Car Price Prediction

1.Objective:

Predicting the price of the car. Maximizing the r2 score with different regression algorithms.

2.Data collection and Data Pre-processing:

- The given car price dataset consists of 205 rows and 26 columns.
- Data Cleaning: Find the missing values, outliers and inconsistencies in the data and should be cleaned before giving to the model.
- **Exploratory Data Analysis (EDA):** Understand the characteristics, distributions and the relationships among variables to get insights about the data. This is to be done through visualisation of the data through several plotting methods like boxplot, count plot, bar plot, scatter plot etc.

3.Feature Engineering:

- There are 26 columns in the dataset it is very important to find the relevant features from the dataset. The features which have significant impact on the car prices based on the domain knowledge or statistical analysis are selected.
- **Feature transformation**: In this model, the Standard Scaler to rescale the data to standardize the distribution of the data was used. The standard scaler rescales the data to a mean of 0 and standard deviation to 1.
- **Feature Extraction**: Principal component analysis is used for the feature extraction. The most important features are extracted out. The explained variance is the percentage of variance that is attributed to each of the selected components.

4.Splitting the dataset:

- Divide the prepared dataset into training and testing data. Here in this model, I used split 80-20 rule.

5.Model Selection and Training:

- In our dataset, the target variable is of continuous data so the regression algorithms were used. The model is trained with Linear Regression, KNN regressor, SGD Regressor, Random Forest Regressor, and Decision Tree Regressor.
- Then, train the model using the Training dataset.

6.Model Evaluation and Fine-tuning:

- Evaluate the training model using the testing dataset to assess its performance
- The metrics used to evaluate the model are Mean squared error, Mean absolute error and r2score. Our aim is to minimise the error and maximise the score.
- The high performing model is further tuned by adjusting the hyperparameters to improve the model's performance.

7.Prediction and Interpretation:

- The ensemble models, such as random forest regressor, Decision tree Regressor, SGD Regressor are the outperformed individual models.
- After hyperparameter tuning, the Random Forest Regressor gave the highest R2 score of about 74%.

Final Model: Random Forest Regressor

From the present analysis, I suggest the car companies to use the Random Forest Regressor model for predicting the Pricing strategy, Market Analysis, Sales forecasting, Competitive analysis and in optimizing the price.

Language and libraries used:

Python - Language

Pandas - Data exploration

Numpy - for performing mathematical operations and for numerical computing

Matplotlib – Data Visualisation

Scikit-learn – Machine learning tasks.