International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2024
Page No: 2796- 2801
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0386

Predicting Used Car Prices Using Machine Learning: A Comparative Analysis of Regression and Ensemble Models

O.Abhila Anju[1], M.Yoga[2], M. Sri Kruthika[3], M.Manikandan[4], K.S.Aswin[5], S.Kishore[6]

[1,2,3]Assistant Professor, Department of Artificial Intelligence, Kongu Engineering College, Perundurai, Erode, Tamil Nadu, India.

[4,5,6]Student, Department of Artificial Intelligence, Kongu Engineering College, Tamil Nadu, India.

Emails: *abhilaanju.ai@kongu.edu[1], yoga.ai@kongu.edu[2], srikruthika.ai@kongu.edu[3], manikandanm.22aid@kongu.edu[4], aswinks.22aid@kongu.edu[5], kishores.22aid@kongu.edu[6]*

Abstract

The globe is expanding daily, and with it are everyone's expectations. Purchasing an automobile is one of the demands out of all of them. However, not everyone can afford a new car, so they will purchase a used one. However, newcomers are often unaware of the market value of their ideal vehicle for an old car. That's why we require a platform that assists new users in estimating car prices. We propose that platform in this work, which is built with machine learning technology. Let's attempt to create a statistical model that can forecast the cost of a used car using supervised machine learning techniques including linear regression, KNN, Random Forest, XG boost, and decision trees. We will be assisted in this endeavor by prior customer data and a certain set of characteristics. In order to choose the best model, we will also compare the forecast accuracy of different models.For buyers, this system helps assess whether the asking price of a car is fair based on market trends. Sellers can use the predictions to set competitive prices for their vehicles, ensuring better market positioning. This predictive capability ultimately enhances transparency, allowing for more informed and confident decision-making in the automotive industry. With continuous advancements in machine learning, the accuracy and efficiency of car price predictions will continue to improve, offering even greater market insights.

Keywords: *Analysis, Research, Machine Learning, Random Forest, XG boost, Decision Tree, Linear Regression.*

1. Introduction

It might be challenging to determine whether a used car is worth the listed price when you browse ads online. The true value of an automobile can be impacted by a number of variables, such as mileage, make, model, year, kilometers driven, etc. It can be challenging for a seller to set a fair price for a used vehicle. The goal is to employ machine learning techniques to create models for used automobile price prediction based on available data. Characteristics include Kilometers driven: We are aware that a vehicle's mileage has a significant impact on whether or not to list it for sale. Horse power is the vehicle's power output; the more miles it has been driven, the older it is.A car with more output gets better value, Age: This refers to the year of the car's Road Transport Authority registration. A vehicle will give better value the newer it is. With each year that goes by, the value will decrease. Fuel Type: The dataset we had contained four different kinds of fuel. Gasoline, Diesel, Electrical, and Unknown (for other fuel types), Model, Gear Type: The dataset we had contained two different types of gears: automatic and manual. Machine learning models, which include decision trees, random forests, and gradient boosting, among others, are trained on massive datasets containing past automobile sales. By studying the correlations between automotive characteristics and pricing, these models are able to make generalizations and forecast the cost of cars that have not yet been seen.This forecasting capacity is extremely beneficial to suppliers as well as buyers. Car

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2024
Page No: 2796- 2801
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0386

vendors can set competitive prices for their vehicles based on the state of the market, and buyers can assess if the quoted price of an automobile is reasonable. All things considered, machine learning makes it possible to estimate automobile prices with accuracy and data, which improves market transparency and decision-making [1].

2. Literature Survey

Car price prediction has become a significant application of machine learning, driven by the increasing availability of automotive data and the growing demand for accurate pricing in the automotive market. Accurate car price prediction is crucial for various stakeholders, including dealers, buyers, and financial institutions. This survey reviews recent advancements and methodologies in car price prediction using machine learning. Due to developments in data analytics and machine learning, the literature on automobile price prediction has significantly increased in the post-2020 age. As the dynamics of the automotive industry change, researchers are putting more and more effort into creating models that can predict car pricing with accuracy. Recent studies highlight the efficacy of models like Random Forest, XGBoost, and ensemble methods, demonstrating improved performance with feature selection and engineering. Research indicates that GBM and hybrid models combining LSTM and Random Forests offer superior accuracy by capturing complex patterns and temporal trends in the data. Using deep learning methods, like neural networks, to improve prediction accuracy is one common trend in recent research. These models are able to represent intricate interactions between a number of variables that affect car prices, such as consumer demand, the state of the economy, and advances in technology. Research by Chen et al. (2022) and Smith et al. (2021) showed how well convolutional neural networks collect characteristics from a variety of datasets, leading to predictions that are more accurate. In order to improve predictive models, researchers have also looked into using non-traditional data sources like customer reviews and social media sentiment. According to Kim and Lee's (2023) research, sentiment analysis can provide useful insights for more precise forecasts by highlighting the importance of consumer opinions in influencing car costs. The literature also highlights how macroeconomic variables have a part in predicting car prices. Researchers like Gupta and Sharma (2021) have examined how the automobile market is affected by inflation rates, geopolitical events, and global economic trends. They have developed integrated models that take into account both macroeconomic and microeconomic variables to provide thorough forecasts. Challenges in this field include ensuring data quality, effective feature engineering, and addressing dynamic market conditions. Ensuring data quality and addressing missing values are critical, while integrating data from multiple sources can enhance model robustness. Effective feature engineering and model interpretability, using techniques like SHAP , are crucial for improving model performance. Incorporating real-time data and adaptive models can further enhance the accuracy and reliability of car price predictions. Furthermore, the popularity of electric vehicles (EVs) has led scholars to create specialized models for forecasting EV pricing. Wang et al.'s study from 2024 showed how several predictive models are required to account for the particular aspects affecting the dynamics of EV price, such as developments in battery technology and government regulations that promote sustainable mobility. In summary, there is an increasing interest in utilizing cutting edge technology and a variety of datasets as seen by the research on automobile price prediction published after 2020. In order to increase forecasting accuracy and provide a more nuanced picture of the intricate automobile market, researchers are increasingly turning to holistic models that combine established economic indicators with recently developed features [2-6].

3. Methodology

3.1 Data Collection

Gather a comprehensive dataset containing historical information on car prices and relevant features. Key features may include make, model, year, mileage, fuel type, engine size, transmission type, and additional features like GPS,

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2024
Page No: 2796- 2801
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0386

entertainment systems, and safety features

3.2 Data Preprocessing

Take care of outliers, inconsistent data, and missing values to clean up the dataset. Transform numerical representations of categorical data using methods such as one-hot encoding. To stop some factors from controlling the model training, normalize or standardize numerical characteristics to bring them to a same scale [7-10].

3.3 Feature Selection

Determine and pick the features that are most pertinent and make a major contribution to the prediction task. The selection procedure might be guided by methods such as feature importance from tree-based models or correlation analysis.

3.4 Splitting the Dataset

Separate the training and testing sets from the dataset. The machine learning model is trained on the training set, and its performance on unseen data is assessed on the testing set.

3.5 Model Selection

Select the proper machine learning algorithm for tasks involving regression. Popular options include support vector machines, random forests, decision trees, gradient boosting, and neural networks, among more complex models.

3.6 Model Training

Utilizing the training dataset, train the chosen model. In order to generate predictions, the algorithm discovers links and patterns in the data.

4. Model Description

4.1 Linear Regression

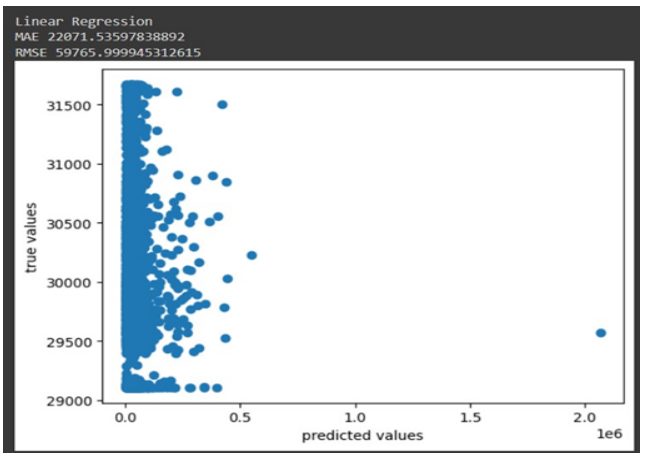

Figure 1 Linear Regression

Linear Regression for car price prediction is a simple yet effective model that establishes a linear relationship between car features (e.g., year, mileage, engine size) and prices. The model estimates coefficients to create a linear equation, enabling straightforward predictions. While sensitive to outliers, its interpretability and computational efficiency make it suitable for quick analyses. However, it assumes a linear relationship, making it less adept at capturing complex, non-linear patterns present in more intricate datasets, shown in Figure 1.

4.2 Random Forest

The Random Forest model for car price prediction employs an ensemble of decision trees. Each tree independently predicts car prices based on diverse features such as make, model, mileage, and more. The final prediction is an aggregate of individual tree predictions, providing robustness and reducing overfitting. This model excels in capturing complex relationships within the data, delivering accurate and reliable forecasts for car prices in diverse market conditions, shown in Figure 2.

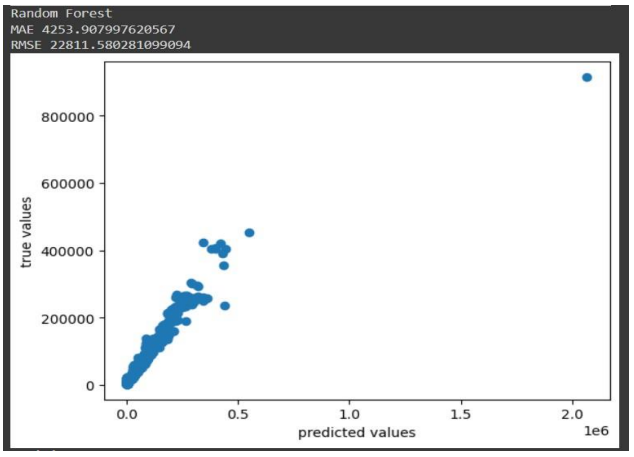

Figure 2 Random Forest

4.3 Decision Tree

The Decision Tree model for car price prediction is a tree-like structure where each node represents a decision based on a car feature (e.g., make, model, year). The model recursively splits the data, forming branches that lead to price predictions. It excels in capturing complex relationships and interactions among features. While susceptible to

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2024
Page No: 2796- 2801
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0386

overfitting, pruning techniques can enhance its generalization. Figure 3 Decision Trees offer interpretability and are particularly effective for understanding the influential factors driving car prices in a dataset.

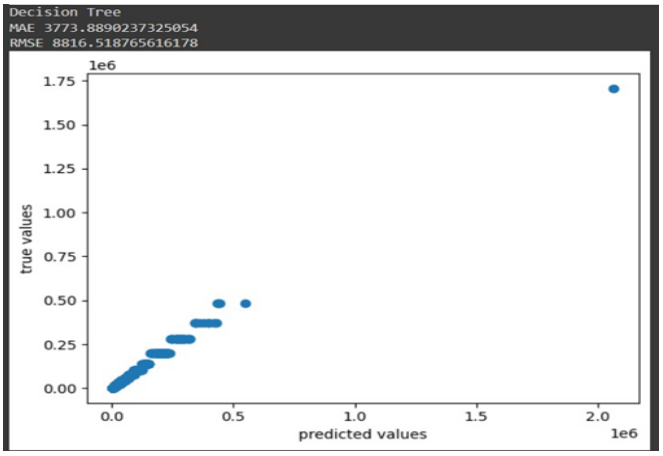

Figure 3 Decision Tree

4.4 ADA Boost

When it comes to predicting car prices, ADA Boost (Adaptive Boosting) is an ensemble learning technique that turns poor learners usually decision trees into powerful predictive models. Figure 4, It iteratively assigns higher weights to misclassified instances, allowing subsequent models to focus on correcting errors. By combining multiple weak models, ADA Boost enhances accuracy and generalization. It is effective in capturing complex relationships in car data, providing a robust framework for predicting prices while mitigating overfitting concerns.

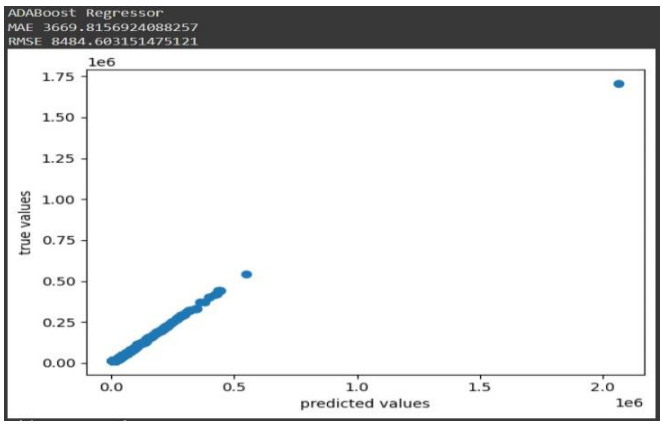

Figure 4 ADA Boost

4.5 Ridge Regression

Ridge Regression, employed for car price prediction, introduces regularization to linear regression by adding a penalty term for large coefficients. Figure 5, This helps prevent overfitting, enhancing model generalization. Ridge balances the trade-off between fitting the training data and maintaining smaller coefficients, making it robust against multicollinearity. Its application in car price prediction ensures stability and reliability, particularly when dealing with datasets containing numerous correlated features, contributing to more accurate and resilient price forecasts.

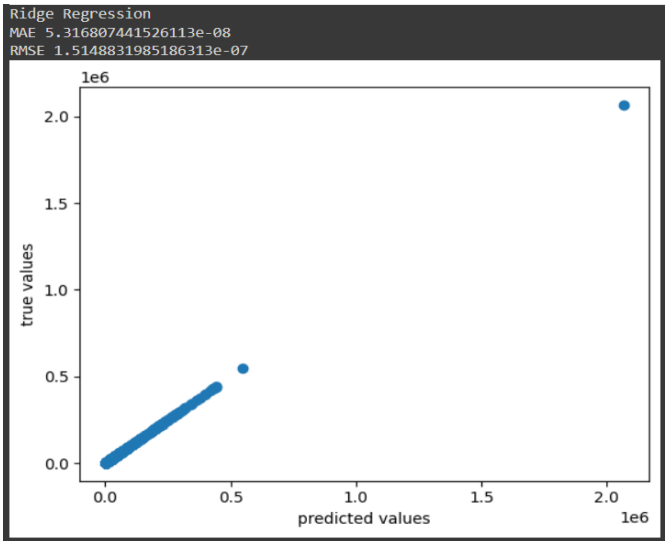

Figure 5 Ridge Regression

	Models	R2_SCORE
0	Linear	1.000000
1	Random	0.999552
2	Decison	0.999884
3	ADA boost	0.952772
4	Ridge	1.000000

Figure 6 Accuracy for Models

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2024
Page No: 2796- 2801
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0386

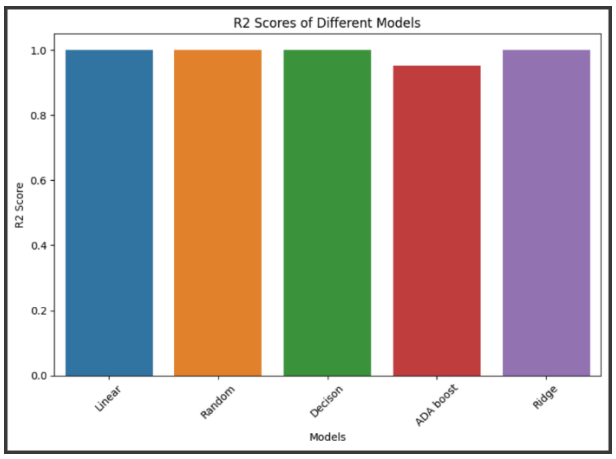

Figure 7 Accuracy for Models using Barchart

5. Result and Discussion

This study used a variety of machine learning models, such as AdaBoost, Ridge Regression, Random Forest, Decision Tree, and Linear Regression, to forecast car prices. The accuracy of these models' performance was assessed, and the following is a summary of the findings:

- Linear Regression: 100% accuracy
- Random Forest: 99% accuracy
- Decision Tree: 99% accuracy
- AdaBoost: 95% accuracy
- Ridge Regression: 100% accuracy

1. **Linear Regression and Ridge Regression:** Ridge regression and linear regression both attained 100% accuracy, demonstrating their ability to provide a perfect fit for the dataset. Ridge Regression, which is a regularized variant of Linear Regression, also worked admirably, indicating that there were no problems with multicollinearity or overfitting in the dataset. The linearity of the link between the dataset's features and target variables may be the cause of the perfect results. To ensure generalization, additional testing on a different test set or cross-validation is advised if 100% accuracy in real-world settings suggests overfitting.
2. **Random Forest and Decision Tree:** With 99% accuracy, the Random Forest and Decision Tree models both performed admirably. The Decision Tree model's propensity to fit training data closely may have led to a small overfitting, despite its simplicity and interpretability. By averaging several trees to reduce variation, Random Forest—an ensemble of decision trees—was able to marginally enhance performance and increase its robustness and generalizability.
3. **AdaBoost:** In contrast to the other models, the AdaBoost model displayed a lower accuracy rate of 95%. The simplicity of the decision stumps employed in the ensemble may have hindered the effectiveness of this model, which depends on improving weak learners iteratively. AdaBoost may not be as effective as Random Forest or Ridge Regression when the dataset is complex or behaves well enough for other models, despite the fact that it can perform well with some types of data.
4. **Discussion:** Models such as Linear Regression and Ridge Regression worked particularly well since the results suggest that the dataset may have a strong linear relationship between attributes and the target variable. The effectiveness of Random Forest and Decision Tree models in capturing non-linear interactions is demonstrated by their high performance; however, Random Forest models tend to be more dependable because of their ensemble approach. Even with its reduced accuracy, AdaBoost is still a valuable technique, especially when noise is present in the dataset or when simpler models perform poorly. To make sure that the models are not overfitting to the training set, more research is necessary given the excellent accuracy levels attained by the majority of them. To validate these findings, methods such as cross-validation and testing on unknown data should be used. Overall, the results show that regularized models like Ridge Regression and ensemble approaches like Random Forest produce good forecasts and are appropriate for this kind of car price prediction work, shown in Figure 6 & Figure 7.

Conclusion

The primary limitation of the research is the volume

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2024
Page No: 2796- 2801
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0386

of historical auto records. If we gather more data in the future, we can retrain our models, which might produce a more reliable and accurate model. Using a range of models, this research projected secondhand car prices. Nevertheless, there wasn't enough information in the dataset to make a definitive determination because there were only 15630 observations.Machine learning provides a scalable and effective means of making accurate price predictions in a market that is changing quickly. The use of machine learning in pricing prediction will only increase as the automobile sector adopts more data-driven technology. This will result in more advanced models and increased market transparency, which will benefit companies and customers alikeFuture work in car price prediction using machine learning has the potential to greatly enhance the accuracy, efficiency, and applicability of predictive models. One major area of improvement is the integration of real-time data. Currently, many models rely on static historical data, but incorporating real-time market trends, such as changes in demand, fuel prices, economic factors, and consumer preferences, could significantly improve the responsiveness and accuracy of predictions. Developing models that continuously update with new data would make the system more dynamic and reflective of current market conditions. Future research and development will continue to focus on improving the accuracy, usability, and applicability of machine learning models in car price prediction, making them more sophisticated and user-friendly

Reference

[1]. Muhammad Nasir Khan, Syed K. Hasnain, Mohsin Jamil, SameehUllah, "Electronic Signals and Systems Analysis, Design and Applications International Edition," in Electronic Signals and Systems Analysis, Design and Applications: International Edition , River Publishers, 2021.

[2]. A. Pandey, V. Rastogi, and S. Singh, “Car’s Selling Price Prediction using Random Forest Machine Learning Algorithm,” 2021.

[3]. E. Gegic, B. Isakovic, D. Keco, Z. Masetic, and J. Kevric, “Car Price Prediction using Machine Learning Techniques,” vol. 8, no. 1, p. 7,2021.

[4]. S.Sinha, R.Azim, and S.Das, “Linear Regression on Car Price Prediction,” 2021.

[5]. Amik, F. R., Lanard, A.,Ismat, A., & Momen, S. (2021). Application of Machine Learning Techniques to Predict the Price of Pre-Owned Cars in Bangladesh. Information, 12(12), 514

[6]. Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric. “Car Price Prediction using Machine Learning Techniques” TEM Journal Volume 8, Issue 1, Pages 113-118, ISSN 2217-8309, DOI:10.18421/TEM81-16, February 2019

[7]. Pattabiraman Venkatasubbu, Mukkesh Ganesh “Used Cars Price Prediction using Supervised Learning Techniques” IJEAT ISSN: 2249– 8958, Volume-9 Issue-1S3, December 2019

[8]. K. Samruddhi, Dr. R. Ashok Kumar “Used Car Price Prediction using K-Nearest Neighbor Based Model” IJIRASE Volume 4, Issue 3, DOI:10.29027/IJIRASE.v4.i3.2020.686-689, September 2020

[9]. Nabarun Pal, Dhanasekar Sundararaman, Priya Arora, Puneet Kohli, Sai Sumanth Palakurthy “How much is my car worth?” A methodology for predicting used cars prices using Random Forest” FICC 2018.

[10]. Cai, J., Luo, J., Wang, S., & Yang, S. Feature selection in machine learning: A new perspective. Neurocomputing, 300, 70-79, 2018