RESEARCH ARTICLE

PRE-OWNED CAR PRICE PREDICTION: A MACHINE LEARNING APPROACH

Thanu V. Marnholkar*, Suraj Vithoba Salam,

A. M. Chopade, L. G. Mulik and N. S. Kulkarni

Department of Computer Science,

Dr. Ghali College, Gadhinglaj, Dist. Kolhapur, M.S. India

*Corresponding author E-mail: thanumarnholkar3110@gmail.com

DOI: https://doi.org/10.5281/zenodo.15125462

Abstract:

The valuation of pre-owned cars is a crucial aspect of the used car market, where accurate price estimation can help buyers and sellers make informed decisions. This study employs Machine Learning (ML) techniques to develop a predictive model for estimating resale values. A dataset of used cars is processed through data cleaning, feature encoding, and exploratory data analysis to identify key factors affecting car prices, such as brand, mileage, fuel type, and transmission type. The model is trained using Linear Regression, a widely used technique for predicting continuous values. Performance evaluation is conducted using metrics such as R^2 score, which measures how well the model explains variance in car prices. The study also explores data visualization techniques to understand market trends and enhance model interpretability. Experimental results indicate that Linear Regression, with optimized features, provides a reasonable estimation of used car prices. This approach offers an efficient and data-driven method for price prediction, enhancing transparency in the pre-owned car market.

Keywords: Pre-Owned Car Price Prediction, Machine Learning, Linear Regression, R^2_score.

Introduction:

The pre-owned car market has expanded significantly in recent years, driving the need for accurate and data-driven price prediction models. Traditional valuation methods rely on expert assessments and market trends, often leading to inconsistencies in pricing. However, advancements in Machine Learning (ML) have enabled the development of more precise and automated models for estimating the resale value of used cars. Several studies, including those by [1] and [2], have demonstrated the effectiveness of ML techniques in predicting car prices based on historical data and key influencing factors.

Among the various ML techniques, Linear Regression is widely used due to its simplicity and interpretability in identifying relationships between independent variables (such as car age, mileage, fuel type, and transmission) and the dependent variable (selling price). Research by [5] and [4] highlights the accuracy of regression models in capturing price trends. Additionally, the [3] serves as a valuable resource for training and optimizing these models.

To evaluate the reliability of predictive models, performance metrics such as R^2 score and Mean Absolute Error (MAE) are commonly used. While Linear Regression provides a strong foundation for car price prediction, alternative methods such as Support Vector Regression [6] and [7] offer advanced capabilities for capturing complex price variations.

This study focuses on developing a Pre-Owned Car Price Prediction system using Machine Learning, specifically Linear Regression, to provide accurate and data-driven price estimates. The model is designed to improve pricing transparency, benefiting both individual buyers and car dealerships by reducing uncertainty in the used car market.

Flowchart:

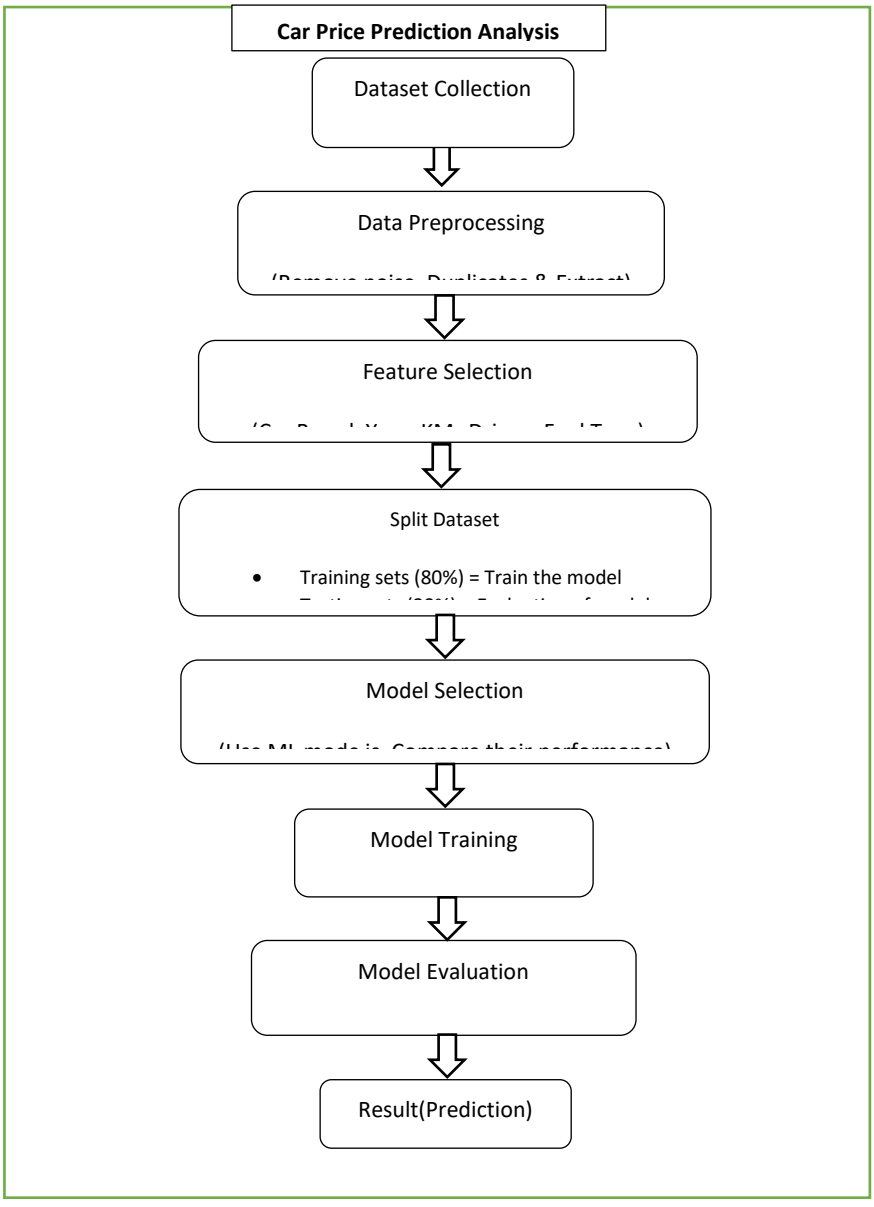

Fig. 1: Flowchart of proposed work

1. Dataset:

```
car_data.head()
```

	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner
0	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
1	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
2	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
3	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
4	swift	2014	4.60	6.87	42450	Diesel	Dealer	Manual	0

Fig. 2: A sample dataset considered for modelling

2. Managing Missing Data: In car price prediction, handling missing data is essential to maintain model accuracy and reliability. Based on [5], missing data in Machine Learning models for car price estimation can significantly impact performance if not managed properly. The study emphasizes the following technique to address missing values. Removing Rows with Missing Values If the number of missing entries is small, dropping incomplete records can prevent introducing noise into the dataset.

3. Data Cleaning: Is an essential step in Machine Learning-based car price prediction to ensure the dataset is accurate, consistent, and ready for model training. [5] and [2] emphasized that poor-quality data can negatively impact model performance, leading to incorrect price estimations. The [3] also highlights the need for cleaning raw car sales data before analysis.

4. Feature Selection: Feature selection is a crucial step in Machine Learning-based car price prediction, as it helps improve model accuracy by identifying the most relevant attributes. [1] and [4] emphasized that choosing the right features leads to better predictions and reduces computational complexity. [2] further demonstrated that Regression Models perform best when redundant and irrelevant features are removed.

5. Model Training: The training process for pre-owned car price prediction involves preparing the dataset, selecting an appropriate Machine Learning (ML) model, and evaluating its performance using key metrics such as R^2 score. Based on previous research, including [1] and[2], Linear Regression has been identified as a fundamental approach for predicting used car prices.

6. Evaluation Matrix Results: Evaluating the performance of a Pre-Owned Car Price Prediction model is essential to ensure accurate and reliable price estimations. This study uses Machine Learning (ML), specifically Linear Regression, to predict car prices and assesses the model's effectiveness using key evaluation metrics. The performance is measured using R^2 score, along with other statistical error metrics. Similar studies by [1] and [2] have demonstrated the importance of evaluating predictive models to improve their accuracy.

Accuracy For Following Models:

```
[33]: from sklearn.linear_model import LinearRegression, Lasso, Ridge
      from sklearn.metrics import r2_score

[34]: linear_regression = LinearRegression()
      linear_regression.fit(x_train, y_train)
      linear_regression_predictions = linear_regression.predict(x_test)
      linear_regression_r2_score = r2_score(y_test, linear_regression_predictions)

[35]: linear_regression_r2_score

[35]: 0.9154724115923335
```

Fig. 3: Accuracy (Linear Regression) : 91.54 %

```
[37]: Lasso_regression = Lasso()
      Lasso_regression.fit(x_train, y_train)
      Lasso_regression_predictions = Lasso_regression.predict(x_test)
      Lasso_regression_r2_score = r2_score(y_test, Lasso_regression_predictions)

[38]: Lasso_regression_r2_score

[38]: 0.8892562128057028
```

Fig. 4: Accuracy (Lasso Regression) : 88.92 %

```
[40]: Ridge_regression = Ridge()
      Ridge_regression.fit(x_train, y_train)
      Ridge_regression_predictions = Ridge_regression.predict(x_test)
      Ridge_regression_r2_score = r2_score(y_test, Ridge_regression_predictions)

[41]: Ridge_regression_r2_score

[41]: 0.9151481581072847
```

Fig. 5: Accuracy (Ridge Regression) : 91.51 %

Results:

Model	R^2 Score	Accuracy(%)
Linear Regression	0.9154724115923335	91.54%
Lasso Regression	0.8892562128057028	88.92%
Ridge Regression	0.9151481581072847	91.51%

Linear Regression achieved highest accuracy of 91.54% , while Ridge Regression came in second with 91.51% . Lasso Regression gave the weakest performance with 88.92%.

Conclusion:

This study, we implemented and evaluated different regression models to predict the prices of pre-owned cars using machine learning

techniques. The models were assessed using the r^2 score to determine their predictive accuracy.Linear Regression achieved an r^2 score of 91.54%, demonstrating strong predictive performance. Ridge Regression performed similarly with an r^2 score of 91.51%, indicating it is an effective model for price prediction. Lasso Regression yielded an r^2 score of 88.92%, slightly lower than Linear and Ridge Regression, but still demonstrating good predictive capability. This study confirms that machine learning regression models, particularly Linear Regression, are highly effective in predicting pre-owned car prices.

References:

1. Gegic, E., Isakovic, B., Keco, D., Masetic, Z., & Kevric, J. (2019). Car price prediction using machine learning technique. *TEM Journal*.
2. Monburinon, N., Chertchom, P., Kaewkiriya, T., Rungpheung, S., Buya, S., & Boonpou, P. (2018). Prediction of prices for old cars by using regression models. *Proceedings of the International Conference on Business and Industrial Research (ICBIR).*
3. Used cars database. (n.d.). Retrieved from https://www.kaggle.com/orgesleka/used-carsdatabase [Accessed: June 4, 2018].
4. Venkatasubbu, P., & Ganesh, M. (2019). Used cars price prediction using supervised learning techniques. *International Journal of Engineering and Advanced Technology (IJEAT), 9*(1S3).
5. Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. *International Journal of Information and Communication Technology (IJICT).*
6. Listiani, M. (2009). *Support vector regression analysis for price prediction in a car leasing application* (Master's thesis). Hamburg University of Technology.
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.* ACM.