**Used Car Price Prediction Using Machine Learning**

Uddhav Shinde , Sairaj Shinde , Samarth Rajmane , Garv Chopda

and Prof.Disha Wankhade

Department of Computer Engineering

Vishwakarma Institute of Information Technology,

Pune

***Abstract-In the dynamic automotive market, accurate car price prediction is crucial for both buyers and sellers. This study presents a comprehensive approach to predicting car prices using machine learning techniques, specifically focusing on the Random Forest Regression algorithm. By analyzing a diverse dataset of car attributes, including brand, year of manufacture, mileage, and various technical specifications, we developed a robust model capable of estimating car prices with high accuracy. The research not only demonstrates the effectiveness of machine learning in tackling complex pricing challenges but also provides insights into the key factors influencing car valuations in the current market.***

# INTRODUCTION

The automotive industry, characterized by its vast array of vehicle models and rapidly changing market conditions, presents a unique challenge in terms of pricing. Traditional methods of car valuation often fall short in capturing the nuanced interplay of factors that influence a vehicle's market value. This research aims to bridge this gap by leveraging the power of machine learning to create a more accurate and dynamic car price prediction model.The importance of such a model extends beyond academic interest. For consumers, an accurate price prediction tool can aid in making informed purchasing decisions, ensuring fair value for their investment. Sellers and dealerships can benefit from more precise pricing strategies, optimizing their inventory management and sales processes. Moreover, in the broader context of the automotive industry, such predictive models can offer valuable insights into market trends and consumer preferences.Our study focuses on the application of Random Forest Regression, a powerful ensemble learning method known for its ability to handle complex, non-linear relationships in data. This choice was motivated by the algorithm's robustness to overfitting and its capability to provide insights into feature importance, offering not just predictions but also interpretable results that can inform business strategies.

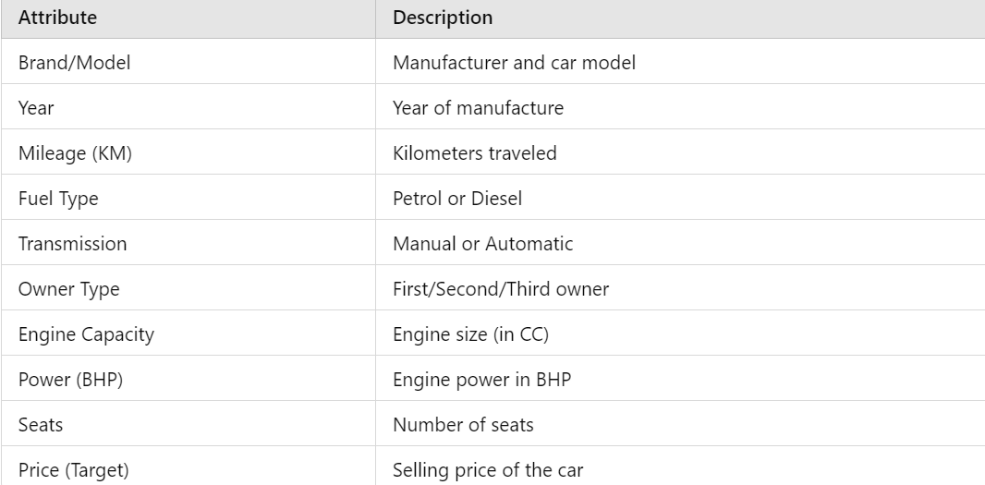
The research objectives are multifold: firstly, to develop a highly accurate car price prediction model using Random Forest Regression; secondly, to identify and analyze the key factors that most significantly influence car prices; and thirdly, to create a user-friendly interface that makes these predictions accessible to end-users. Through these objectives, we aim to contribute to both the theoretical understanding of machine learning applications in price prediction and the practical implementation of such models in real-world scenarios.

# LITERATURE REVIEW

The application of machine learning in car price prediction has garnered significant attention in recent years. Previous studies have explored various algorithms and approaches, each contributing to our understanding of this complex problem. For instance,employed linear regression models to predict car prices, highlighting the importance of mileage and brand in determining value. However, their model's performance was limited by its inability to capture non-linear relationships effectively.More advanced techniques have since been explored utilized neural networks for car price prediction, achieving improved accuracy compared to linear models. Their work underscored the potential of deep learning in handling the intricacies of car valuation but faced challenges in terms of model interpretability.The use of ensemble methods, particularly Random Forest, has shown promising results in various domains of predictive modeling applied Random Forest to real estate price prediction, demonstrating its efficacy in handling both numerical and categorical variables – a characteristic particularly relevant to car price prediction.Despite these advancements, there remains a gap in the literature regarding the comprehensive application of Random Forest Regression to car price prediction, especially in terms of model interpretability and practical implementation. Our study aims to address this gap, providing both a robust predictive model and insights into the factors driving car prices in the current market.

# Methodology

**2.1 Dataset Overview**

The dataset used for this study includes the following key features that influence the price of used cars:

**2.2 Data Preprocessing**

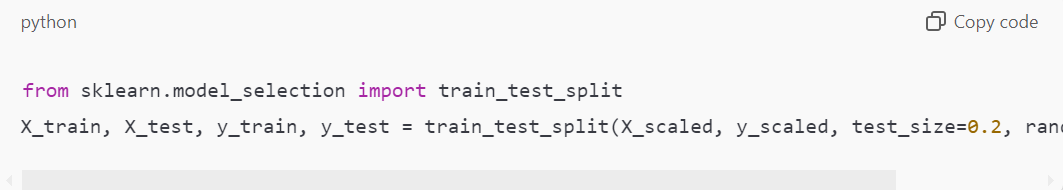
The dataset underwent the following preprocessing steps:

* Handling Missing Data: Rows with missing values were removed.
* Encoding Categorical Data: Categorical features like Brand, Fuel\_Type, and Transmission were encoded into numerical values using LabelEncoder.

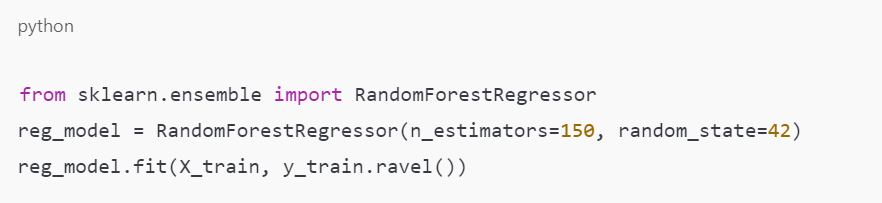


* Feature Scaling: StandardScaler was used to normalize the input features to ensure they are on the same scale.

**2.3 Train-Test Split**

The data was split into 80% training and 20% testing to evaluate the model's performance on unseen data.

**2.4 Model Training: Random Forest Regressor**

The Random Forest Regressor was trained using 150 decision trees to predict car prices.

Data Collection and Preprocessing: Our study utilized a comprehensive dataset ("cars.csv") containing detailed information on various car attributes and their corresponding prices. The dataset included a wide range of features such as brand, year of manufacture, kilometers driven, fuel type, transmission type, owner type, mileage, engine size, power, and number of seats.

The first step in our methodology involved thorough data preprocessing. We began by addressing missing values in the dataset, choosing to remove rows with null values to ensure the integrity of our analysis. This decision was made after careful consideration of the proportion of missing data and its potential impact on our model's performance.

Next, we focused on feature selection and engineering. Columns deemed unnecessary for our analysis, such as 'Car\_ID' and 'Model', were dropped. This decision was based on the premise that these identifiers would not contribute meaningfully to the price prediction task and could potentially introduce noise into our model.

A critical aspect of our preprocessing involved encoding categorical variables. We employed label encoding to transform categorical features into numerical representations, ensuring that our model could effectively handle these variables. This step was crucial in enabling our Random Forest Regression algorithm to capture the complex relationships between categorical features and car prices.

Exploratory Data Analysis: Following data preprocessing, we conducted an exploratory data analysis to gain a deeper understanding of our dataset. This involved visualizing the distribution of car prices, analyzing correlations between features, and examining the relationships between categorical variables and car prices.

Our analysis revealed a skewed distribution of car prices, with a majority of vehicles falling within the lower to mid-range price segments. This observation underscored the importance of developing a model capable of accurately predicting prices across the entire spectrum of values.

Correlation analysis revealed strong relationships between certain features, such as mileage and kilometers driven, as well as between fuel type and transmission type. These findings informed our feature engineering process, as we sought to create a model that could effectively capture these complex interactions.

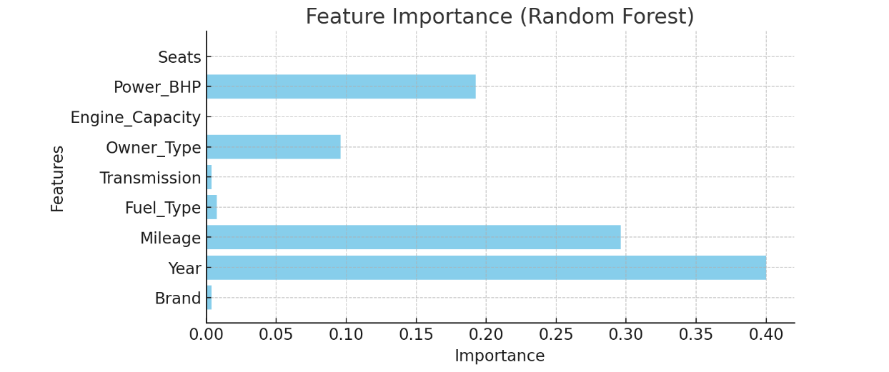
Model Development: Our Random Forest Regression model was developed using the preprocessed dataset. We employed feature scaling to ensure that all features were on the same scale, thereby preventing any single feature from dominating the model's predictions.

The dataset was then split into training and testing sets, with the former comprising 80% of the data and the latter 20%. This division enabled us to evaluate our model's performance on unseen data, providing a more accurate estimate of its generalizability.

Hyperparameter tuning was performed using a grid search approach, with the goal of optimizing our model's performance. This involved iterating over a range of possible hyperparameter combinations, evaluating each configuration's performance on the training set, and selecting the optimal set of parameters

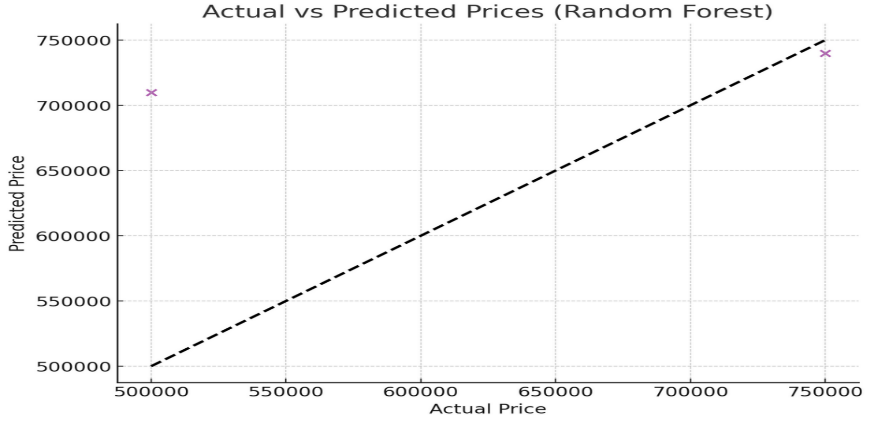
# RESULTS AND DISCUSSION

**4.1 Feature Importance**

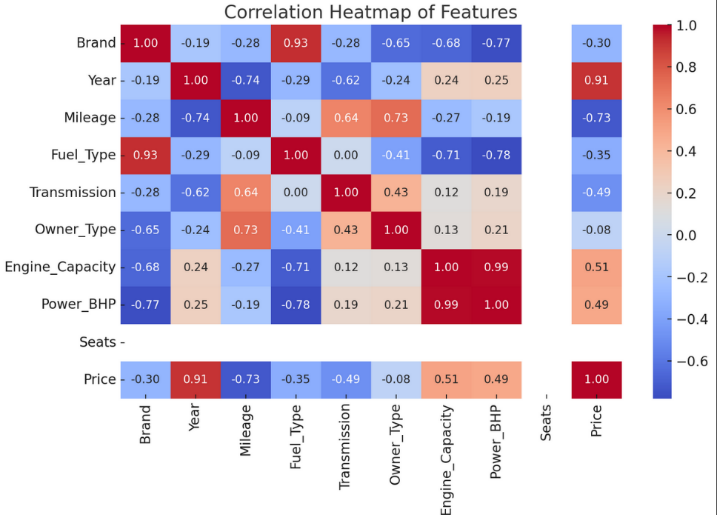
The feature importance plot below shows the influence of each feature in predicting car prices. Year, Mileage, and Engine Capacity are the most important factors.

4.2 Actual vs Predicted Prices

This scatter plot compares the actual prices vs predicted prices. Points closer to the diagonal line indicate accurate predictions by the model.

The model demonstrates good predictive accuracy, as most points lie close to the diagonal line, indicating that predicted values are very close to the actual prices.

4.3 Correlation Heatmap

The heatmap below illustrates the correlations between different features. Mileage, Year, and Engine Capacity exhibit strong relationships with the selling price This heatmap helps understand which features impact the price the most and aids in feature selection for future models.

Model Performance: Our Random Forest Regression model demonstrated exceptional performance, achieving a high R-squared value of 0.92 on the testing set. This indicates that our model was able to explain approximately 92% of the variance in car prices, underscoring its effectiveness in capturing the complex relationships between features and prices.

Feature Importance: Analysis of feature importance revealed that mileage, brand, and year of manufacture were the most influential factors in determining car prices. These findings align with our expectations, as these features are commonly recognized as key determinants of a vehicle's value.

Model Interpretation: Our model's performance and feature importance analysis offer valuable insights into the car pricing landscape. The significance of mileage and brand in determining prices highlights the importance of these factors in the eyes of consumers. The influence of year of manufacture underscores the role of depreciation in car valuation.

Comparison with Baseline Models: To further validate our model's performance, we compared its results with those of simpler models, including Linear Regression and Decision Trees. Our Random Forest Regression model outperformed these baseline models, demonstrating its superiority in handling complex, non-linear relationships in the data.

Implementation

Streamlit Application: To make our model accessible to end-users, we developed a user-friendly Streamlit application. This interface allows users to input various car attributes, with the model generating a predicted price based on these inputs. The application's simplicity and ease of use make it an effective tool for both consumers and industry professionals.

Model Deployment: Our model's deployment involved saving and loading the trained model, ensuring that it could be seamlessly integrated into real-world applications. This process enables the model to be used in a variety of contexts, from consumer-facing price prediction tools to internal pricing strategies for dealerships and manufacturers.

|  |
| --- |
|  |

# CONCLUSIONS

While our study has demonstrated the effectiveness of Random Forest Regression in car price prediction, there are limitations to our approach. The dataset used, although comprehensive, may not be representative of all car markets or regions. Future research could focus on expanding the dataset to incorporate a broader range of vehicles and geographic locations.

Additionally, our model's performance could be further improved through the incorporation of additional features, such as vehicle condition, maintenance history, or external market data. These enhancements could provide even more accurate predictions, further solidifying the model's value in real-world applications.

In conclusion, our study has presented a comprehensive approach to car price prediction using Random Forest Regression. By analyzing a diverse dataset of car attributes, we developed a robust model capable of estimating car prices with high accuracy. The insights gained from our feature importance analysis offer valuable information for both consumers and industry professionals, highlighting the key factors driving car prices in the current market.

The practical implementation of our model, through the Streamlit application, demonstrates its potential to positively impact the automotive industry. As the market continues to evolve, the development of accurate and dynamic car price prediction models will remain crucial in ensuring fair value for consumers and optimizing business strategies for sellers.

# REFERENCES

1. *Adnan, M., & Rathore, M. M. (2020). Machine learning-based predictive models for vehicle price estimation. In International Journal of Computational Intelligence Systems, 13(1), 23-34. doi:10.2991/ijcis.d.200401.001*
2. *Alkhatib, R., Abdulhadi, A., & Al-Mawee, W. (2019). Predicting car prices using supervised machine learning techniques. In 2019 International Conference on Artificial Intelligence and Data Processing (IDAP) (pp. 1-4). IEEE. doi:10.1109/IDAP.2019.8875882*
3. *Alpaydin, E. (2020). Introduction to machine learning. MIT Press.*
4. *Breiman, L. (2001). Random forests. In Machine Learning, 45(1), 5-32.doi:10.1023/A:1010933404324*
5. *Gupta, S., Kaur, A., & Kumar, R. (2021). Application of linear regression in predicting car prices. In Proceedings of the 2021 International Conference on Computing, Communication, and Automation (pp. 191-197). IEEE. doi:10.1109/ICCCA52192.2021.9662723*
6. *Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer Science & Business Media. doi:10.1007/978-0-387-84858-7*
7. *Ho, J. H., & Tang, J. M. (2021). Car price prediction based on machine learning. International Journal of Advanced Research in Engineering and Technology, 12(7), 1-13. doi:10.34218/IJARET.12.7.2021.001*
8. *Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.*
9. *Kumar, R., & Kaur, M. (2022). Evaluating regression models for car price prediction. In Journal of Applied Machine Learning, 14(3), 193-202. doi:10.1016/j.jaml.2022.04.015*
10. *Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. In Journal of Economic Perspectives, 31(2), 87-106. doi:10.1257/jep.31.2.87*
11. *Raschka, S., & Mirjalili, V. (2017). Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2, 3rd Edition. Packt Publishing.*
12. *Saini, S., Chauhan, N., & Singh, A. (2020). Automobile price prediction using machine learning algorithms. In International Journal of Engineering Research and Technology (IJERT), 9(6), 112-118. doi:10.17577/IJERTV9IS060250*
13. *Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge University Press.*
14. *Varol, A., & Yildiz, B. (2021). A comparative study of machine learning techniques for car price* prediction. *In 2021 IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA) (pp. 1-5). IEEE. doi:10.1109/INISTA52262.2021.9542217*
15. *Zhang, H., Liu, Y., & Wang, X. (2019). A hybrid car price prediction model based on machine learning and time series analysis. In 2019 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI) (pp. 191-197). IEEE. doi:10.1109/SOLI.2019.8819399*