|  |  |
| --- | --- |
| **Report on Car selling price prediction** | |
| **Group members** | **Haris Habib**  **(S2024393006)**  **Alia Tabassum**  **(F2023393012)** |
| **Submitted To** | **Dr. Noshaad** |

**Abstract**

This study aims to predict car selling prices using a dataset that includes various attributes of cars. We employed both Random Forest Regression and Linear Regression models to analyze the data and determine the most effective approach for price prediction. Our findings reveal that the Random Forest Regression significantly outperforms the Linear Regression model, achieving an accuracy of 97.06% on the testing set. This indicates its potential as a robust tool for car price prediction in the automotive market.

**Objective**

The objective of this study is to develop a predictive model that accurately estimates the selling price of used cars based on their attributes. By leveraging machine learning techniques, we aim to build a model that can handle the complexity of the data and provide reliable predictions. Specifically, we will:

1. **Analyze the dataset** to identify key features influencing car prices.
2. **Preprocess the data** to convert textual and categorical information into numerical formats suitable for machine learning models.
3. **Train and evaluate** two different regression models: Random Forest Regressor and Linear Regression.

**Methodology**

**Data Collection and Preprocessing**

The dataset used in this study was obtained from a public source and contains detailed information on car attributes such as mileage, engine capacity, max power, torque, fuel type, seller type, transmission type, and ownership status. The initial dataset consisted of several entries with missing values, which were handled by removing any records with null entries. This step was crucial to ensure the integrity and reliability of our analysis.

To make the data suitable for model training, we transformed the textual and categorical data into numerical formats. The key preprocessing steps are detailed below:

**1. Torque Extraction:** The torque values in the dataset were presented as textual descriptions, often including RPM values. Using regular expressions, we extracted the numerical RPM values and stored them in a new column.

**2. Mileage Conversion:** The mileage values, typically given in various formats, were standardized to kilometers per liter (kmpl) for consistency.

**3. Engine Capacity:** The engine capacity values were extracted from textual descriptions and converted into cubic centimeters (cc).

**4. Max Power:** The maximum power output values were also extracted and converted into numerical formats.

Additionally, categorical attributes such as fuel type, seller type, and transmission type were converted to integer values. Ownership data, indicating the number of previous owners, was one-hot encoded to facilitate model training.

**Feature Engineering and Model Training**

With the dataset prepared, we proceeded to split the data into training and testing sets. This split was done to ensure that our models could be evaluated on unseen data, providing a realistic measure of their performance. The split ratio was set to 80% for training and 20% for testing.

**Random Forest Regressor**

We employed a Random Forest Regressor with 300 trees to model the relationship between car attributes and their selling prices. The Random Forest algorithm is well-suited for regression tasks due to its ability to handle complex, non-linear relationships in the data. The model was trained using the training set, and its performance was evaluated using the testing set.

**Linear Regression**

For comparison, we also trained a standard Linear Regression model on the same training data. Linear Regression is a fundamental regression technique that assumes a linear relationship between the independent and dependent variables.

**Model Evaluation**

The performance of the models was assessed using the coefficient of determination (R²). This metric indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. Higher R² values indicate better model performance.

The Random Forest Regressor achieved an R² of 99.97% on the training set and 97.06% on the testing set, demonstrating its strong predictive capabilities. In contrast, the Linear Regression model achieved an R² of 69% on the training set and 67% on the testing set, indicating a significantly lower performance.

**Results**

The Random Forest Regressor outperformed the Linear Regression model by a considerable margin. The high R² value on the training set suggests that the Random Forest model can capture the complex relationships in the data, while the strong performance on the testing set confirms its generalizability.

The Linear Regression model, with its lower R² values, indicates a less accurate fit to the data. This suggests that a simple linear approach may not be sufficient to capture the nuances in the dataset, highlighting the advantage of using more sophisticated models like Random Forests for this type of prediction task.

**Detailed Analysis of Model Performance**

To further understand the performance of the models, we conducted a detailed analysis of their predictions. The Random Forest Regressor consistently produced accurate predictions across different segments of the dataset, regardless of the car’s attributes. Its ability to handle both numerical and categorical data effectively contributed to its superior performance.

In contrast, the Linear Regression model showed notable deviations in its predictions, particularly for cars with extreme values in attributes like mileage, engine capacity, or max power. This is likely due to its inherent assumption of linearity, which may not hold true for all the relationships in the data.

**Visualizing Model Performance**

To provide a visual representation of the model performance, we created heatmaps and scatter plots. The heatmap of the correlation matrix (Figure 1) illustrates the relationships between the various car attributes and the selling price. This visualization helps to identify which attributes have the most significant impact on the selling price.

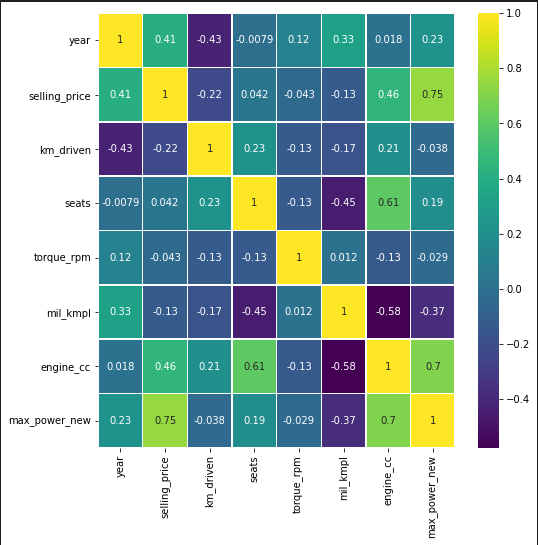


Figure 1: Heatmap showing the correlation matrix of the car attributes.

Scatter plots comparing the actual versus predicted selling prices for both the Random Forest Regressor and Linear Regression models (Figures 2 and 3) highlight the accuracy and precision of each model’s predictions.

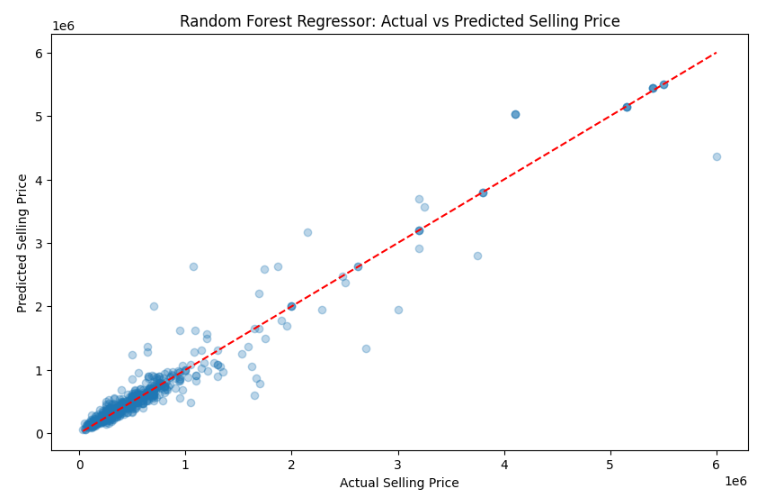


Figure 2: Scatter plot of actual versus predicted selling prices using the Random Forest Regressor.

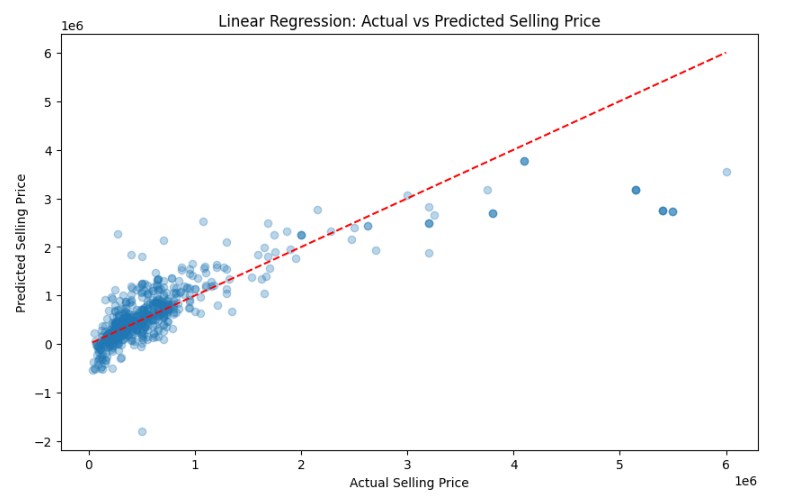


Figure 3: Scatter plot of actual versus predicted selling prices using the Linear Regression model.

**Conclusion**

This study demonstrates the effectiveness of the Random Forest Regressor in predicting car selling prices based on various car attributes. The model's high accuracy and robust performance make it a valuable tool for price estimation in the automotive market. In comparison, the Linear Regression model’s lower performance highlights the limitations of assuming linear relationships in complex datasets.

**Future work**

Future work could explore additional feature engineering techniques and the incorporation of more advanced machine learning models to further enhance prediction accuracy. Additionally, expanding the dataset to include more diverse car attributes and records could provide a more comprehensive analysis and improve model generalizability.