Used Car Price Prediction

Name:	**Chavan Vishwaraj Sopan**
Registration No./Roll No.:	21086
Institute/University Name:	IISER Bhopal
Program/Stream:	DSE
Problem Release date:	February 02, 2022
Date of Submission:	November 18, 2023

1 Introduction

The primary objective of this project is to develop a robust framework that can accurately predict the price of a given used car based on its various features. The dataset contains information about used cars of different brands of India. The columns in the given dataset are as follows:

Brand,Location,Year,Km Driven,Fuel Type,Transmission, Owner Type,Mileage,Engine,Power,Seats

It is a regression problem not a classification problem, as the goal is to predict a continuous value (the price in this case) rather than class labels.Therefore, there are no distinct classes here. We were provided with three datasets, namely

- training data
- training data targets
- test data

We used various Machine Learning algorithms and selected the best model for our datasets.

2 Methods

The first step involved the preprocessing of data and data cleaning. We first loaded the training and training data targets and combined them.Then we performed mean inputing for dealing with the missing values. Then we dropped the rows with null values and checked for outliers. After that, we removed outliers for various features. Then to make the target variable linear with respect to features, we took the log of instances (log Price). The second step involved Encoding of categorical columns. For Brand and Location columns, we used Target Encoding and for Fuel Type, Transmission, and Owner Type, we used One Hot Encoding. In the third step, we performed various machine learning algorithms with hyperparameter tuning using Grid SearchCV.

3 Experimental Setup

We used 8 Regression algorithms with Hyperparameter tuning with GridSearchCV

1. Linear Regression
2. Decision Tree Regression
3. Random Forest Regression
4. Ridge Regression

5. Support Vector Regression

6. K Nearest Neighbour Regression

7. Gradient Boosting Regression

We used these seven algorithms.
First, we had to perform encoding techniques. We performed target encoding on Brand and Location columns as there were many unique categorical values. On Fuel, Transmission, and Owner columns, we performed One Hot Encoding.

Then we performed Hyperparameter tuning and implemented our algorithms and recorded the performance.

4 Results and Discussion

The correlation matrix for training data is

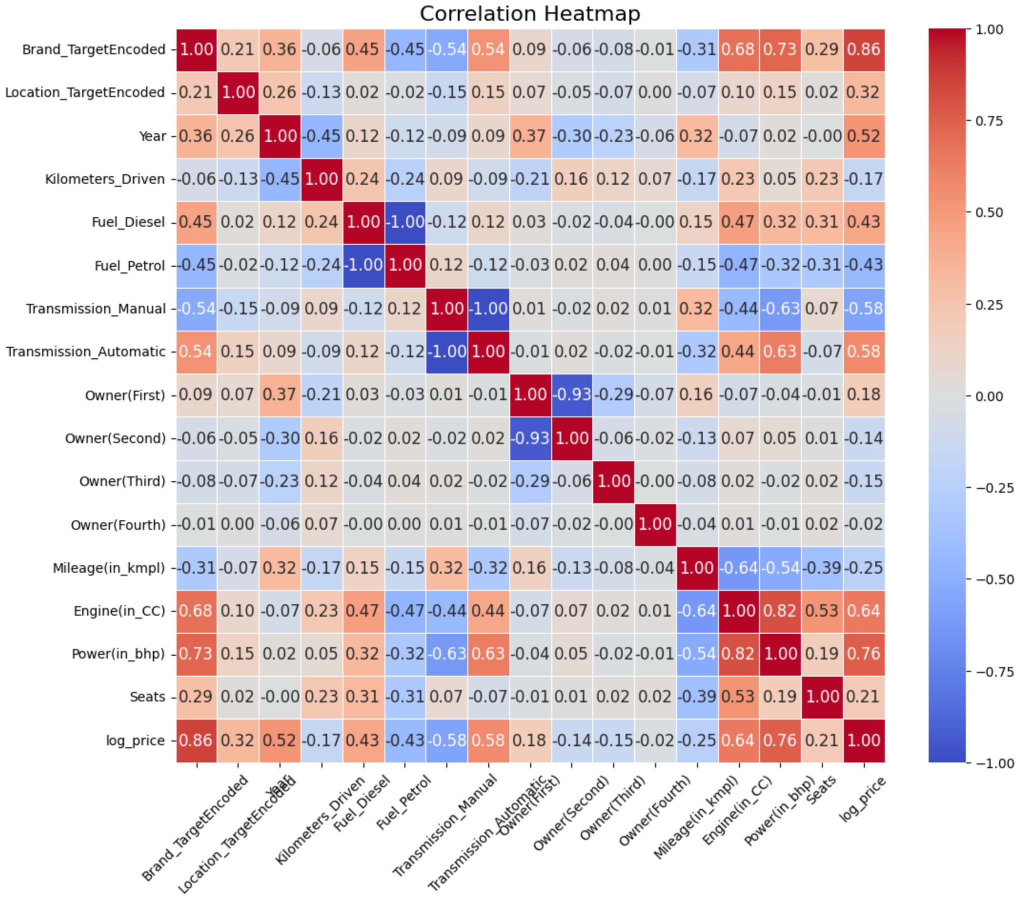

Figure 1: Correlation Matrix for Training Data

After Implementing our model we had the following performance scores for different models

Table 1: Performance Of Different Regressors

Regressor	**Best Hyperparameter**	**MSE**	**R-Square**
Linear Regression	fit intercept:True	10.62	0.89
Decision Tree Regression	max depth:None	18.11	0.84
Random Forest Regression	max depth:None,n estimators:200	4.06	0.95
Ridge Regression	alpha:10	10.66	0.89
Support Vector Regression	C:10,kernel:rbf	6.78	0.93
K Nearest NeighbourRegression	nneighbors:5,weights:distance	8.2	0.91
Gradient Boosting Regression	learning rate: 0.1,max depth: 5,n estimators: 200	3.4	0.96

After implementing our models Gradient Boosting Regression Model is the best our model with following scores

- **Best Hyperparameter**—learning rate: 0.1,max depth: 5,n estimators: 200
- **Mean Square Error**—2.91
- **R Square**—0.96

The plot of Actual Price vs Predicted price for Gradient Boosting Regression is

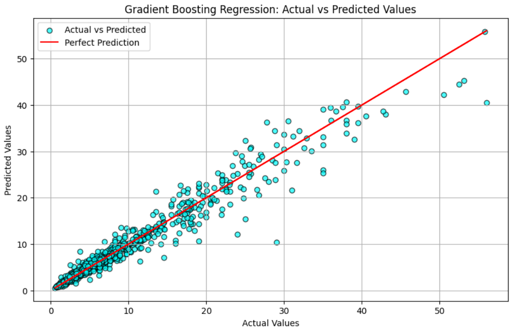

Figure 2: Plot for Gradient Boosting without PCA

5 Conclusion

The Gradient Boosting is best for our model.

References

Aurelien Geron. Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow: Concepts, Tools and Techniques to Build Intelligent Systems
Data Science Projects with Python: A Case Study Approach to Successful Data Science Projects Using Python, Pandas, and Scikit-learn
Scikit-Learn MLin algorithm