IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025

DOI: 10.17148/IJARCCE.2025.1411105

A Data-Centric Review of Predictive Models in Second-Hand Car Valuation

Ashwin Krishna N[1], Dr. G. Paavai Anand[2]

M.Tech., CSE, SRMIST, Vadapalani, Chennai, India[1]

Asst. Professor, CSE, SRMIST, Vadapalani, Chennai, India[2]

Abstract: After 2021, over 90 million passenger automobiles were produced, marking a significant increase in auto production. This growth has led to a flourishing used car market, which has become a highly lucrative sector. One of the most critical and fascinating areas of research within this market is automobile price prediction. Accurate price prediction models can greatly benefit buyers, sellers, and businesses in the used car industry. This paper presents a detailed comparative analysis of two supervised machine learning models: K-Nearest Neighbour and Support Vector Machine regression techniques, to predict used car prices. We utilized a comprehensive dataset of used cars sourced from the Kaggle website for training and testing our models. The K Nearest Neighbour algorithm is known for its simplicity and effectiveness in regression tasks. On the other hand, the Support Vector Machine regression technique uses a different approach, finding the optimal hyperplane that best fits the data. Both methods have their strengths and weaknesses, which we explored in this study. Our results indicated that both KNN and SVM models performed well in predicting used car prices, but with slight variations in accuracy. Consequently, the suggested models fit as the optimum models and have an accuracy of about 83 percent for KNN and 80 percent for SVM. The results indicate that the KNN model slightly outperforms the SVM model in predicting used car prices

Keywords: K Nearest Neighbour, Machine Learning, Prediction, Support Vector Machine, Used Cars Accuracy.

I. INTRODUCTION

The car industry has experienced earth-shattering improvement over the past decade, coming full circle inside the era of over 70 million traveller vehicles in 2021 alone. This surge led to a booming new car market but also contributed to a growing but has additionally given rise to a energetic and expanding assistant promote for utilized automobiles. As the utilized car promote flourishes, accurately anticipating vehicle costs has gotten to be a significant locale of interested for both buyers and merchants. Generally, vehicle fetched desire depended on straight backslide models that, while coordinate, frequently fought to capture the complex, non-linear associations characteristic in assessing data. These models, grounded in bona fide taken a toll data and basic highlights such as mileage, age, and condition, as regularly as conceivable required precision when associated to complex and wide datasets. In afterward a long time, the field has seen a vital move towards the application of machine learning strategies, which offer the potential to overhaul figure accuracy by managing with non-linear plans and large-scale data more suitably. Among these methods, the K-Nearest Neighbour calculation and Reinforce Vector Machine backslide have accumulated noteworthy thought. KNN, with its effortlessness and ampleness, predicts vehicle costs based on the region of data centres, while SVM focuses to recognize the perfect hyperplane that best separates data into diverse classes, subsequently advancing figure execution through its taking care of non-linear associations. This study explores the comparative execution of K-Nearest Neighbours and support Vector Machines in anticipating utilized car costs. Utilizing information from the Kaggle store, we assess the exactness of these models beneath different preparing and testing scenarios. Our discoveries show that whereas both models show promising comes about, SVM illustrates a slight advantage in exactness over KNN. This inquiries about points to contribute to the continuous talk on prescient modelling within the car division, emphasizing the benefits of progressed machine learning methods in improving the precision of utilized car cost estimations.The following are the variables used:

A. Car Name: The name and model of the car.
B. Registration Year: The year the car was registered
C. Insurance Validity: The validity period of the car's insurance.
D. Fuel Type: The type of fuel the car uses (e.g., petrol, diesel).
E. Seats: The number of seats in the car.
F. Kms Driven: The total kilometres driven by the car.
G. Ownership: The number of previous owners of the car.
H. Transmission: The type of transmission (e.g., manual, automatic).
I. Manufacturing Year: The year the car was manufactured.

DOI: 10.17148/IJARCCE.2025.1411105

J. Mileage (kmpl): The car's mileage in kilometres per litre.
K. Engine(cc): The engine capacity in cubic centimetres.
L. Max Power(bhp): The maximum power output of the car in brake horsepower.
M. Torque (Nm): The torque produced by the car's engine in Newton meters.
N. Price (in lakhs): The price of the car in lakhs (a unit of currency).

II. LITERATURE SURVEY

We examined several studies on the expanding used automobile market and the importance of accurate car price prediction. Recent research highlights the growing effectiveness of machine learning in automobile cost prediction. Gigic et al. (2019) explored ensemble methods combining Random Forest, Support Vector Machines, and Artificial Neural Networks for estimating car prices. Their model demonstrated the effectiveness of ensemble approaches in handling high dimensional data and capturing complex patterns by leveraging the strengths of each algorithm [1]. Using a Kaggle dataset, K. Samruddhi and Dr. R. Ashok Kumar (2021) developed a supervised machine learning model employing K Nearest Neighbor regression. Their model showed notable performance with small datasets, though other studies have noted variations in KNN's effectiveness [2]. Pallavi Bharambe et al. (2021) examined three regression techniques Lasso, ridge, and linear using Kaggle data. Ridge regression emerged as the most effective method in their study, emphasizing the importance of choosing the right regression approach and the need for meticulous feature selection and preprocessing for reliable predictions [3]. Artificial Neural Networks have been identified as highly successful tools for predicting used automobile prices, as demonstrated by Aravind Sasidharan Pillai (2022). His model, using data from 140,000 vehicles, outperformed traditional models and showcased superior accuracy for pricing forecasts [4]. Pudaruth (2020) investigated various machine learning methods, including decision trees and linear regression, for forecasting second hand car prices. The study found that while performance across approaches was similar, achieving high accuracy remained challenging. The research suggests exploring more advanced algorithms to improve predictions [5]. To address market challenges caused by the COVID-19 pandemic, Budiono et al. (2022) proposed using the K-Nearest Neighbours model for predicting used car values. Their model, based on data from a website, demonstrated high accuracy and minimal error, aiming to improve price predictions and resolve trust issues between buyers and sellers [6].

III. METHODOLOGY

The Used Cars data set was taken and data processing has done to filter the data and to remove some unnecessary data. The model was trained with the processed data using KNN algorithm to predict the sales of used cars with higher accuracy. Fig 1 shows the structured outline for proposed Methodology.

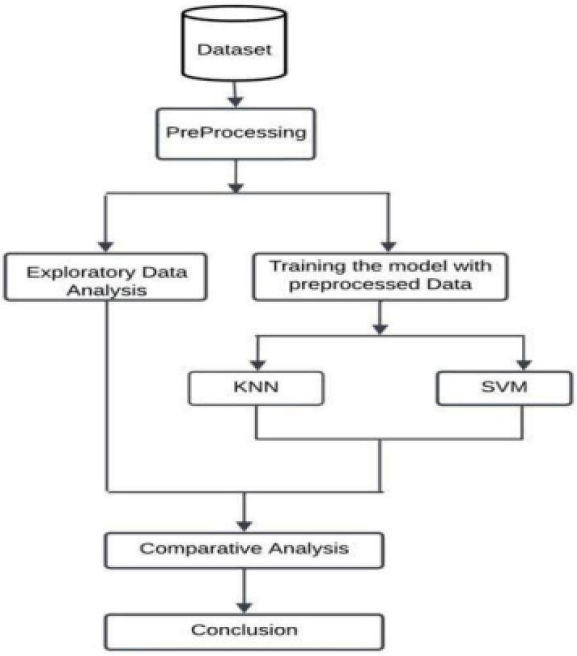

Fig. 1. Structured Outline Of Proposed Methodology

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025

DOI: 10.17148/IJARCCE.2025.1411105

A. Dataset Collection:
It is the process of gathering the information from the source for the evaluation. The Used Cars data set is collected from a website Kaggle which is in a CSV format. The dataset contains 12 variables, including: name, year, selling price, km driven, fuel, seller type, transmission, owner, mileage (km/ltr/kg), engine, max power, and seats.
B. Preprocessing:
After gathering the Used Cars data collection, data processing was done to filter and eliminate certain extraneous information. Using the processed data, the model was trained using KNN and SVM algorithms, used car sales can be predicted more accurately.
C. Exploratory Data Analysis:
Exploratory Data Analysis is the process of analysing the dataset to summarize its main characteristics. This usually involves visualizations and statistical analysis to understand the distribution of the data, detect outliers, and uncover patterns. D. Model Training Using Pre-processed Data:
The next step is to train machine learning models using the pre-processed data. Two specific algorithms are represented here: • K-Nearest Neighbours (KNN): A simple, distance-based classification algorithm that predicts the label of a new data point based on the labels of its nearest neighbours. • Support Vector Machine (SVM): A more advanced algorithm that seeks to find the optimal hyperplane separating different classes in the dataset.
E. Comparative Analysis:
After training the models, their performance is evaluated and compared. This comparison may use various metrics such as accuracy, precision, recall, or other evaluation methods to determine which model performs better on the test data. F. Conclusion:
The final stage involves drawing conclusions based on the comparative analysis. This could include selecting the best model for deployment, providing insights into the strengths and weaknesses of each approach, or suggesting potential future improvements.

IV. IMPLEMENTATION

Machine Learning is a branch of artificial intelligence focused on designing algorithms that allow systems to learn from data and make informed predictions or decisions without being explicitly programmed. This field includes various methods like supervised, unsupervised, and reinforcement learning to handle different types of tasks. K-Nearest Neighbours (KNN) is a straightforward supervised learning algorithm used for classification and regression. It works by evaluating the k closest data points to a given instance and assigning the instance a label or value based on the majority or average of these neighbours. Support Vector Machines (SVM) is a powerful supervised learning technique used for classification and regression. It seeks to find the optimal hyperplane that best divides different classes in the data, aiming to maximize the margin between the closest points from each class, known as support vectors

Table 1. shows the performance metrics of svm classifier

Accuracy Score	0.80121
Precision Score	0.80466
F1 Score	0.80163

Table 2. shows the performance metrics of knn classifier

Accuracy Score	0.83333
Precision Score	0.83341
F1 Score	0.83329

Table I shows the performance metrics of the SVM classifier, which achieved an accuracy of 0.80121, a precision of 0.80466, and an F1 score of 0.80163. Table II illustrates the KNN classifier's better performance, with an accuracy of 0.83333, a precision of 0.83341, and an F1 score of 0.83329. These findings suggest that the KNN classifier is more effective than the SVM classifier for predicting car prices using this dataset.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025

DOI: 10.17148/IJARCCE.2025.1411105

Table 3. actual price and predicted price

Serial NO	Actual price	Predicted Price
1	8.99	6.4
2	13.56	9.3
3	5.45	4.7
4	5.12	4.59
5	9.3	7.5
6	8.02	5.8
7	10.95	8.6
8	8.99	6.4
9	7.45	4.3
10	10.45	8.3

Table III The analysis of car price predictions using the K-Nearest Neighbours (KNN) algorithm reveals important insights into the model's performance. By comparing actual car prices to their predicted counterparts, we can assess how accurately the model forecasts these values. Some predictions align closely with the actual prices, demonstrating that the model can indeed make reliable predictions under certain conditions. The actual price of 7.99 lakhs is predicted as 7.96 lakhs, which is relatively accurate

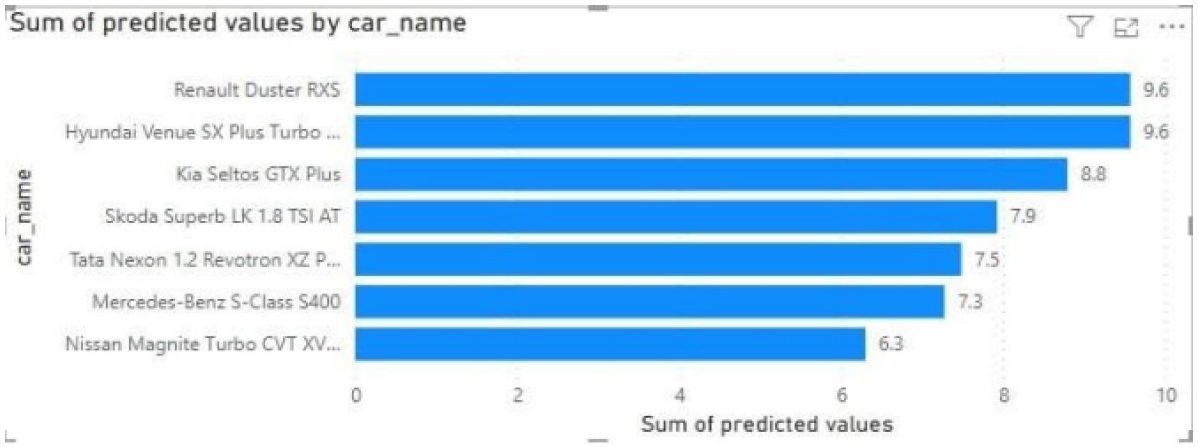

Fig. 2 Predicted Price for Different Car Models using KNN

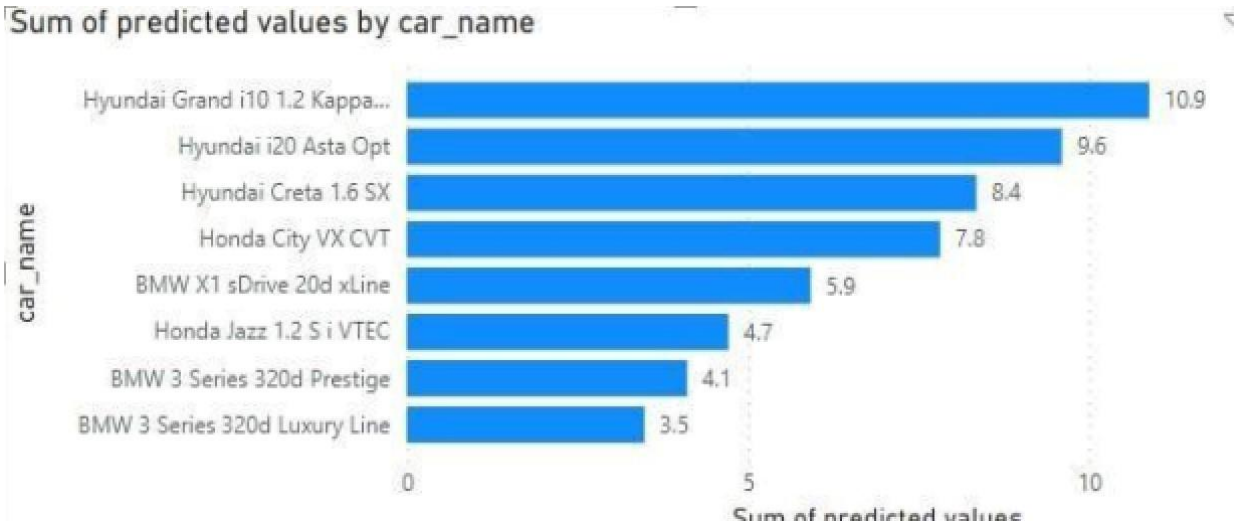

Fig. 3 Predicted Price for Different Car Models using SVM

DOI: 10.17148/IJARCCE.2025.1411105

The fig.2 and fig.2 is a bar chart you provided displays the summed predictions of used car prices categorized by model using knn and svm. Each bar represents the aggregate predicted value for a specific model, ranked from highest to lowest. This visualization aids in identifying which car models hold higher estimated values, likely indicating either elevated demand or pricing according to the model predictions derived from the dataset. The fig.2 summarizes At 10.88, Hyundai Grand i10 1.2 Kappa Sportz BSIV had the highest Sum of predicted values and was 212.52% higher than BMW 3 Series 320d Luxury Line, which had the lowest Sum of predicted values at 3.48. Hyundai Grand i10 1.2 Kappa Sportz BSIV accounted for 19.83% of Sum of predicted values. Across all 8 car name, Sum of predicted values ranged from 3.48 to 10.88. The fig.3 summarizes Renault Duster RXS and Hyundai Venue SX Plus Turbo DCT tied for highest Sum of predicted values at 9.57, followed by Kia Seltos GTX Plus. Nissan Magnite Turbo CVT XV Premium Opt BSVI had the lowest Sum of predicted values at 6.30. Across all 7 car name, Sum of predicted values ranged from 6.30 to 9.57.

V. COMPARATIVE ANALYSIS

When predicting used car prices, the choice between K-Nearest Neighbours (KNN) and Support Vector Machines (SVM) hinges on factors such as dataset size, dimensionality, and computational resources. KNN achieved an accuracy of 83%, while SVM reached 80%, both surpassing the 70% accuracy obtained from the exploratory data analysis (EDA). KNN, though simple and effective with smaller datasets, can be computationally expensive and less suitable for high-dimensional data due to its distance calculations and memory requirements. Conversely, SVM excels in handling high-dimensional data and capturing complex, nonlinear relationships through kernel functions, albeit at a higher computational cost and the need for careful parameter tuning. Given the higher accuracy of KNN in this case, it appears to be a better fit for the dataset, yet the decision should also consider model interpretability, computational costs, and generalization to unseen data.

VI. CONCLUSION

This study compared the performance of K-Nearest Neighbour (KNN) and Support Vector Machine (SVM) classifiers in predicting used car prices using a Kaggle dataset. The results demonstrated that the KNN classifier outperformed the SVM classifier, achieving an accuracy of 83.33%, a precision of 83.34%, and an F1 score of 83.32%. In contrast, the SVM model recorded an accuracy of 80.12%, a precision of 80.46%, and an F1 score of 80.16%. These findings suggest that the KNN model is better suited for this dataset. Future research could explore combining different machine learning techniques, refining feature selection methods, and improving data preprocessing to enhance prediction accuracy. Additionally, other machine learning models could be tested to identify the most effective approach for predicting used car prices. Enhanced feature engineering and the incorporation of more complex algorithms might also contribute to improved performance. Overall, this study highlights the potential of KNN for used car price prediction and underscores the importance of continuous refinement in machine learning methodologies to achieve better results

REFERENCES

[1]. Akerlof, G. A. (1970). The market for “lemons”: Quality uncertainty and the market mechanism. *Quarterly Journal of Economics, 84*(3), 488–500. https://doi.org/10.2307/1879431
[2]. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
[3]. Broder, A. Z. (1997). On the resemblance and containment of documents. *Proceedings of the Compression and Complexity of Sequences*, 21–29. https://doi.org/10.1109/SEQUEN.1997.666900
[4]. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine.
[5]. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451
[6]. Indyk, P., & Motwani, R. (1998). Approximate nearest neighbors: Towards removing the curse of dimensionality. *Proceedings of the 30th Annual ACM Symposium on Theory of Computing*, 604–613. https://doi.org/10.1145/276698.276876
[7]. Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy, 74*(2), 132–157. https://doi.org/10.1086/259131
[8]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.
[9]. Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy, 82*(1), 34–55. https://doi.org/10.1086/260169
[10]. UCI Machine Learning Repository. (2019). *Automobile dataset & car evaluation dataset*. University of California, Irvine. https://archive.ics.uci.edu/ml

