Used Car Price Prediction

Name:	**Gargi Surendra Yeole**
Registration No./Roll No.:	21110
Institute/University Name:	IISER Bhopal
Program/Stream:	DSE
Problem Release date:	August 18 ,2022
Date of Submission:	November 19, 2023

1 Introduction

Car price prediction is the process of using data analysis and machine learning algorithms to forecast the fair market value of a vehicle. We want to predict car prices to provide transparency and accuracy in the used car market. This empowers buyers to make informed decisions, helps sellers set fair prices, and contributes to a more efficient and trustworthy automotive marketplace.The training dataset contains 5417 instances and 11 features and the test dataset contains 602 instances and 11 features.The dataset consists of a features like seats, model, year, mileage, fuel type, location,transmission, engine, power and owner Type.The datasets contains the missing values.Missing values in both training set and test set and those are mileage(1),engine(34),power(34),seats(38) and mileage(1),engine(2),power(2),seats(4) respectively. So I have imputed the missing values using the statistical method i.e using mean. Later on,I have change categorical values to numeric values using one hot encoding so that ML Model can work. Figure

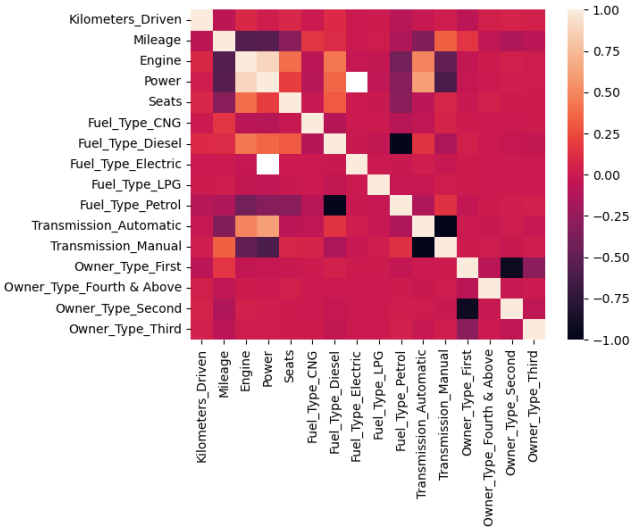

Figure 1: Correlation Matrix

2 Methods

The project explores various regression techniques to predict the prices of used cars based on different features. While classic regression methods such as Linear Regression, Decision Tree Regression, and K-Nearest Neighbors (KNN) Regression are part of the experimental analysis, more complex ensemble

methods like Random Forest Regression and Support Vector Regression (SVR) are also included for their potential to capture nonlinear relationships and improve prediction accuracy.

GitHub

Different Methods Explored for Experimental Analysis:

- Linear Regression: A straightforward method assuming a linear relationship between input features and the target variable, used as a baseline for comparison.
- Decision Tree Regression: A tree-based model that recursively partitions the feature space to make predictions.
- K-Nearest Neighbors (KNN) Regression: Predicts the target variable by averaging the values of the k-nearest neighbors in the feature space.
- Random Forest Regression: An ensemble learning method utilizing multiple decision trees to improve prediction accuracy.
- Support Vector Regression (SVR): Uses support vectors to create a hyperplane that best fits the data, particularly effective for non-linear relationships.

3 Experimental Setup

Evaluation Criteria:

- Mean Squared Error (MSE): Measures the average squared differences between predicted and actual prices.
- Root Mean Squared Error (RMSE): Represents the square root of MSE, providing an interpretable measure of prediction error in the same units as the target variable.
- R-squared (R2): Quantifies the proportion of variance in the target variable explained by the model.

Significant Parameters or Hyperparameters:

- Linear Regression: No significant hyperparameters need tuning as it involves fitting a linear relationship.
- Decision Tree Regression: Parameters like max depth, min samples split regulate the tree's depth and structure.
- Random Forest Regression: Key parameters include n estimators (number of trees) and max depth (maximum depth of trees) and max features (sqrt,auto,log2)
- Support Vector Regression (SVR): Hyperparameters include the choice of kernel type (kernel), C (regularization parameter), and gamma (kernel coefficient)
- K-Nearest Neighbors (KNN) Regression: Crucial hyperparameter is the number of neighbors (n neighbors)and other hyperparameters are weight(uniform,distance) and p

Libraries Used:

The project leverages the scikit-learn library in Python for implementing various regression models.Scikit-learn provides extensive tools for machine learning, encompassing regression algorithms, hyperparameter tuning methods by GridsearchCV and model evaluation metrics.

Additional supporting libraries, such as pandas for data manipulation, numpy for numerical operations, and matplotlib/seaborn for data visualization, complement the modeling process.

Table 1: Performance Of Different Classifiers Using All Features

Regressor model	MAE	RMSE	MSE	R2
Linear Regression	4.5323	7.4486	55.4830	0.5550
Random Forest Regressor	0.9100	1.8442	3.4013	0.9727
K-Nearest Neighbor	0.0409	0.3937	0.1550	0.9988
Decision Tree	2.4784	4.6749	21.8552	0.8247
Support Vector Regressor	4.3409	8.7705	76.9226	0.3830

4 Results and Discussion

Observations:

- K Nearest Neighbor outperforms other models: It exhibits the lowest MAE, RMSE, and MSE, indicating better accuracy in predicting used car prices. Additionally, it has the highest R2 score, suggesting that it explains the variance in the target variable more effectively compared to other models.
- Support Vector Regressor is not the good model for for my project as it has highest MAE, RMSE and MSE score.
- Random Forest Regressor: It follows K Nearest Neighbor (KNN) closely in terms of performance metrics, indicating good predictive ability. It achieves a lower performance than KNN but still outperforms Linear Regression, Support Vector, and models.

5 Conclusion

The primary aim of this project was to identify the most effective predictive model for used car price prediction among the chosen machine learning techniques. After evaluating all the models, the analysis indicates that the K-Nearest Neighbor model emerges as the optimal choice for predicting used car prices. The K-Nearest Neighbor model consistently demonstrates the ability to accurately forecast prices by effectively capturing the intricate relationships among various features. With the same set of hyperparameters, the KNN model showcases exceptional performance, achieving high accuracy (low MAE, RMSE, MSE score)in predicting prices, indicating its superiority over other models.

6 References

Towards DataScience

Wikipedia

Used Car Price Prediction Paper