Predicting Prices of Used Cars Using Machine Learning

Muhammad Raees

Computational and Data Sciences

George Mason University

Fairfax, USA

mraees@gmu.edu

***Abstract*— The market for used cars is challenging to understand because of the many factors that affect a car's market worth. This research project uses the machine learning method, linear regression model, to estimate the price of used cars based on their basic features. The main goal is to give potential buyers and sellers a data-driven strategy to help them make well-informed decisions. The project addresses the concerns and hesitations often associated with used car purchases, particularly the fear of overpaying or acquiring a vehicle with unknown issues. By using linear regression, this paper strives to uncover the important factors that significantly influence a car's resale value, enabling reliable price predictions. With data-backed pricing forecasts, it gives buyers the tools they need to negotiate more effectively. On the other hand, sellers can also benefit by making sure they don't undervalue the car and are able to set prices that are both competitive and attractive. This, in turn, can lead to increased profitability and quicker sales. The models exhibit significant predictive power, with R-squared values of 0.910 for training data and 0.922 for test data, explaining approximately 91% of the price variance. Notably, key predictors like manufacturing year and engine power emerged as vital drivers of price. These findings empower buyers and sellers by providing data-backed insights for well-informed decision-making.**

Keywords— used cars price prediction, linear regression model.

I. Introduction

Accurately determining a vehicle's genuine market worth can be difficult for both buyers and sellers due to the complexity of the used car market. Purchasing used automobiles makes buyers to be hesitant due to worries about overpaying and hidden concerns. Buyers can have a reliable method to assess used cars' fair market value when they apply a pricing prediction approach supported by data-driven insights. The project aims to develop a machine learning model, a linear regression model, to predict used automobile values according to their basic characteristics.

This project has two benefits: it gives buyers more negotiation leverage by providing data-supported pricing projections; also, it helps sellers set competitive prices without undervaluing their products. Faster sales and more profitability follow from this. This project investigates the most important variables that impact an automobile's resale value and looks at whether the location of the sale has an impact on the vehicle's cost. By exploring these topics, the project aims to provide useful instruments for well-informed decision-making.

In Section II, a review of existing research concerning the prediction of used car prices is provided. Section III delves into the details of the dataset and the preprocessing steps carried out. Section IV discusses the results of the exploratory data analysis on the dataset. In Section V, the paper introduces the proposed model, which utilizes linear regression for predicting car prices, along with a comprehensive account of the accuracy of the models in relation to both training and test data. Finally, Section VI offers concluding remarks.

II. Literature Review

[1] uses the 'Used Car Database' dataset to investigate the prediction of used car prices using Random Forest. One of the testing dataset's notable results is a good R-squared score of 83.63%, which indicates effective price prediction.

The paper's strong point is the well-chosen algorithm and feature selection approach. Integrating real-time market data, investigating multiple methods, and evaluating model

generalization across various markets and vehicle classifications can be the main areas of future study.

Monburinon et al.'s [2] study addresses the challenge of accurately pricing used cars. With a dataset of 371,528 car observations, the study explores the application of regression models, including Random Forest Regression, Multiple Linear Regression, and Gradient-Boosted Regression.
Results indicate that Gradient-Boosted Regression performed exceptionally well with an MSE of 0.28, while Random Forest Regression closely followed with an MSE of 0.35. Multiple Linear Regression had the highest MSE at 0.55.
The paper's commendable use of ensemble methods, such as bagging, to enhance model performance is a highlight. Monburinon et al.'s research is a valuable reference for those seeking to improve pricing accuracy in the used car market, benefitting both buyers and sellers.

In the paper by Pudaruth [3], the application of machine learning to predict used car prices in Mauritius is investigated using over 400 car data points. Various models, including Multiple Linear Regression, K-nearest neighbors, Decision Trees, and Random Forest, are employed. Notably, transforming prices into logarithmic values improved regression results. Multiple linear regression model produced a regression coefficient of 0.819, which increased to 0.851 after transforming prices into logarithmic values. For K-Nearest Neighbors (KNN**)** the data was segmented based on car make (Nissan and Toyota), and KNN was applied with different values of K (1, 3, 5, and 10) for each make. Results revealed that KNN performed better for Nissan cars, indicating more consistent pricing for Nissan vehicles.
Decision Trees highlighted the attribute "YEAR" as the most influential factor, with a 64% success rate for the decision tree model. The random forest model exhibited similar results to the decision tree, achieving a success rate of 59%.

III. Data Collection and Preprocessing

This study is based on a large dataset that was obtained from Kaggle, a well-known site for collaboration and the sharing of data. The dataset is from an Indian company, Cars4u, that sells used cars and it includes 7,253 car observations, a sizable sample size, and each observation has 14 unique features. A car's market value is largely determined by a variety of attributes, many of which are included in the dataset. The dataset includes the following features - ‘Serial no.’, ‘Name’, ‘Location’, ‘Year’, ‘Kilometers driven’, ‘Fuel type’, ‘Transmission’, ’Owner type’, ‘Mileage’, ’Engine’, ‘Power’, ‘seats’, ‘New Price’, ‘Price’. Preprocessing and data cleaning techniques were used to make sure that any outliers, missing values, or inconsistencies that may have affected the accuracy of the models were removed.

- Changed column name of milage to 'fuel economy.'
- Converted data types of fuel economy, engine, power, and new price to numerical.
- Removed outlier where the kilometer driven was 6,500,000km.
- Replaced the value 0 with NaN for cars where the seats and fuel economy were equal to 0.
- Split the column ‘name’ into ‘brand’ and ‘model.’
- Imputed missing values for seat, fuel_economy, engine, power, and New_Price by taking medians of the particular model of the car.
- Dropped the column ’S.No.’ (serial number).
- Dropped two observations where the fuel type was electric.
- Dropped observations that had null values for price and new price.

IV. Data Exploration

After preprocessing the data, it was time to explore the data visually to gather insight. Multiple types of graphs were used such as bar charts, boxplots, and scatterplots to explore the data and find relations between explanatory variable and the target variable.

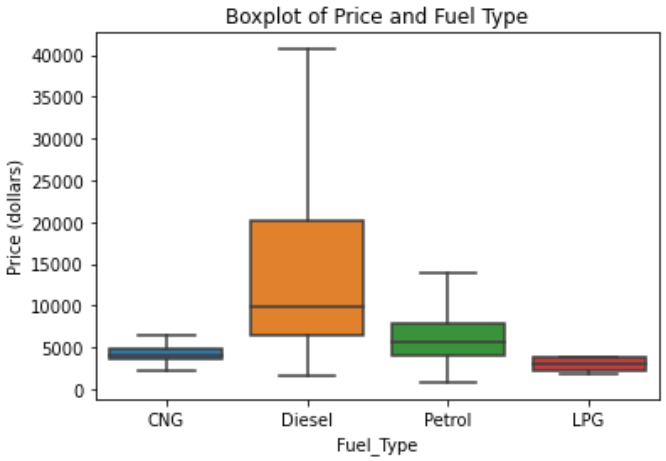

Fig. 1. Boxplot of Price and Fuel Type

Figure 1 shows us that diesel cars are more expensive than other fuel types like Petrol and Gas.

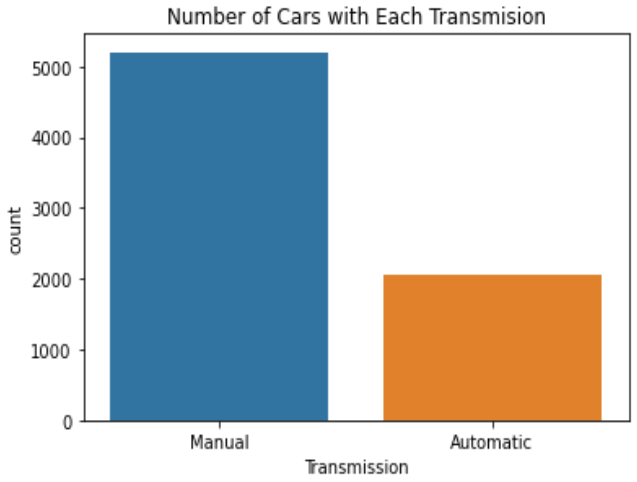

Fig. 2. Bar chart of Number of Cars with Each Transmission

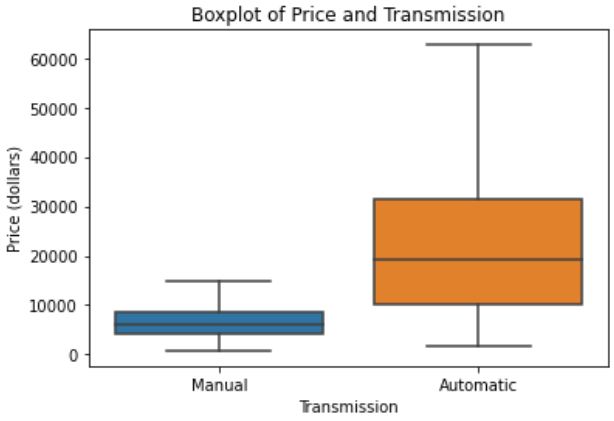

Fig. 3. Boxplot of Price and Transmission

Figure 2 and 3 shows that people still prefer to buy cars with manual transmissions mostly mainly because Automatic cars are more expensive than manual cars.

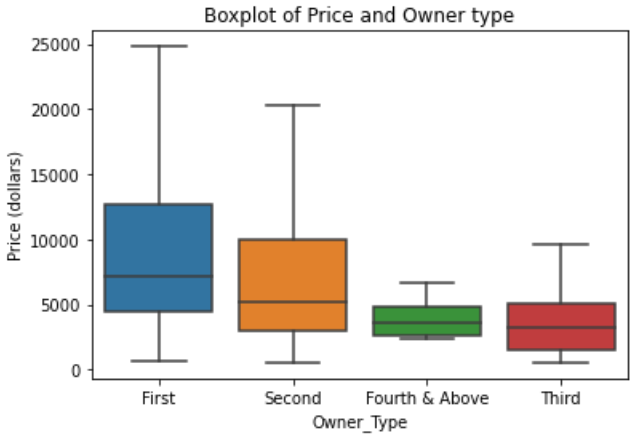

Fig. 4. Boxplot of Price and Owner Type

Figure 4 shows that first-owner vehicles are more expensive than second- or third-owner vehicles, which is expected because when a vehicle is sold repeatedly, its value declines.

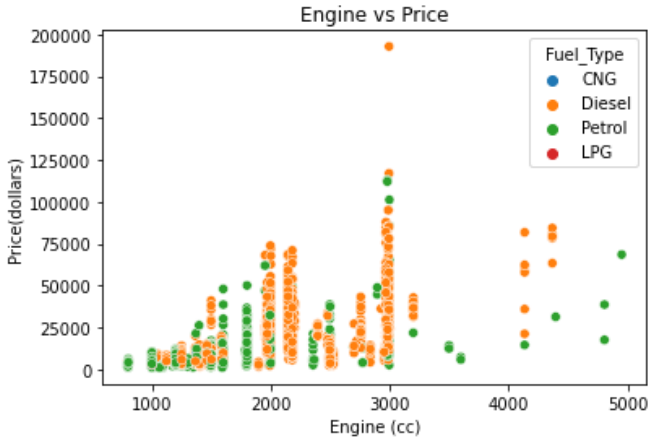

Fig. 5. Scatterplot of Engine vs Price

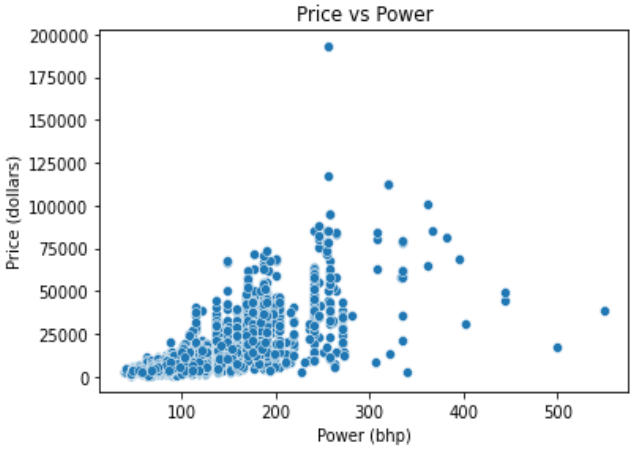

Fig. 6. Scatterplot of Power vs Price

Figure 5 and 6 shows that as power and engine size increases, the price of the car also increases. The cars with engines 2000-3000cc are most likely to run on diesel.

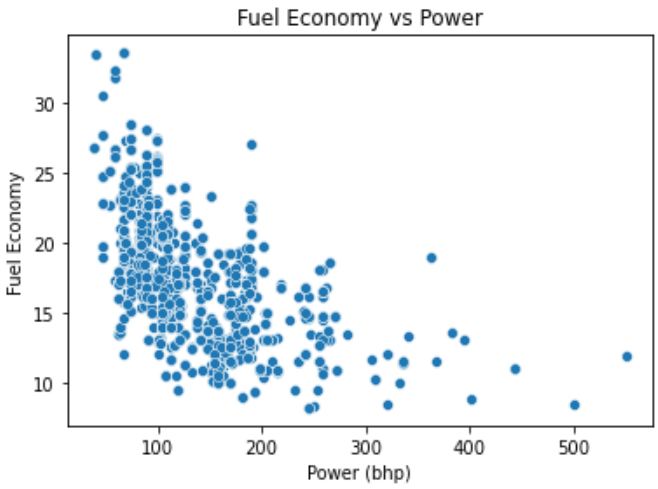

Fig. 7. Scatterplot of Fuel Economy vs Power

Figure 7 shows that as the power of the car increases, the fuel economy decreases. The more powerful the engine, the more fuel it consumes making it less fuel efficient.

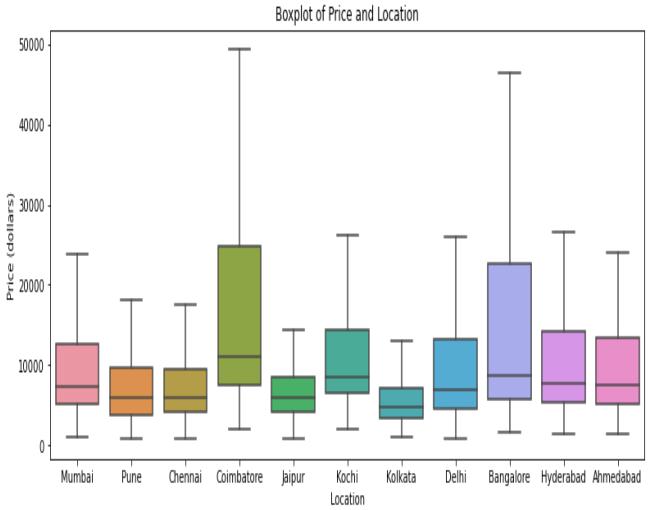

Fig. 8. Boxplot of Price and Location

Figure 8 shows that expensive cars are mostly in Coimbatore and Bangalore, while car prices in the other cities are generally similar.

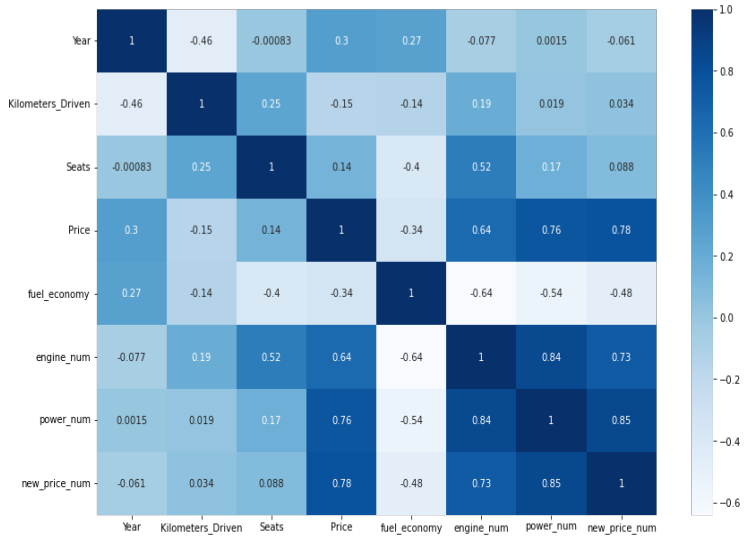

Fig. 9. Correlation Matrix

Figure 9 is a correlation matrix of the explanatory variables. New Price has a strong positive correlation with Engine and Power. Power and engine also have a strong positive correlation. New price and engine columns were dropped to remove multicollinearity.

V. Model

In this section, linear regression model is used for predicting used car prices using a collection of explanatory factors. The dataset was split into a training set, which included 70% of the data, and a test set, which included the remaining 30%. The dataset had been modified with dummy variables for categorical variables. The modeling process was conducted with the primary objectives of achieving accurate price predictions and understanding the influence of each feature on the target variable.

In addition, because the price was right-skewed, a logarithmic scale was used. This approach allows for the modeling of percentage changes in price, which can be valuable when examining relative price differences. The logarithmic transformation can reduce the influence of extreme values and skewness.

A. Model Accuracy:

The performance of this model was assessed using common evaluation metrics. The R-squared values for both the training and test data are notably high, with an R-squared of 0.912 for the training set and 0.914 for the test set. These values indicate that the model explains approximately 91% of the variance in used car prices. This suggests that the model provides a robust explanation of the factors affecting car prices and exhibits good generalization to unseen data.

The F-statistic, which tests the overall significance of the model, yielded a probability nearly approaching zero,

indicating that the model is statistically significant in explaining the variance in car prices.

The Mean Absolute Error (MAE), which measures the average absolute difference between the model's predictions and the actual prices, is approximately $2,347 in the test dataset. This suggests that, on average, the model's predictions deviate by $2,347 from the actual prices. This level of accuracy is significant in the context of used car pricing, providing a valuable tool for both buyers and sellers in making well-informed decisions.

B. Significant Predictors:

The p-values associated with the explanatory variables provide insights into the predictors that significantly influence the price of used cars. Variables such as 'Year,' 'Seats,' 'Power,' 'Fuel Economy,' 'Kilometers Driven,' 'Location,' 'Fuel Type,' 'Transmission,' and 'Car Category' all have p-values less than 0.05. This means that these variables are highly significant in predicting used car prices.

C. Variable Impact:

Further analysis reveals the magnitude of the impact of certain variables on used car prices. For instance, a one-unit increase in the manufacturing year results in a substantial 13% increase in the selling price of the car, all else being constant. Similarly, a one-unit increase in engine power results in a 0.5% increase in the selling price. Moreover, a one-unit increase in the fuel economy of the engine results in [exp(0.0182) - 1] * 100 $\approx$ 2% decrease in the selling price of the car when everything else is constant.

VI. Conclusion

In this study, linear regression model was used to predict used car prices. The model exhibited strong predictive power, explaining approximately 91% of the variance in car prices, and performed well on unseen data. Key predictors, including manufacturing year and engine power, were identified as significant influencers of price.

These findings empower buyers and sellers in the used car market, offering data-backed insights for making informed decisions. The model and results provide a valuable resource for assessing fair market values, negotiating prices, and ultimately enhancing transparency and efficiency in the used car market.

For future work, the intention is to develop Random Forest and Decision Tree models. These models will be compared against the existing Multiple Regression model to determine their effectiveness in predicting used car prices. The comparison aims to identify the strengths and weaknesses of each approach and potentially improve the predictive power of the analytical tools.

Reference

[1] Pal, N., Arora, P., Sundararaman, D., Kohli, P., & Palakurthy, S. S. (2017). How much is my car worth? A methodology for predicting used cars prices using Random Forest.

[2] Monburinon, N., Chertchom, P., Kaewkiriya, T., Rungpheung, S., Buya, S., & Boonpou, P. (2018). Prediction of prices for used car by using regression models. *2018 5th International Conference on Business and Industrial Research (ICBIR)*, 115-119.

[3] Pudaruth, S. (2006). Predicting the Price of Used Cars using Machine Learning Techniques.