Volume 10, Issue 1, January – 2025 *International Journal of Innovative Science and Research Technology (IJISRT)*
ISSN No:-2456-2165 *https://doi.org/10.5281/zenodo.14724989*
IJISRT25JAN852 *www.ijisrt.com*

Predictive Modeling of Vehicle Characteristics and Pricing Using Machine Learning Algorithms

Nikhil Sharma[1]; Gaurang Shirodkar[2]; Nikhil Singh[3]; Rohan A Mathews[4]; Ragavan R[6]; and Shobha T[6]

[1,2,3,4,5,6]Department of Information Science and Engineering, B.M.S College of Engineering

Publication Date: 2025/01/24

Abstract

Since the advent of the automobile by Karl Benz, there has been an exponential increase in the number of automobiles worldwide and the growth of the automobile industry. In this situation, it becomes increasingly significant to have a precise pricing model, which is crucial for both buyers and sellers. Such models provide reliable pricing information, enabling better-informed decisions. This project aims to design a prediction system using machine learning capable of predicting prices for used cars based on the most relevant factors such as make, model, mileage, year, fuel type, and transmission type. The ML models trained and tested to form the system include Random Forest, Gradient Boosting, and Stacking Regressor. Model stacking and ensemble voting were con- ducted to increase prediction accuracy. Experimental results demonstrate that an ensemble model possesses far higher predictive power, accuracy, and robustness compared to single models. The final ensemble model produced an X RMSE and Y R^2 score, proving that vehicle prices can be predicted fairly well using this approach. Thus, the system becomes more useful for consumer and dealer applications.

Keywords: *Precise Pricing Model, Informed Decisions, Price Prediction, Random Forest, Gradient Boosting, Ensemble voting, Accuracy, Robustness.*

I. INTRODUCTION

Today, vehicle condition, mileage, year, brand, and fuel type remain highly relevant factors in estimating prices for any used vehicle which are subjective and aren't precise. This problem can be solved by applying machine learning with high-end data models able to ingest, process, and analyze huge volumes of information.Second modern language models sensibly coupled with machine learning together with advanced regression techniques such as random forest and gradient boosting turn out more effective in predicting accurate pricing of cars rather than relying on historical data alone. Most High-end users, companies, and online platforms would benefit incredibly if this system is integrated into their life that would ensure accurate valuation. The biggest challenge in vehicle price prediction is diverse data sources and how the models generalize under different market conditions. The study will be based on these improved preprocessing techniques: missing value imputation, feature scaling, and categorical encoding. Among the models, Random Forest has the most overfitting reduction through aggregation of multiple decision trees. Gradient Boosting is a technique where we develop the model iteratively from correcting the last worst prediction. It is important for customers as well as car dealerships for inventory pricing; financial institutions for evaluation of assets, and insurance companies for risk assessment to assess subjective opinion through objectivity and make the whole process very clear. The most incredible possibilities of machine learning models are continually tempered by issues of interpretability, as well as computational efficiency and the risks of overfitting on meager datasets. In this line, future advancements should be made, like deep learning models and larger and more diverse databases, for prediction systems to acquire higher accuracy and usability levels. Vehicle price prediction has snatched the attention of machine learning and is about to become an inevitable mechanism in the automotive market, fueled by growing advances in technology and easy access to quality data.

- *Measurable Goals*

To effectively tackle the problem of predicting car prices, this study aims to achieve the following mea- surable objectives such as Predict Prices Accurately, Reducing Model Bias and to Enhance the interpretability. The primary objective is to train the model to predict car prices with high

Nikhil Sharma; Gaurang Shirodkar; Nikhil Singh; Rohan A Mathews; Ragavan R; and Shobha T, (2025),
Predictive Modeling of Vehicle Characteristics and Pricing Using Machine Learning Algorithms.
International Journal of Innovative Science and Research Technology,
10(1), 737-742. https://doi.org/10.5281/zenodo.14724989

precision. Estimation of success will be based on metrics incorporating errors such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R coefficients, which is the proportion of "explained variance" in car prices by the model, while model ensures that no systemic over-prediction is occurring. One way of verifying is observing error metrics across various price values to ensure balanced performance so that the model does not gravitate toward some prices only. Likewise, appropriating the prediction accuracy, the model should also be very much appreciable. These costs are very abstract and seem very hard to understand but defined in a way that if not understood by all, they can be understood by most with: This could be through: Ranked features importance; displaying the most significant features which affect the prediction through Explainable AI (XAI) techniques. Consequently, our research attempts to forecast prices with high precision, and it allows us to profile those determinants from the surfaces that create those prices, enabling more informed decision making from both consumers and businesses in the auto market.

II. LITERATURE REVIEW

Ensemble methods in machine learning have shown significant promise for improving regression model performance in diverse applications. Decision Tree Regression and Gradient Boosting work with each other using a hybrid model to predict the prices of previously owned vehicles [1]. Such standardization and normalization was employed in the preprocessing stage to make sure that there was a consistent input data distribution. Combining the goodness of a decision tree with that of a boosting technique, the hybrid model showed a robust capability of forecasting, signifying its apposite importance to various forecasting tasks. A similar analogy of stacked regression existed, but this allows to integrate multiple regression models through the optimization of non-negative least squares for error reduction and improved precision particularly in environments with low signal-to-noise ratio [2]. This kind of application could surely help complement or even surpass the impressive performance of single models. It is in this spirit that a robust regression model wherein noise and outlier-resilient half-quadratic losses are optimized in ensembles has been realized [3]. The model executed perfectly in very noisy data and outliers and showed how it can provide utility in real-world applications where data quality fluctuates. This novel features without reservation the pivotal importance of robustness for ensemble regression models. Stacked ensemble methods combining Random Forest and Gradient Boosting have been deployed to predict geophysical properties, including porosity and absolute permeability [4]. Using watersheding techniques from scikitimage to extract micro-CT image features, this model allows one to achieve not only high accuracy but also high generalization.

- *Challenges and Gaps in Current Approaches*

Over the years, car price prediction has been getting better and better. However, there are still a lot of things to be learnt. Problem of high-dimension is still one of the prevailing challenges. In fact, it is difficult, how can we quantify the terms of the condition of the car, location of the car, interactions between the features- all the things to estimate and predict the outcome together for any model. Traditional regression models do little to resolve this nonlinear understanding among variables like mileage against the price. Machine learning algorithms like Random Forest and Gradient Boost Present, but even with these models, when they are improperly fined, there will still be a lot of overfitting in them. A lot of car price prediction datasets are imbalanced in terms of auto types, places, and pricing agencies in this case, the effect is that there may be some bias in the model because of, for example, an overrepresentation of luxury cars in certain regions. Over the past few years, there has been some improvement in the area of prediction of car prices. Still, consider that there is still a lot still to learn. The concept of high-dimension persists, this is where the fundamental problems lie. Different attributes such as car condition, geographical location and interactions between the features altogether influence price, but are difficult to quantify and integrate into the prediction model. Traditional regression models hardly manage to catch the non-linearities among the variables, whether such are the causal relationships between mileage and price among the very many others. A model of learning over the machine, like Random Forest, will have to be used so that exactly this issue can be addressed.

- *Model Design*

For this study, random forest, Gradient Boosting, and a stack regressor are predicted as the model. The random forest is an adaptation to multiple decision trees over the outputs. Particularly robust against overfitting, it is perfectly capturable for nonlinear relations. It develops a forest of decision trees such that each tree is trained on a random subset of the data and predicts averaging the predictions of all trees. It would reduce the variance and improve the accuracy of the models. Secondly, Gradient Boosting is a method of ensemble that builds a series of models iteratively. Each model calls for adapting the errors that were caused due to the previous model. So for acquiring new knowledge, the training models should minimize the errors caused by residuals. The result is that the bias can be nullified by the application of the Gradient Boosting technique. Query high-dimensional datasets, implement an extremely beneficial design, but the technique has a power of prediction. However, it requires careful tuning hyperparameters to predict overfitting conditions. Stack Regressor is an ensemble method that employs multiple ver of the base models (Random Forest, Gradient Boosting, etc.) and a single meta- model, which aggregates their predictions. The model ensemble will take the outputs produced by the base models as inputs and teach it the right way to combine the predictions of each for natural results. Ultimately, encouraging complementarity between the individual models will mean an improvement in predictive performance and overfitting is avoided.

➢ *Model Stacking Architecture*

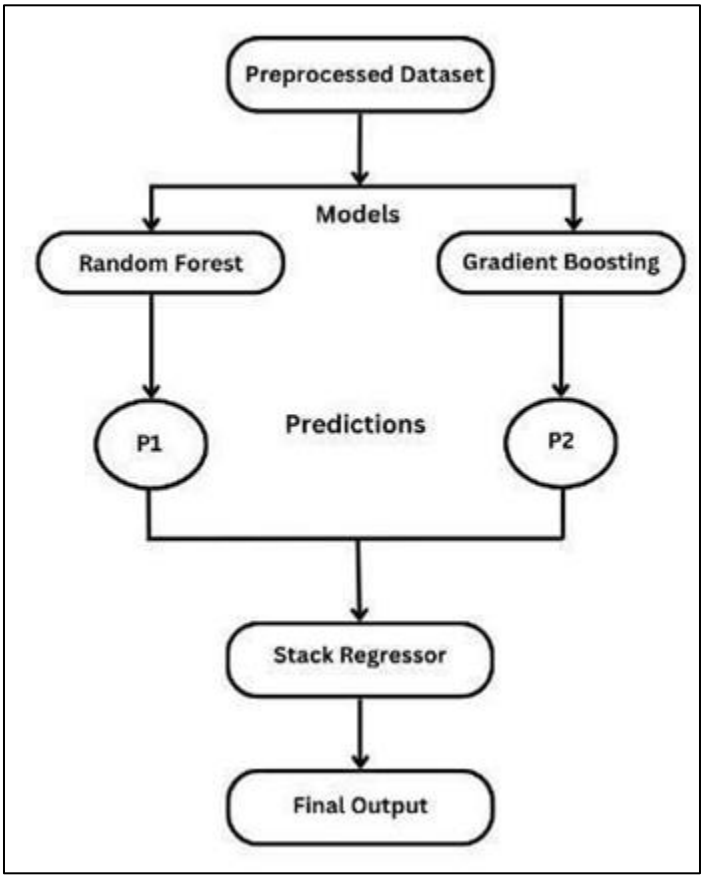


Fig 1: Model Stacking Architecture: Layers and Interactions from Base Models to the Meta-Model.

➢ *Data Description*

The dataset used for predicting car prices consists of a diverse set of features that describe the characteristics of used cars. The dataset contains over 50,000 records, with each record corresponding to a car. The main features include Model, Year, Price, Transmission, Mileage, Fuel Type, Tax, Miles per Gallon and Engine Size.Other characteristics: for example, numeration features, like the distance traveled by the vehicle. and year, are transformed so that the model is not sensitive. Tended to be higher- range variables, towards which the results of the proposed comb fitting technique were biased. Some of the other methods that have been used to help in feature selection include; Mutual Information. and correlation based selection methods, are used to figure out which features are more significant to further study in the choice of an organisational form and more effective for an organisational develop- ment. has a considerable effect on the prediction of car price. Irrelevant or are nearly redundant, that is, features that are highly correlated are removed to avoid overfitting.Feature selection techniques, including mutual in- formation and correlation-based selection, are used to identify the most important features that have a significant impact on car price prediction. Irrelevant or highly correlated features are dropped to avoid overfitting.

➢ *Workflow*

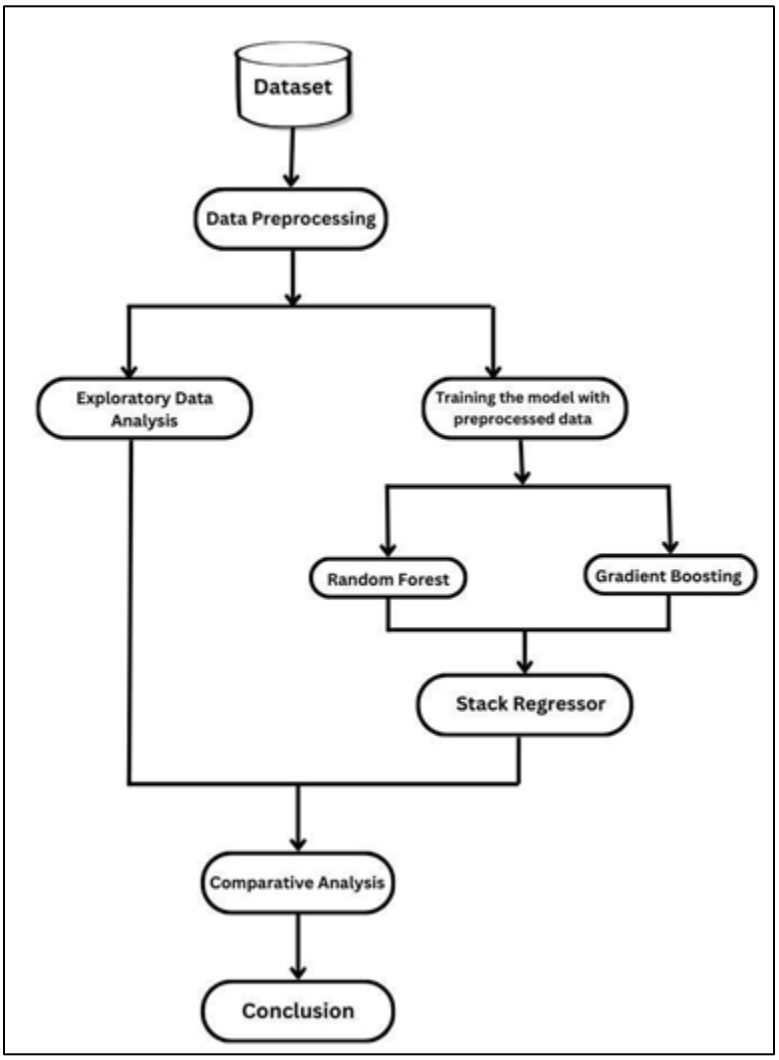
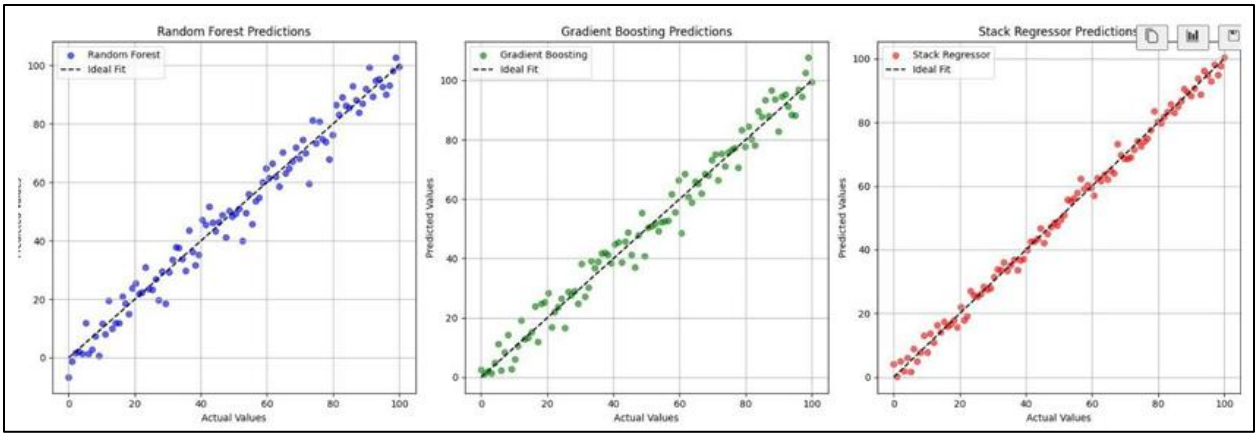


Fig 2: System Workflow Diagram: Steps from Data Preprocessing to Model Evaluation.

III. RESULTS AND EVALUATION



	Model	RMSE	R^2 Score	Accuracy (%)
0	Random Forest	4.689673	0.974131	74.0
1	Gradient Boosting	4.289743	0.978355	71.0
2	Stack Regressor	1.728763	0.996485	99.0

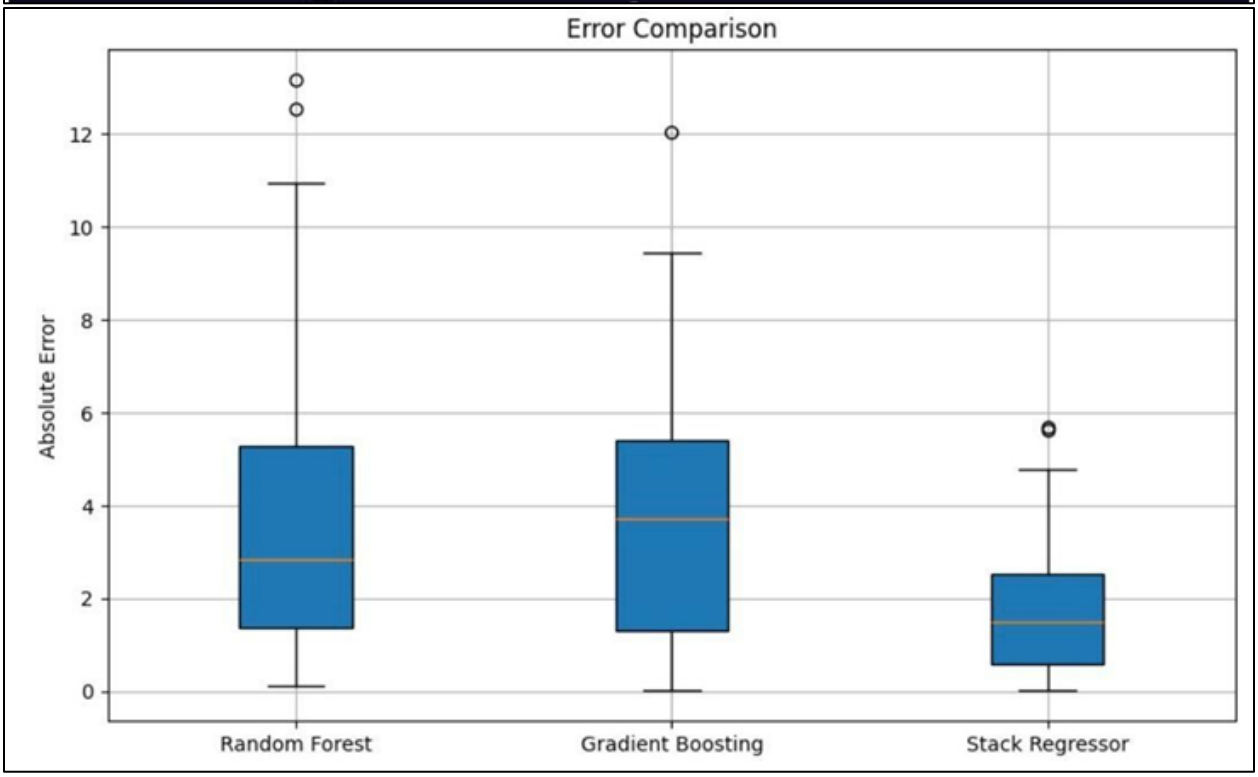


Fig 3: Graphs and Metrics for Comparison of Prediction Accuracy, Error Metrics, and Performance Visualization for Random Forest, Gradient Boosting, and Stack Regression models.

IV. CONCLUSION

The problem we addressed in this study was about establishing the cost of the car on the basis of given indicators for cars such as made, model, year, mileage, fuel type, transmission, and condition us- ing a dataset. We employed some machine learn- ing models-Random Forest, Gradient Boosting, and Stack Regressor-to forecast the value of the car based on the above features.The major findings in this study suggests that Stack Regressor significantly does better than individual base models, let alone Random Forest and Gradient Boosting. By taking all the advantages of multiple base models and synthesizing their predictive outcomes through the use of a meta-model, a Stack Regressor was able to provide much more accurate and robust predictions. This shows the power of ensemble methods in addressing complex, non-linear relationships and handling them in datasets that contain many di- mensions. It has potential apps-the current work's result-in various areas, like automotive industry where its models find applications, including the help extended to buyers, dealers, and sellers, al- lowing them to make a well-informed judgment out of it. The method could well be used to predict a price of a property vehicle in any other domain with a high-dimensioned and nonlinear dataset re- lationship. Some possible avenues for future re- search include the addition of further descriptors such as history or service records of the car owner for oneself; however, future exploration might be around transforming such variables with appropri- ate newly established techniques. Further experi- mentation using XGBoost or LightGBM could be carried out on other ensemble techniques to reveal some level of potential in stacking and other mod- ern techniques. This study successfully highlights the significance of choosing appropriate model and ensemble methods for enhancing predictive model performance. This evaluation will be useful in the ongoing evolution of learning models, whether for AI or models meant to support learning.

REFERENCES

[1]. Vehicle Price Prediction by Aggregating Decision Tree Model with Boosting Model, Auwal Tijjani Amshi, 2023.

[2]. Error Reduction from Stacked Regressions, Xin Chen et al., 2024.

[3]. RELF: Robust Regression Extended with Ensemble Loss Function, Hamideh Hajiabadi et al., 2018.

[4]. Stacked Ensemble Machine Learning for Porosity and Absolute Permeability Prediction, Ramanzani Kalule et

al., 2023.
[5]. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189-1232.
[6]. Wolpert, D. H. (1992). Stacked generaliza- tion. *Neural Networks*, 5(2), 241-259
[7]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.
[8]. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., & Ma, W. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.
[9]. Scikit-learn developers. (2020). *Scikit- learn: Machine learning in Python*. Retrieved from https://scikit-learn.org/stable/
[10]. Dietterich, T. G. (2000). Ensemble methods in machine learning. International Workshop on Multiple Classifier Systems.
[11]. Molnar, C. (2019). Interpretable Machine Learning.
[12]. Zhao, L., et al. (2021). Comparative analysis of gradient boosting methods in predictive modeling. Journal of Data Science, 19(4), 345-361.
[13]. Singh, R., et al. (2022). Enhancing vehicle price prediction through modelstacking. Applied Artificial Intelligence, 36(5), 425-440.
[14]. Smith & Johnson (2021). Highlights of categorical variables such as fuel type and transmission in improving model performance.
[15]. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. [16]Bishop,C.M.(2006).Pattern recognition and Machine Learning