RISHIKA RAVICHANDRAN

DATA SCIENCE INTERN

SAPIENCE EDU CONNECT PVT. LTD.

WEEK 2

**ANALYSIS REPORT**

1. **Evaluation Metrics for Each Model**

* Linear Regression
* Mean Squared Error (MSE): 94596219903.38426
* R-Squared (R²): 0.5686872253788728
* Mean Absolute Error (MAE): 173245.8345193564
* Logistic Regression
* Accuracy: 0.8636363636363636
* Precision:0.8693759071117562
* Recall: 0.8581661891117478
* F1-Score: 0.8637346791636625
* Decision Tree
* Mean Squared Error (MSE): 0.13131313131313133
* R-Squared (R²): 0.4747201306057176
* Mean Absolute Error (MAE): 0.13131313131313133
* Training Accuracy: 0.9093862815884477
* Testing Accuracy: 0.8686868686868687
* Precision: 0.8623595505617978
* Recall: 0.8796561604584527
* F1-Score: 0.8709219858156029
* Random Forest
* Mean Squared Error (MSE): 18248431290.3928
* R-Squared (R²): 0.9167960248265615
* Mean Absolute Error (MAE): 77372.3264428092

Suitable metric for each task

* Regression:
* Mean Squared Error (MSE): Measures the average squared difference between actual and predicted values. A lower MSE indicates better model performance, but it is sensitive to large errors due to squaring.
* R-Squared (R²): Represents the proportion of variance explained by the model. A higher R² (closer to 1) indicates better predictive power.
* Mean Absolute Error (MAE): Measures the average absolute difference between actual and predicted values. Unlike MSE, MAE is less sensitive to large errors.
* Classification:
* Accuracy: The proportion of correctly predicted labels. It can be misleading if classes are imbalanced.
* Precision: The proportion of true positive predictions out of all positive predictions.
* Recall: The proportion of true positives out of all actual positives.
* F1-Score: The harmonic mean of precision and recall, balancing both measures. It is useful when false positives and false negatives are equally important.

1. **Insights from Feature Importance**

* Linear Regression:
* Most influential features: max\_power, year, engine
* Negative influence: transmission, km\_driven(categorical variables)
* Max Power (268,635.64) and Year (129,357.46) had the highest positive impact, meaning newer cars with higher power tend to have higher prices.
* Engine Capacity (34,222.22) and Mileage (13,221.03) also positively influenced price.
* Negative coefficients for Owner (-14,990.76), Fuel Type (-38,289.05), and Transmission (-90,649.32) indicate that cars with more previous owners, certain fuel types, and manual transmissions tend to have lower prices.
* Logistic Regression:
* The model correctly classifies most cars as high or low based on the median price threshold.
* A high accuracy suggests that the model is well-trained and generalizes effectively on the dataset.
* The balance between precision and recall suggests that both false positives and false negatives are minimized.

1. **Insights from Model comparisons:**

Decision Tree vs. Random Forest:

* The Random Forest model performed better than the Decision Tree due to ensemble learning, reducing overfitting.
* Feature importance in Random Forest showed similar trends to Decision Tree but with more refined weight distribution.

1. **Most suitable model for this dataset:**

* Regression: Random Forest Regressor was the best-performing model. It has:
* Higher R² and lower MSE compared to other models.
* Ability to handle both linear and non-linear relationships effectively.
* Reduced overfitting compared to the Decision Tree.
* Classification: Logistic Regression was best.
* Simple and interpretable
* Less prone to overfitting