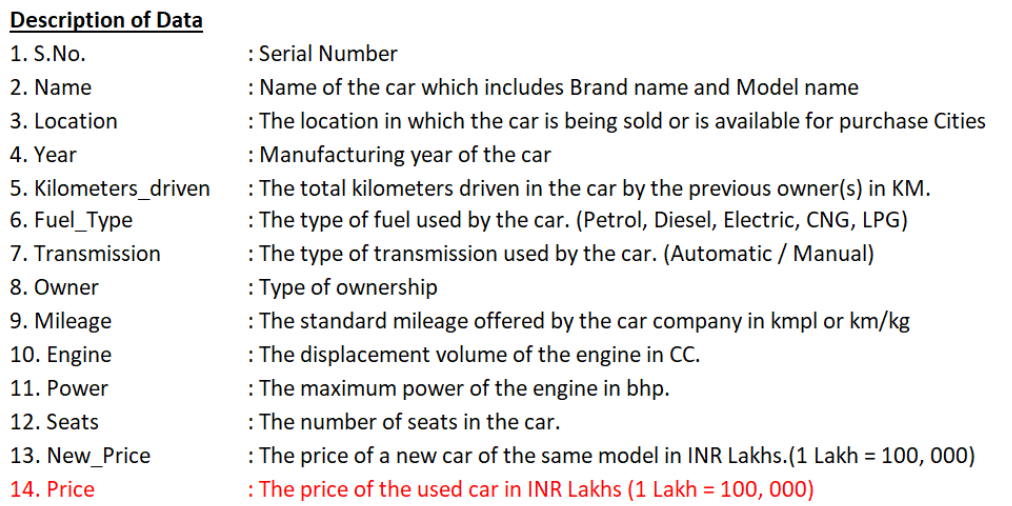
# **Used Car Price Prediction Report**

**1. Introduction**

**This report summarizes the findings and methodology used in predicting used car prices using various regression models. The goal was to build a robust predictive model by evaluating several algorithms, interpreting results using statistical and machine learning tools, and drawing business-relevant conclusions.**

**2. Dataset Properties**

****

* **Target Variable: Price, with a log-transformed version Price\_log used for regression modeling.**
* **Feature Engineering:**
  + **One-hot encoding applied to categorical variables.**
  + **Multicollinearity reduction using VIF analysis.**
  + **Standardization of numeric features using StandardScaler.**

**3. Models Evaluated**

**Nine regression models were evaluated:**

* **LinearRegression**
* **DecisionTree**
* **RandomForest**
* **AdaBoost**
* **KNN**
* **XGBoost**
* **SVR**

**A separate OLS Regression model was also built using statsmodels for statistical inference.**

**Hyperparameter tuning was done using GridSearchCV with 5-fold cross-validation.**

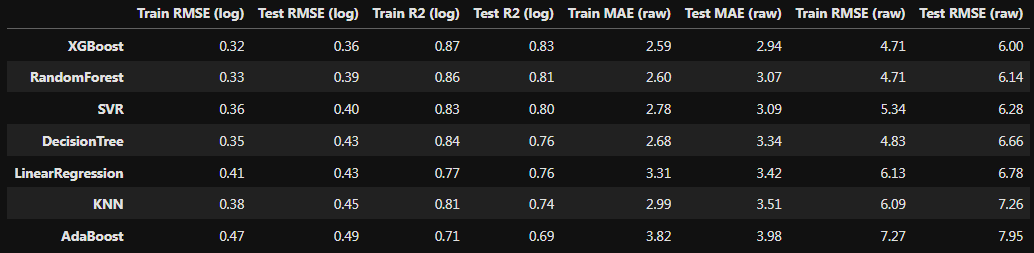
**4. Evaluation Metrics**

**Models were evaluated on:**

* **Log-scale metrics:**
  + **RMSE (log)**
  + **R2 Score (log)**
* **Raw-scale metrics (after inverse log transformation):**
  + **MAE (raw)**
  + **RMSE (raw)**

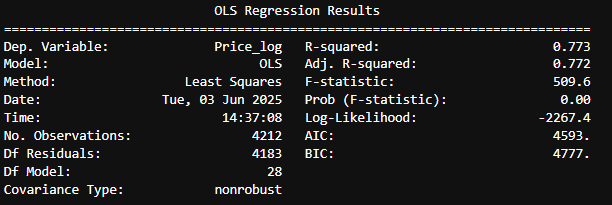
**5. Model Comparison Summary (Test Set)**

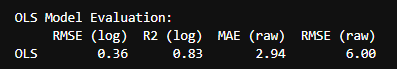
**The best performing model based on RMSE (raw) was highlighted.**

****

**6. OLS Model Results (Separate)**

**OLS was used primarily for statistical diagnostics and interpretability. Despite being similar to LinearRegression, OLS provides coefficient significance, R-squared, and confidence intervals.**

****

****

**7. Residual Analysis**

**Histogram plots of residuals for each model show how prediction errors are distributed. Normally distributed residuals with low variance suggest better model performance.**

**8. SHAP Interpretability**

**SHAP values were computed for:**

* **RandomForest**
* **XGBoost**

**Visualizations:**

* **SHAP Summary Plot**
* **SHAP Bar Plot**
* **SHAP Dependence Plot**

**These plots highlighted which features had the most influence on price predictions.**

**9. Conclusion & Inferences**

* **Ensemble models (RandomForest, XGBoost) outperformed linear models.**
* **SHAP analysis helped identify top contributing features.**
* **OLS provided statistical confidence in feature importance.**
* **The model pipeline is robust and scalable.**

**Recommendations:**

* **Use the XGBoost model in production if interpretability is less critical.**
* **Use RandomForest for a balance between performance and explainability.**

**Appendix**

* **All models saved using joblib.**
* **Pipeline supports future integration with new data or deployment systems.**

**End of Report**

**2025-06-03**