Predicting Used Car Prices using Machine Learning

Submitted By: Dheeraj Arora
(102103508)

Submitted To: Dr. Raman Goyal
(Assistant Professor)

Introduction and Problem Statement

- The used car market is extensive and highly variable, with prices influenced by numerous factors such as make, model, year of manufacture, mileage, and fuel type.
- Accurate price predictions can benefit both sellers and buyers. Sellers can set competitive and fair prices, while buyers can ensure they are getting a good deal.
- The *primary goal* of this project is to develop a machine learning model that can predict the price of a used car based on its features.
- This involves collecting and cleaning data, selecting relevant features, building and evaluating models, and refining them for better performance.

Data Overview and Cleaning

Description of the dataset:

- Number of entries - 15411
- Number of features - 13 (car_name, brand, model, vehicle_age, km_driven, seller_type, fuel_type, transmission_type , mileage, engine, max_power, seats, selling_price)
- Source of the data - https://www.kaggle.com/datasets/riddhivernekar/cardekho-dataset

Data cleaning steps

- Dropping Irrelevant Columns
- Removed rows with Z-scores exceeding a threshold of 3 to handle outliers.
- Converted the categorical column 'model' into numerical features using one-hot encoding
- Replaced categorical values in 'seller_type', 'fuel_type', and 'transmission_type' with numerical labels for model compatibility.
- Verified for missing values using car.isnull().sum().

Model Building

Model Used:

- *Linear Regression* - Chosen for its simplicity and interpretability. It's a commonly used algorithm for predicting continuous values like car prices.

Steps to Build the Model:

- Data Collection and Preparation: Gather and preprocess car data, handling missing values and encoding categorical variables.
- Feature Engineering: Identify and create relevant features that influence car selling prices.
- Model Selection: Choose a suitable regression model, starting with linear regression and exploring advanced options.
- Model Training: Split the data and train the model to learn the relationship between features and prices.
- Model Evaluation: Assess model performance using metrics like R-squared and mean squared error, and fine-tune hyperparameters.

Model Evaluation

- The model's performance was evaluated using the R^2 score.
- An R^2 score of ***0.83*** was observed which indicates that 83% of the variability in car selling prices can be explained by the model, signifying a good fit.

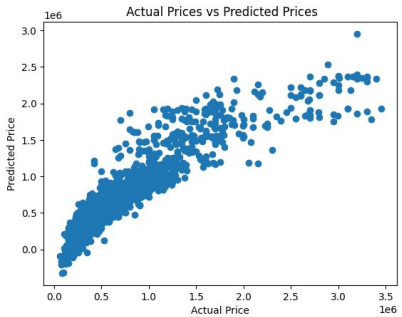

Fig.1 : Comparison between actual and predicted car selling price

Results and Future Work

Key Results:

- The model can predict car prices with a high degree of accuracy.
- **Best Achieved R^2 Score**: 0.83, indicating a good fit and reliable predictions.

Future Work:

- **Explore Other Machine Learning Algorithms**: Investigate more complex algorithms like Random Forests, Gradient Boosting, or neural networks to potentially improve prediction accuracy.
- **Include Additional Features**: Enhance the model by incorporating more features such as the car's location, number of previous owners, service history, and more, which might impact the price.
- **Deploy the Model as a Web Application**: Make the model accessible to a wider audience by deploying it as a web application. This would allow users to input car details and receive price predictions online.