Car Price Prediction Project - Analysis & Insights

1. Data Preprocessing

Steps Taken:

- Loaded the dataset and checked for missing values.
- Handled missing or incorrect values appropriately.
- Encoded categorical variables (e.g., car brands, fuel types) using one-hot encoding.
- Standardized/normalized numerical features for better model performance.

Insights:

- Proper data preprocessing ensured that the dataset was clean and ready for machine learning models.
- Encoding categorical variables allowed the models to interpret non-numeric features effectively.

2. Outlier Detection & Treatment

Steps Taken:

- Used boxplots to visualize and detect outliers in numerical variables.
- Capped extreme values using percentile-based capping or transformation techniques.

Insights:

- Outlier detection helped in identifying extreme values that could negatively affect model performance.
- Capping ensured that the dataset maintained its integrity while avoiding overfitting to extreme values.

3. Feature Selection

Steps Taken:

- Initially used a correlation heatmap but found it ineffective for feature selection.
- Applied feature importance using the Random Forest Regressor to identify significant variables affecting car prices.

Insights:

- The correlation heatmap was not a reliable method for feature selection due to multicollinearity among variables.
- Feature importance scores from Random Forest revealed that engine size, horsepower, curb weight, and fuel type had a strong impact on car price predictions.

4.Model Implementation & Evaluation

Steps Taken:

Implemented five regression models:
- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor
- Gradient Boosting Regressor
-Support Vector Regressor (SVR)

- Split the dataset into training (80%) and testing (20%) sets.
- Evaluated models based on:
  - R-squared (R^2) → Measures how well the model explains variance.
  - Mean Squared Error (MSE) → Measures the average squared difference between predicted and actual prices.
  - Mean Absolute Error (MAE) → Measures the average absolute difference between predicted and actual prices.

Insights:

- Random Forest Regressor performed the best, achieving an R^2 of 0.958 with the lowest MSE and MAE, making it the most accurate model.
- Gradient Boosting Regressor was the second-best, performing slightly lower than Random Forest.
- Linear Regression & Decision Tree Regressor showed moderate accuracy, capturing some trends but not as powerful.
- Support Vector Regressor performed the worst, failing to learn meaningful patterns from the data.

5. Hyperparameter Tuning

Steps Taken:

- Performed hyperparameter tuning on Random Forest using GridSearchCV to optimize parameters like **n_estimators, max_depth**, and **min_samples_split.**
- Evaluated the tuned model's performance.

Insights:

- The tuned Random Forest model achieved $R^2 = 0.9585$, with improved MSE and MAE, confirming that tuning enhanced accuracy.
- The model generalized well, reducing errors and improving prediction reliability.

Final Takeaways

Key Factors Affecting Car Prices:

- Engine Size, Horsepower, Curb Weight, and Fuel Type were the most important features impacting car prices.

Best Model for Car Price Prediction:

- Random Forest Regressor provided the best accuracy and lowest error rates, making it the most reliable model for this dataset.

⚙ Impact of Hyperparameter Tuning:

- Fine-tuning improved model performance, reducing errors while maintaining high predictive power.