JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Analysis of Used Car Price Prediction Using Machine Learning

Ramireddy Siva
Asst. Professor, Department of CSE
AITS, Tirupati.
rsnr777@gmail.com

Gumma Venkata Krishna
UG Scholar, Department of CSE
AITS, Tirupati.
gummakrishna30@gmail.com

Chenna Sree Lakshmi
UG Scholar, Department of CSE
AITS, Tirupati.
chennasreelakshmi9@gmail.com

Sunkara Sathish
UG Scholar, Department of CSE
AITS, Tirupati.
sathishsunkara279@gmail.com

Shaik Barkat Ali
UG Scholar, Department of CSE
AITS, Tirupati.
barkatalishaik167@gmail.com

***Abstract*—The high rate of expansion of the pre-owned automotive market has enhanced the demand of viable and precise means of estimating the prices of vehicles. It is not easy to determine the resale value of a used car because it has to rely on various parameters including the age of the car, car mileage, the type of fuel and the transmission system and ownership. Traditional methods of pricing are largely based on subjective information by use of manual pricing and recommendations; this may lead to inconsistent and unrealistic pricing. Machine learning methods present an alternative approach which is data-driven by defining the patterns of pricing based on the past historical vehicle data. The paper provides comparative analysis of regression based machine learning models used to predict the price of used cars, such as Linear Regression, Random Forest Regressor and Gradient Boosting Regressor. These data preprocessing methods include feature encoding, outlier handling, and normalization which are embedded in the system. The standard regression evaluation metrics are the models used to measure model performance. Experimental experiments have shown that ensemble learning paradigms are much better than the traditional linear algorithms in their ability to describe non-linear associations between participants. The suggested system should help online platforms, dealerships, and customers make prices of their deals informed.**

***Index Terms*—Used Car Price Prediction, Machine Learning, Regression, Random Forest, Gradient Boosting**

I Introduction

Carmobile industry is an important part of the contemporary systems of transport and the need of used cars has increased dramatically based on economic viability and expanded accessibility. In comparison to the new cars, which have comparatively fixed prices, the used cars have very dissimilar prices as per their various characteristics among them being the manufacturing year, the total distance covered, brand value, type of fuel consumed by the car, level of transmission and past ownership history.

Proper prediction of the resale value can help the purchasers, the vendors as well as the car dealers. This provides buyers with protection against overpricing, enables sellers to adopt realistic and competitive prices and also provides dealers with standardised valuation prac- tices. However, standard pricing systems are quite manualized, dependent on past experience, and have minimal rules involved, which is much harder to scale and highly susceptible to subjectivity.

Machine learning gives a relevant solution to this problem because it can use past vehicle information to make its own dedication to sophisticated pricing patterns. With the usage of regression-based models, it is possible to induce credible predictions of prices with a much better prediction. The paper is devoted to the examination of various regression methods and selection of the most appropriate one regarding car prices prediction of used cars.

I-A Motivation

The primary motivation for this work arises from the difficulties faced by participants in used vehicle marketplaces when estimating fair resale prices. Due to limited market knowledge and continuously changing trends, both buyers and sellers often struggle to determine appropriate vehicle values. An automated machine learning-based prediction system can address these challenges by providing accurate and unbiased price estimates, thereby improving transparency and trust in the used car market.

I-B Contributions of the Paper

The major contributions of this research are as follows:

- Development of a complete machine learning workflow for predicting used car resale prices.
- Implementation and comparative analysis of regression models, including Linear Regression, Random Forest, and Gradient Boosting.
- Application of real-world data preprocessing techniques such as handling missing values, encoding categorical features, removing outliers, and feature scaling.
- Evaluation of model performance using standard regression metrics to identify the most effective approach.

II Problem Statement

Used car price prediction is formulated as a regression problem where the output variable represents a continuous resale value. The pricing process is influenced by several interdependent and non-linear factors, along with changing market conditions. The objective of this study is to design an efficient predictive model capable of estimating vehicle resale prices based on important attributes such as:

- Vehicle manufacturing year and age
- Total kilometers driven
- Fuel category
- Transmission type
- Ownership status

The goal is to enhance prediction accuracy while minimizing estimation errors, making the system suitable for practical real-world applications.

III Literature Survey

Numerous studies and educational materials have contributed to the current machine learning methods to solve regressions and prediction tasks. The references cited in this paper are implemented to support a particular aspect of the proposed system, that is, the pre- processing, regression modeling, ensemble learning, boosting, evaluation, and tool implementation.

Witten et al. [1] described key concepts underlying data mining and machine learning systems, and focused on supervised data mining techniques to create predictive models. Breiman[2] has given a proposal where random forests enhance the effectiveness of pre- diction through the combination of decision trees and minimization of variance.

Hastie et al. [3] described statistical learning techniques and gave an insight on the regression techniques and model generalization. The Kaggle dataset [4] is taking the real-life records of auction- automobile prices that are used to test and validate in the predictive modeling projects.

Brownlee [5] introduced Gradient Boosting and explained why techniques to boost can be effective compared to simple regression models on complicated datasets. Bishop [6] has covered the basics of machine learning and the significance of feature representation when doing regression.

Russell and Norvig [7] presented the background of artificial intelligence, and it guided in the perception of how automated decision-making system is used. Geron [8] offered examples of useful regression models and preprocessing pipelines in Scikit-learn.

Other crucial machine learning principles identified by Domingos [9] are bias- variance trade-off, quality of features and overfitting prevention. Gradi- ent Boosting Machines were formally presented by Friedman [10] and its efficiency in the regression process was demonstrated.

The concepts of weak learnability were discussed by Schapire [11] and had an impact on the methods of boosting. The AdaBoost learning algorithm which is a significant ensemble method introduced by Friedrich Freund and Shapire [12] is a concluding algorithm that generates stronger learners as a combination of weak learners.

Quinlan [13] proposed decision tree learning techniques that serve as a basis to most tree-based ensemble models. Kohavi [14] was concerned with cross-validation and model assessment to choose the most appropriate model to use with unseen data.

A review of data analysis and clustering by Jain et al. [15] helps to gain a better insight into the pattern of data sets and feature distribution. The nature of dataset imbalance is important to address and Chawla et al. proposed SMOTE on enhancing training stability [16].

Ng [17] has explained machine learning algorithms and feature scaling and model training process which are appropriate and coincide with the workflow of this project. Another theory that can be researched in future work when seeking advanced prediction is neural networks and learning machines, as explained by Haykin [18].

LeCun et al. [19] introduced foundations of deep learning research as well as demonstrated the ways of using neural networks in large-scale prediction. Platt [20] talked about probabilistic modeling techniques that will enhance interpretability.

Smola and Scho¨lkopf [21] described Support Vector Re-gression that is applicable in non-linear regression issues. Early investigations on SVR and regression modelling were done by Drucker et al. [22].

Kuhn and Johnson [23] offered useful information on pre-dictive modeling with a focus on feature engineering and feature appraisal. Chollet [24] clarified deep learning based methods that can further predict in case there is adequate data.

Both the regression model and preprocessing tools are available as standard imple- mentations in scikit-learn documentation [25]. NumPy documentation [26] facilitates the work of numerical calculations and array operations necessary in ML. Pandas documentation[27] aids in data cleaning, data manipulation and preprocessing.

The tools allowing visualization of the dataset relations and graphical representation of the results are offered in Matplotlib documentation [28]. XGBoost was proposed by Chen and Guestrin [29] and is commonly applied in boosting-based regression and can significantly improve the performance.

The LightGBM documentation [30] offers efficient gradient boosting applications that can be applied to large data sets. An enhancement in mixed feature schemes was recently proposed in CatBoost documentation [31] proposing a boosting algorithm that effectively works with categorical variables, adding value through boosting algorithm technology.

IV Dataset Description

The data used in this paper is obtained via the real-life used vehicles listings and it is recorded in the project repository. It consists of a blend of both numerical and categorical attributes which illustrate the different attributes of the vehicles including the name of the manufacturer, year of production, distance covered, type of fuel and gear box structure as well as ownership history.

IV-A Dataset Features

Table I summarizes key dataset attributes. These features strongly influence vehicle resale value and help the model learn relationships between input variables and price.

IV-B Dataset Overview Image

Fig. 1 provides an overview of the dataset distribution, helping to understand patterns in the data.

TABLE I
DATASET FEATURES AND DESCRIPTION

Feature	Description
Brand	Manufacturer name of the vehicle
Year	Year of manufacture
Km Driven	Total kilometers driven
Fuel Type	Petrol/Diesel/CNG/Electric
Transmission	Manual/Automatic
Owner	First/Second/Third owner
Selling Price	Target value (vehicle price)

Overview of Used Vehicle Dataset

Brand	Year	Kilometers Driven	Fuel Type	Transmission	Owners	Selling Price
Maruti	2015	45,000	Petrol	Manual	First	350,000
Hyundai	2013	65,000	Diesel	Automatic	Second	480,000
Honda	2017	28,000	Petrol	Manual	First	620,000
Toyota	2012	80,000	CNG	Manual	Third	1298 CC
Ford	2016	55,000	Diesel	Manual	Second	520,000

Fig. 1. Overview of Used Vehicle Dataset

V Exploratory Data Analysis (EDA)

Exploratory Data Analysis is carried out to understand the structure of the dataset and examine relationships between different variables. The observations show that vehicles manufactured in recent years generally have higher resale values, while increased mileage is commonly associated with lower selling prices.

V-A Key Observations

- Selling price decreases as kilometers driven increases.
- Newer vehicles generally have higher resale price.
- Transmission type and fuel type affect resale value.

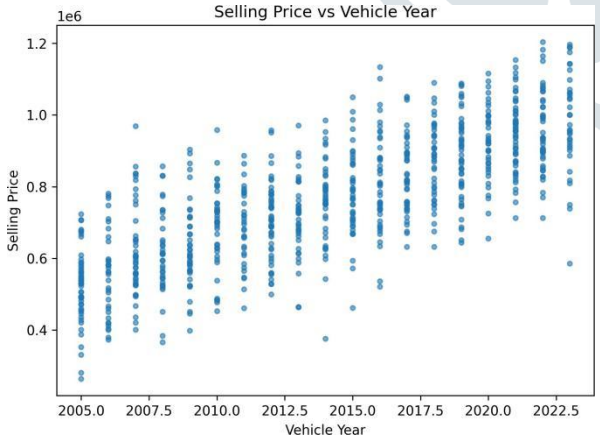

Fig. 2. Selling Price vs Vehicle Year

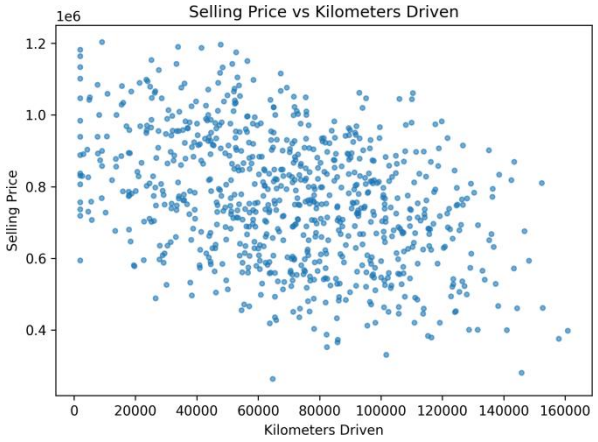

Fig. 3. Selling Price vs Kilometers Driven

VI Data Preprocessing

Data preprocessing improves dataset quality and model performance. Missing values were handled using imputation. Categorical features were encoded into numerical values and feature scaling was applied to normalize the dataset.

VII Feature Engineering

Feature engineering improves learning by creating meaningful features. In this project, derived features such as vehicle age were used for better prediction accuracy.

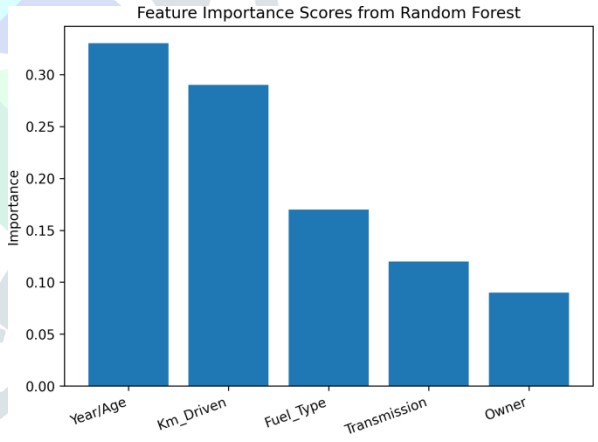

Fig. 4. Feature Importance Scores from Random Forest

VIII Methodology

This study adopts a step-by-step machine learning framework for model development and evaluation. The dataset is partitioned into training and test sets in an 80:20 ratio. Regression algorithms are trained using the training subset and validated on the test data to measure prediction accuracy. The model with the best performance is chosen for deployment.

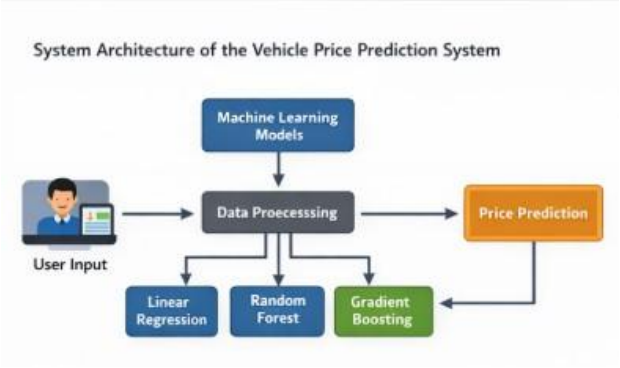

Fig. 5. System Architecture of the Vehicle Price Prediction System

IX Machine Learning Models Used

This study applies multiple regression-based machine learning algorithms to estimate used vehicle prices. The following models are implemented and evaluated:

- Linear Regression model
- Random Forest–based regression model
- Gradient Boosting–based regression model

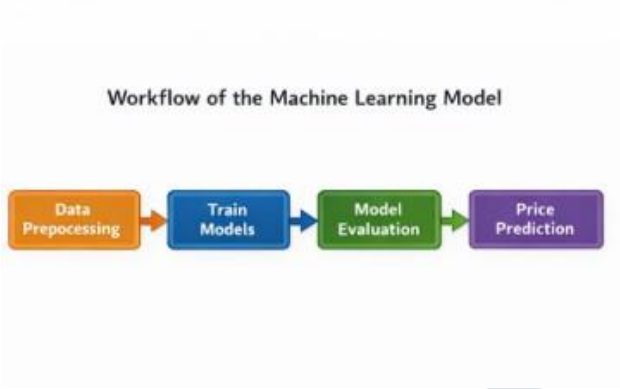

Fig. 6. Workflow of the Machine Learning Model

X Performance Evaluation

To evaluate model performance, multiple regression metrics were used.

X-A Mean Absolute Error (MAE)

Mean Absolute Error measures the average magnitude of prediction errors by computing the absolute difference between the observed values and the model predictions.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (1)$$

X-B Mean Squared Error (MSE)

Mean Squared Error evaluates prediction performance by squaring the differences between actual and predicted values, thereby assigning greater weight to larger errors.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2)$$

X-C Root Mean Squared Error (RMSE)

Root Mean Squared Error is obtained by taking the square root of the Mean Squared Error, allowing the error to be interpreted in the same scale as the target variable.

$$RMSE = \sqrt{MSE} \quad (3)$$

X-D Coefficient of Determination (R-squared)

The R-squared metric indicates the proportion of variance in the dependent variable that is explained by the predictive model.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (4)$$

XI Results and Discussion

The results indicate that ensemble-based models perform better than basic linear regression due to their ability to handle non-linear relationships.

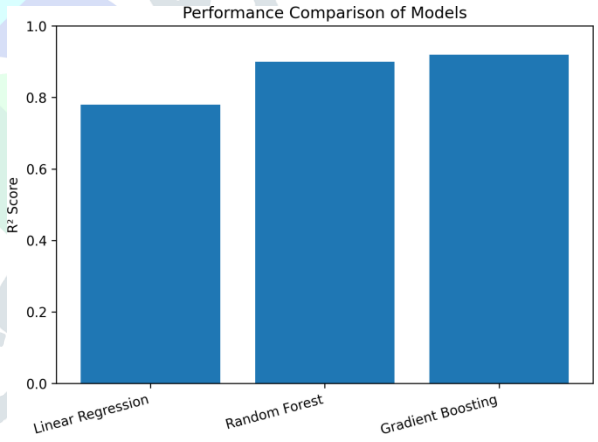

Fig. 7. Performance Comparison of Models

XII Conclusion and Future Work

This study presented a machine learning-based framework for predicting used car resale prices using real-world automobile data. The approach involved data preprocessing, feature engineering, and training multiple regression models to evaluate their predictive capabilities. Among the tested models, Gradient Boosting demonstrated superior performance compared to traditional linear regression techniques.

The proposed system can be effectively applied in online used car platforms, dealership valuation systems, insurance assessments, and automated pricing tools. However, the current

dataset does not account for real-time market demand, regional price variations, or detailed service history. Future enhancements may include integrating live market data, incorporating location-based features, exploring advanced models such as XGBoost, and deploying the solution as a web or mobile application.

References

[1] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2016.
[2] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.
[4] Kaggle, "Used Car Price Prediction Dataset," 2023.
[5] J. Brownlee, "A Gentle Introduction to Gradient Boosting," 2020. [Online]. Available: https://machinelearningmastery.com/ gentle-introduction-to-gradient-boosting-algorithm-machine-learning/
[6] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
[7] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Pearson, 2010.
[8] A. Ge´ron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, 2019.
[9] P. Domingos, "A Few Useful Things to Know About Machine Learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.
[10] J. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
[11] R. E. Schapire, "The Strength of Weak Learnability," *Machine Learning*, vol. 5, pp. 197–227, 1990.
[12] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.
[13] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.
[14] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in *Proc. IJCAI*, 1995, pp. 1137–1143.
[15] A. Jain, M. Murty, and P. Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.
[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
[17] A. Ng, "Machine Learning," Stanford University Lecture Notes, 2018.
[18] S. Haykin, *Neural Networks and Learning Machines*. Pearson, 2009.
[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, pp. 436–444, 2015.
[20] J. Platt, "Probabilistic Outputs for Support Vector Machines," in *Advances in Large Margin Classifiers*, MIT Press, 1999.
[21] A. Smola and B. Scho¨lkopf, "A Tutorial on Support Vector Regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.
[22] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support Vector Regression Machines," in *Proc. NIPS*, 1996.
[23] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Springer, 2013.
[24] F. Chollet, *Deep Learning with Python*. Manning Publications, 2017.
[25] Scikit-learn, "Machine Learning in Python Documentation," 2023. [Online]. Available: https://scikit-learn.org/stable/
[26] NumPy Developers, "NumPy Documentation," 2023. [Online]. Available: https://numpy.org/doc/
[27] Pandas Development Team, "Pandas Documentation," 2023. [Online]. Available: https://pandas.pydata.org/docs/
[28] Matplotlib Developers, "Matplotlib Documentation," 2023. [Online]. Available: https://matplotlib.org/stable/
[29] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. ACM SIGKDD*, 2016, pp. 785–794.
[30] LightGBM, "LightGBM Documentation," 2023. [Online]. Available: https://lightgbm.readthedocs.io/
[31] CatBoost, "CatBoost Documentation," 2023. [Online]. Available: https: //catboost.ai/