

Car Price Prediction Using Machine Learning

Rui Chen ^a

Department of Statistics and Applied Probability, University of California, Santa Barbara, California, U.S.A.

Keywords: Machine Learning, Price Prediction.

Abstract: Car price prediction is always a critical issue for the car industry, with significant implications for consumers, dealers, and manufacturers. This study aims to compare the performance of three machine learning models - Linear Regression (LR), Decision Tree (DT), and K-nearest neighbors (KNN) - to predict car prices using a comprehensive dataset of vehicle characteristics. The research integrates diverse methodologies, evaluates model performance using the R-squared (R^2) metric, and discusses the implications for practical applications. The DT model, enhanced by feature importance analysis, achieved the highest R^2 on the test set, indicating its strong ability to capture complex patterns within the data. These findings underscore the potential of advanced machine learning techniques to provide more accurate and reliable pricing models. By improving price predictions, this research can support stakeholders in developing more effective pricing strategies, ultimately benefiting consumers with fairer prices and helping dealers and manufacturers optimize their revenue and inventory management.

1 INTRODUCTION

In the rapidly evolving automotive market, the precise prediction of car prices remains a critical challenge with significant implications for consumers, dealers, and manufacturers alike. As consumers seek the best value for their purchases, and dealers aim to optimize inventory and pricing strategies, the need for precise valuation models becomes increasingly evident. This research is motivated by the impact that accurate pricing can have on consumer satisfaction and business profitability.


The automotive industry relies heavily on accurate pricing strategies to ensure fair market value for vehicles, optimize inventory management, and enhance consumer trust. Traditional pricing models often fail to account for the intricate dependencies among variables like vehicle age, mileage, brand, and condition, leading to suboptimal pricing strategies. However, Machine learning has opened up new avenues for increasing the precision of auto price forecasts. Based on a reliable dataset of diverse vehicle characteristics, this study aims to examine and contrast the accuracy of Linear Regression (LR), Decision Tree (DT), and K-nearest neighbors (KNN)

models in forecasting automobile pricing. By analyzing a robust dataset, this research contributes to demonstrating the strengths and limitations of these models in real-world scenarios, providing a comparative analysis that will serve as a valuable reference for consumers, dealers and industry manufacturers.

2 LITERATURE REVIEW

Making use of machine learning to anticipate vehicle prices has been the focus of numerous studies, each adding unique methodologies and reasoning about the variables influencing automobile pricing and the precision of various machine learning models used to make predictions. In the automotive industry, machine learning has become a useful tool that assists dealers and consumers in making informed decisions.

The car industry is a major pillar of global economies, significantly contributing to the gross domestic product (GDP) of many nations. Several studies have explored the relationship between automotive production and economic stability. One such study examined the evolution of motor vehicle production across various continents from 2018 to

^a <https://orcid.org/0009-0008-2487-2862>

2022, highlighting how the industry is deeply impacted by macroeconomic factors such as the pandemic. The research demonstrated that the automotive industry fared better in the pre-pandemic period compared to the pandemic period (Toma, 2023).

As one of the largest sectors in terms of employment and technological innovation, the industry's impact extends far beyond the vehicles it produces. The integration of modern supply chain practices, such as modular procurement, has made the automotive sector a driver of global economic development. However, the COVID-19 pandemic disrupted these supply chains, leading to shortages of key components, such as semiconductors, which has further impacted car prices (Radić. N & Radić. V, 2021). Another study emphasized the ripple effects of supply chain disruptions in the automotive industry, underscoring how these disruptions not only led to a decline in production but also contributed to price volatility in the market (Asghar et al., 2021). As the global economy recovers from the effects of the pandemic, the used car market has experienced rapid growth. Many buyers, unable to afford new vehicles, have turned to used cars as a more affordable option. According to research, the surge in used car demand, spurred by the shortage of new cars and a rise in consumer purchasing power, has driven up the prices of used cars (Das Adhikary et al., 2022). With the rising demand, car sellers have taken advantage by listing vehicles at inflated prices, further emphasizing the need for accurate car price prediction models to help buyers make informed decisions.

A number of machine learning algorithms, each with specific advantages and disadvantages, have been used to forecast automobile values, helping buyers and sellers evaluate vehicles more accurately. According to a study, LR models are useful for estimating the cost of used automobiles and emphasize the significance of a vehicle's attributes such as its make, model, condition, and mileage (Muti & Yıldız, 2023). LR, while effective, often struggles with non-linear data patterns, leading researchers to explore more advanced models like Random Forests (RF) and DT for better accuracy. Another study used Random Forest models with more than 200 DTs to predict used car prices, achieving high accuracy rates (Ranjith, 2021). The study showed that because RF can handle complicated interactions between variables and many features, it performs better than other regression models. Another study evaluates the increasing complexity of China's used car market by using machine learning models, including LightGBM, to analyze key factors from five datasets,

ultimately constructing a predictive model that enhances used car sales strategies. (Wang et al., 2022). Additionally, a 2022 research introduces an intelligent framework using artificial neural networks to estimate used car prices, outperforming traditional models like random forests in accuracy, as validated with large datasets of U.S. vehicles. (Pillai, 2022). Moreover, ensemble machine learning methods like Random Forest, Support Vector Machine, and Artificial Neural Network were used in a study on Bosnia and Herzegovina's car price prediction. Using this method produced a model with an accuracy of 87.38%. (Gegic et al., 2019).

However, the rise of electric vehicles (EVs) and advancements in digitalization are reshaping the automotive industry. The EU and Germany's focus on the decarbonization of the automotive sector highlights the impact of environmental regulations on car prices. As consumers shift towards EVs, machine learning models must adapt to include variables such as battery life, charging infrastructure, and government incentives (Nettekoven, 2023).

Although there are already various researches on car price prediction, there isn't a comprehensive comparative analysis of different algorithms that can predict car prices. Also, different prediction methods may perform differently in certain situations in reality. Therefore, this paper will use experiments to show which algorithm has the best performance in general and how those algorithms can perform better than each other in different situations.

3 METHODOLOGYS

3.1 Data Description and Preparation

The dataset from Kaggle was used in this investigation because it has a number of attributes, including make, model, horsepower, mileage. The brands and makes of cars might serve as a representation of what consumers choose to purchase in the present day. Because of its richness and significance, it is a good fit for creating reliable prediction models.

The first step in data preparation involved cleaning the dataset, where missing values were addressed through imputation. For continuous variables, median values were used, while mode values were applied to categorical variables. Feature engineering was also performed to enhance model performance by creating new features from existing data, such as categorizing continuous variables like mileage and age into bins.

One-hot encoding was used to change categorical information into a format that machine learning algorithms could understand. To make sure that scale discrepancies wouldn't allow one property to dominate the others, numerical features were standardized.

3.2 Model Development

Three machine learning models were selected for this study. LR served as a baseline, offering simple interpretations of data by modeling a linear associations between car prices and vehicle attributes. DT can manage non-linear association between car features, so it was a perfect fit for capturing the intricate dynamics involved in automotive pricing. KNN was included for its effectiveness in cases where similar historical data points can predict future data points, leveraging feature similarity for price prediction.

4 EXPERIMENT RESULTS AND DISCUSSION

4.1 Experiment Evaluation

To understand how well the models used fit the data, the R-squared (R^2) was used to evaluate those models. R^2 stands for the proportion of dependent features' variance that is predictable from independent features. Higher values in R^2 indicate a better fit between the model and the data. Additionally, scatter plots were generated to visually assess the alignment between the predicted and actual values, offering a more intuitive understanding of the model's performance. Furthermore, to gain deeper insights into how the DT model arrived at its predictions, a feature importance analysis was conducted. This analysis identifies which features had the most influence on predicting car prices. The tools utilized were Python and modules like Pandas, Scikit-learn, Matplotlib. Pandas were used to process data, Scikit-learn was used to create and assess models, and Matplotlib was used to visualize data.

4.2 Model Performance

The models' performance was evaluated based on the R^2 metric, with the DT model achieving the highest R^2 of 0.90. This suggests that the DT model was the most effective at capturing the complex patterns within the dataset, explaining 90% of the variance in

car prices. The LR model, with an R^2 of 0.81, served as a reliable baseline but struggled with the non-linear aspects of the data. The KNN model, with an R^2 of 0.72, was less effective than both the DT and LR models in explaining the variance in car prices. Table 1 shows the results of LR, DT and KNN.

Table 1: R^2 Values for LR, DT and KNN.

| Model | R^2 |
|---------------------|-------|
| Linear Regression | 0.81 |
| Decision Tree | 0.90 |
| K-Nearest Neighbors | 0.72 |

4.3 Visual Analysis

Scatter plots were created to compare the actual values for each model versus the anticipated values in order to better examine the models' performance. The degree to which each model's forecasts matched the actual automobile pricing is shown visually in these plots. Figure 1, Figure 2 and Figure 3 shows the scatter of LR, DT and KNN.

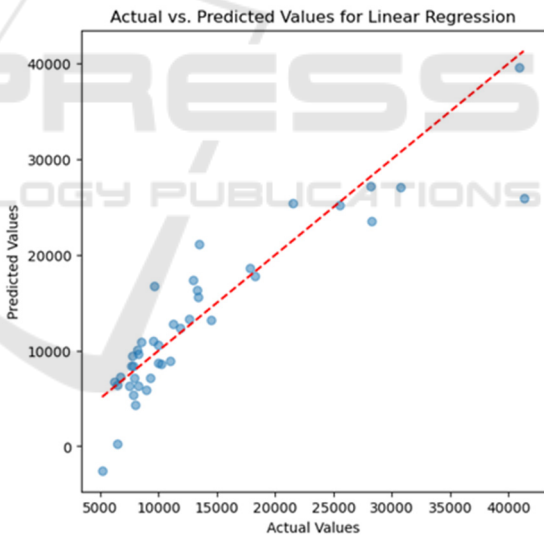


Figure 1: Scatter plots for LR (Photo/Picture credit: Original).

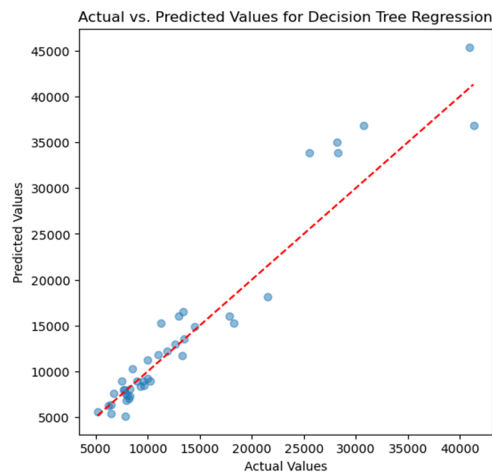


Figure 2: Scatter plots for DT (Photo/Picture credit: Original).

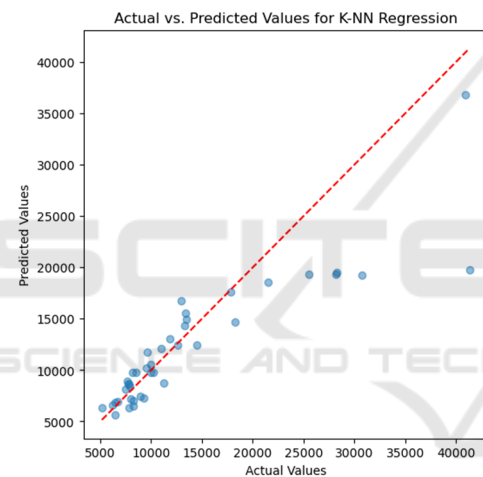


Figure 3: Scatter plots for KNN (Photo/Picture credit: Original).

These plots indicate that the DT model provided predictions that are closely aligned with the actual car prices, as shown by the points clustering along the diagonal line, which indicates a strong fit. The LR and KNN models, while generally effective, exhibited more variability, with predictions deviating more from the actual values, particularly in the case of the KNN model, which showed the greatest dispersion from the diagonal.

4.4 Feature Importance Analysis

In addition to the visual analysis for the three models, a feature importance analysis was conducted for the DT model. The importance of each feature was calculated based on its contribution to the prediction of car prices, as shown in Figure 4.

The results indicate that engine size was the most significant feature, with a feature importance score of approximately 0.65. This aligns with expectations, as engine size typically plays a key role in determining a car's performance and value. Larger engines generally correlate with more powerful and expensive vehicles. Curb weight was the second most important feature, with a score of approximately 0.27. Heavier cars are often associated with luxury models or robust vehicles, contributing to a higher price point. Other features, such as stroke and mileage on the highway, had relatively low importance scores (around 0.02 each), suggesting that while they do contribute to the pricing model, they are not as significant as performance-related features like engine size and curb weight. Several other features, such as car width, car height, horsepower, car peak rpm, and mileage in the city, exhibited negligible importance. This suggests that these characteristics play a minor role in price determination, at least within the context of the dataset used in this study.

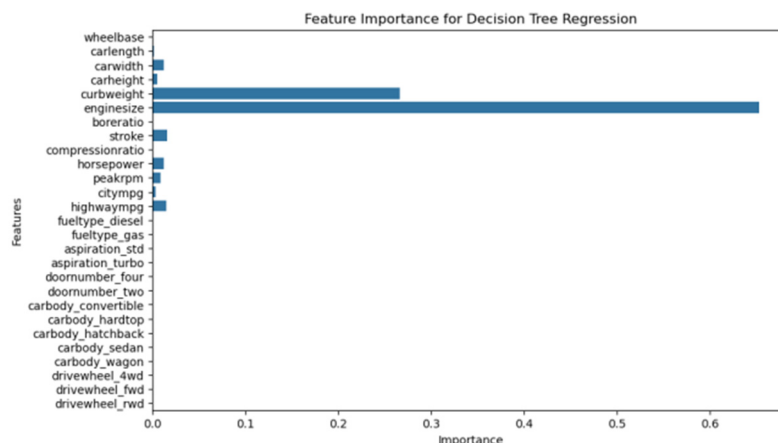


Figure 4: Feature Importance (Photo/Picture credit: Original).

4.5 Practical Implications

The automotive sector can benefit from this study's findings in practice. The DT model works well with dynamic pricing technologies that need to analyze several aspects in real time because of its capacity to manage intricate, non-linear interactions. However, care must be taken to avoid overfitting, especially when using the model for higher car price predictions where residual errors were noted. Additionally, the feature importance analysis suggests that engine size and curb weight should be primary considerations when building pricing models for vehicles, as they account for the majority of the predictive power.

In contrast, the LR model, though simpler, could be beneficial in scenarios where interpretability and processing speed are prioritized over precision. This model may require additional tuning or regularization to improve its generalizability, as it struggles with non-linearity in the data.

The KNN model, while computationally intensive and having a lower R^2 score, offers a robust alternative with better generalization capabilities. KNN is useful in contexts where data consistency is variable, as it leverages the similarity between data points to make predictions. However, it is less effective in capturing the broader trends in the data, as seen in its lower performance relative to the DT model.

5 LIMITATIONS AND FUTURE OUTLOOK

This study's primary limitation is its reliance on a static dataset, which does not account for temporal fluctuations in the automotive market. Car prices are influenced by various external factors, such as economic conditions, changes in consumer preferences, and technological advancements, all of which can change over time. The models developed here do not incorporate temporal dynamics, which could lead to reduced accuracy in real-world applications where these factors play a significant role.

Another limitation is the geographical specificity of the dataset, which may limit the application of the models to other regions with different market conditions. For instance, factors such as local economic conditions, tax regulations, and consumer preferences can vary significantly between regions or countries, making price predictions less accurate.

When it comes to model complexity, the DT and KNN models are prone to overfitting, especially when dealing with small or noisy datasets. When a model is overfitted, it may perform well on training data but not on unknown data. This issue could be mitigated by employing techniques such as pruning in DTs or adjusting the number of neighbors in KNN.

Future research could address these limitations by incorporating time-series analysis to capture temporal changes in car prices. This would involve developing models that can adapt to changes in economic conditions, market trends, and other temporal factors. Additionally, expanding the dataset to include data from multiple regions would enhance the generalizability of the models, allowing them to be applied more broadly.

Moreover, future studies could explore the integration of unstructured data, such as text from online car listings or social media sentiment, into the models. This could provide a more comprehensive understanding of the factors influencing car prices and lead to more accurate predictions.

Finally, investigating more sophisticated machine learning algorithms, such as deep learning, may help the models perform better. Because those more complex algorithms are able to recognize more interactions between variables, more robust predictions can be made.

6 CONCLUSIONS

This paper offers a thorough examination of KNN, DT, and LR in the prediction of vehicle pricing. The DT model emerged as the most effective, demonstrating its ability to capture complex patterns and interactions within the dataset. The study also highlights the practical applications of these models in the automotive industry, particularly in dynamic pricing tools and real-time valuation systems. However, the study also acknowledges its limitations, particularly the lack of temporal dynamics and geographical diversity in the dataset. Future research could address these issues by incorporating time-series analysis and expanding the dataset to include multiple regions. Additionally, integrating unstructured data and applying more advanced machine learning techniques to the data could further enhance the accuracy of the models. The findings of this study contribute to the ongoing development of machine learning models for car price prediction, providing valuable insights for both academic researchers and industry practitioners. By improving the accuracy and robustness of these models, it is

possible to develop more effective pricing strategies, enhance consumer trust, and optimize inventory management in the automotive industry.

REFERENCES

- Asghar, M., Mehmood, K., Yasin, S., Khan, Z. M., 2021. Used cars price prediction using machine learning with optimal features. *Pakistan Journal of Engineering and Technology*, 4(2), 113-119.
- Das Adhikary, D. R., Sahu, R., Pragyna Panda, S., 2022. Prediction of used car prices using machine learning. In *Biologically Inspired Techniques in Many Criteria Decision Making: Proceedings of BITMDM 2021* (pp. 131-140). Singapore: Springer Nature Singapore.
- Gegic, E., Isakovic, B., Keco, D., Masetic, Z., Kevric, J., 2019. Car price prediction using machine learning techniques. *TEM Journal*, 8(1), 113.
- Muti, S., Yıldız, K., 2023. Using linear regression for used car price prediction. *International Journal of Computational and Experimental Science and Engineering*, 9(1), 11-16.
- Nettekoven, Z. M., 2023. Automotive industry transformation and industrial policy in the EU and Germany: A critical perspective (No. 208/2023). Working Paper.
- Pillai, A. S., 2022. A Deep Learning Approach for Used Car Price Prediction. *Journal of Science & Technology*, 3(3), 31-50.
- Radić, N., Radić, V., 2021. Macroeconomic Consequences Caused by the COVID-19 Pandemic—Case Study of the Automotive Industry. *LIMEN* 2021, 79.
- RANJITH, V., 2021. Used Car Price Prediction Using Machine Learning.
- Toma, S. G., 202. The Evolution of the World Motor Vehicle Production in the Period 2018-2022. *Ovidius University Annals, Economic Sciences Series*, 23(2), 175-179.
- Wang, A., Yu, Q., Li, X., Lu, Z., Yu, X., Wang, Z., 2022. Research on Used Car Valuation Problem Based on Machine Learning. In *2022 International Conference on Computer Network, Electronic and Automation (ICCNEA)* (pp. 101-106). IEEE.