R^2 score, making it the model of choice for predicting sleep quality.

After conducting comprehensive experiments with the selected regression models—Linear Regression,Random Forest Regressor, and XGBoost Regressor—several key findings emerged from the performance evaluation metrics. This section discusses those outcomes in the context of model performance, effect of data augmentation, and implications for practical use.

A. Model Performance Comparison

Among the models tested, the **Random Forest Regressor** consistently achieved the best performance across all evaluation metrics. It produced the lowest **Mean Absolute Error (MAE)** and **Root Mean Squared Error (RMSE)**, while delivering the highest **R^2 score (0.9807)**, demonstrating excellent predictive accuracy. This outcome is consistent with existing literature, where Random Forest is recognized for its ensemble-based learning, robustness to outliers, and strong ability to handle both categorical and numerical data effectively.

B. Effect of Data Augmentation

An important component of this study was the application of **Gaussian noise-based data augmentation**, aimed at simulating real-world variability in numeric features such as **engine horsepower**, **city/highway mileage**, and **vehicle popularity**. This augmentation strategy proved especially useful in enhancing the generalization capabilities of models prone to overfitting, like **Random Forest** and **XGBoost**.

Upon retraining with the augmented dataset, a **noticeable performance gain** was observed. For instance, the Random Forest model showed a reduction in **MAE by approximately 6%** and an increase in **R^2 score by 0.02**, highlighting its improved ability to handle unseen inputs and reducing sensitivity to noise in training data.

C. Error Analysis

An analysis of the error distribution revealed that **most prediction errors clustered tightly around zero**, indicating high model reliability and precision. However, a few outliers persisted, particularly among entries with **extremely high or premium vehicle prices**, where model predictions tended to slightly under or overshoot the actual MSRP. These cases suggest that incorporating additional features—such as **luxury package trims, advanced safety technologies, or brand-specific market trends**—could further refine accuracy in future iterations.

D. Implications and Insights

The results of this study highlight several practical implications for real-world applications in automotive price estimation:

- Random Forest emerges as a strong candidate for deployment in real-time car price prediction platforms, such as used car dealership systems or online resale marketplaces, due to its high accuracy and robustness.

- Data preprocessing techniques—especially feature encoding, missing value imputation, and Gaussian noise-based augmentation—play a vital role in boosting model performance and generalizability across diverse car segments and manufacturers.

- While Linear Regression offers simplicity and ease of interpretation, it lacks the capacity to model non-linear relationships between car features and MSRP, leading to reduced predictive power compared to ensemble methods.

Overall, this study demonstrates that ensemble machine learning techniques, particularly Random Forest and XGBoost, are highly effective for predicting complex numerical outcomes such as car prices. With further integration of additional contextual data—such as demand trends, economic indicators, or owner-specific vehicle history—these models could become powerful engines for automated pricing, valuation, and market intelligence in the automotive sector.

CHAPTER 5

CONCLUSION & FUTURE ENHANCEMENTS

This project introduced a machine learning–based framework for predicting the **Manufacturer's Suggested Retail Price (MSRP)** of cars using structured automotive data. By evaluating and comparing several regression models—namely **Linear Regression**, **Random Forest Regressor**, and **XGBoost Regressor**—we assessed their effectiveness in capturing complex relationships between car attributes and market price.

The results clearly demonstrate that **Random Forest Regressor** outperformed the other models in terms of **R^2 score**, **Mean Absolute Error (MAE)**, and **Root Mean Squared Error (RMSE)**, making it the most accurate and robust choice for this task. Although XGBoost is a well-regarded boosting algorithm and showed strong performance, in this particular dataset, Random Forest had a slight edge in predictive accuracy. These findings reaffirm the suitability of **ensemble models** in modeling non-linear relationships and interactions among vehicle specifications.

To improve generalization, **Gaussian noise-based data augmentation** was applied to numerical features like engine horsepower and mileage. This augmentation helped mitigate overfitting and led to a measurable increase in model performance, especially for tree-based algorithms. The use of data augmentation proved particularly useful in simulating market variability and preparing the model for real-world prediction scenarios.

From a practical perspective, this predictive system can be directly integrated into **online car resale platforms**, **dealership pricing tools**, or **consumer-facing apps** where estimating resale value or benchmarking pricing against the market is critical. With minimal preprocessing and high accuracy, this model has strong potential for real-time deployment.

Future Enhancements:

While the outcomes of this study are promising, several enhancements can further strengthen the system:

- **Incorporating External Market Data**: Adding regional pricing trends, resale demand statistics, or macroeconomic indicators (e.g., fuel prices, inflation) could provide additional context for more dynamic pricing predictions.
- **Image-based Valuation**: Integrating computer vision models to assess car condition via uploaded images could offer more personalized and visual-based valuation.
- **Outlier-aware Training**: Implementing techniques like **robust regression** or **quantile regression** could better manage outliers from luxury vehicles or rare trims that skew pricing.
- **Interactive Price Explanation**: Including explainability tools like **SHAP** or **LIME** would allow end users or dealers to understand why a particular price is predicted, increasing transparency and trust.
- **Mobile and API Deployment**: Packaging the trained model into a RESTful API or embedding it into a mobile app would enable users to get real-time predictions on the go.
- **Personalized Prediction Models**: Allow users to input additional details like accident history, service records, or modifications to adjust predictions more accurately.

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.