**CAR DHEKO - USED CAR PRICE PREDICTION**

**Project Overview**

The goal of this project was to develop a machine learning model to predict used car prices. The project involved several key steps, including data cleaning, feature selection, model training, and evaluation. An interactive Streamlit application was also developed for real-time price predictions based on user input.

**Data Cleaning and Preprocessing**

**1. Loading and Initial Cleanup:**

- The dataset was loaded and irrelevant or redundant columns were dropped. These included columns that were not useful for predicting prices or had excessive missing values.

- Columns with more than 80% missing values were removed to ensure the quality of the dataset.

**2. Handling Missing Values:**

- For numeric columns, missing values were imputed using the mean for normally distributed data and the median for skewed data.

- For categorical columns, missing values were replaced with the mode (most common value).

**3. Data Transformation:**

- Price values, initially in string format with currency symbols, were converted to numeric values.

- Various columns that contained numerical data as strings (e.g., ‘Alloy Wheel Size’, ‘Displacement’) were cleaned and converted to numeric formats.

**4. Outlier Detection and Removal:**

- Outliers were identified using the Interquartile Range (IQR) method. Extreme values were capped to minimize their impact on the model.

- Columns with a single unique value were dropped as they did not provide useful information.

**5. Encoding Categorical Variables:**

- Categorical variables were converted to numerical values using label encoding to prepare them for machine learning algorithms.

**Feature Selection**

**1. Feature Importance:**

- A Random Forest Regressor was used to determine the importance of each feature in predicting car prices.

- The top 15 most important features were selected based on their contribution to the model’s predictions.

**2. Feature Reduction:**

- The dataset was reduced to include only the most important features, simplifying the model and improving interpretability.

**Model Development and Evaluation**

**1. Model Training:**

- Several machine learning models were trained and evaluated, including:

- Linear Regression: A simple model that assumes a linear relationship between features and the target variable.

- Decision Tree Regressor: A model that splits the data based on feature values to make predictions.

- Random Forest Regressor: An ensemble of decision trees that improves prediction accuracy by averaging multiple tree outputs.

- Gradient Boosting Regressor: An advanced ensemble method that builds models sequentially to correct errors made by previous models.

**2. Hyperparameter Tuning:**

- The Random Forest and Gradient Boosting models were fine-tuned using RandomizedSearchCV to find the best combination of parameters for improved performance.

**3. Model Evaluation:**

To assess and compare the performance of various regression models for predicting car prices, we examined three critical metrics:

1. **Mean Absolute Error (MAE):** Reflects the average magnitude of errors in predictions, without considering their direction. A lower MAE indicates better predictive accuracy.

2. **Mean Squared Error (MSE):** Measures the average of the squares of the errors. It emphasizes larger errors due to squaring, making it sensitive to outliers. Lower MSE indicates better performance.

3. **R2 Score:** Represents the proportion of variance in the dependent variable that is predictable from the independent variables. A higher R2 Score indicates a better fit of the model to the data.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **MAE** | **MSE** | **R2 Score** |
| Linear Regression | 1.34e+05 | 3.38e+10 | 0.825 |
| Decision Tree | 8.99e+04 | 2.32e+10 | 0.880 |
| Random Forest (Tuned) | 6.56e+04 | 1.04e+10 | 0.946 |
| Gradient Boosting (Tuned) | 6.26e+04 | 9.03e+09 | 0.953 |

**Gradient Boosting Regressor (Tuned):**

-MAE: 62,639.89 – This is the lowest MAE among the models, indicating the smallest average error in predictions.

- MSE: 9,027,908,000 – The smallest MSE shows that it has the least deviation from the actual values, with fewer large errors compared to other models.

- R2 Score: 0.953 – The highest R2 Score means this model explains the most variance in the target variable, demonstrating the best overall fit.

**Advantages:**

- High Accuracy: Provides the most accurate predictions with minimal average error.

- Robust Performance: Effective in handling complex datasets with non-linear relationships.

**Disadvantages:**

- Computational Complexity: May require more computational resources and time to train compared to simpler models.

**Random Forest Regressor (Tuned):**

- MAE: 65,579.35 – Slightly higher MAE than Gradient Boosting, indicating a bit more average error in predictions.

- MSE: 10,399,920,000 – Also slightly higher MSE, reflecting more variance in prediction errors.

- R2 Score: 0.946 – Very close to the Gradient Boosting model, indicating strong predictive performance.

**Advantages:**

- Good Performance: Excellent performance with good accuracy and relatively low error metrics.

- Feature Importance: Provides insights into feature importance, which can be useful for understanding model behavior.

**Disadvantages:**

- Less Accurate than Gradient Boosting: Slightly less accurate than Gradient Boosting in capturing complex patterns.

**Decision Tree Regressor:**

- MAE: 89,988.72 – Higher MAE compared to ensemble methods, indicating larger average prediction errors.

- MSE: 23,168,780,000 – Higher MSE shows more variance and larger errors in predictions.

- R2 Score: 0.880 – Lower R2 Score than ensemble methods, showing less explanatory power of the variance in the target variable.

**Advantages:**

- Interpretability: Easy to interpret and understand, with straightforward decision rules.

- Simple and Fast: Relatively fast to train and predict compared to more complex models.

**Disadvantages:**

- Overfitting: Prone to overfitting, especially with complex datasets, leading to poorer generalization on unseen data.

**Linear Regression:**

- MAE 134,256.47 – Highest MAE among all models, indicating the largest average error in predictions.

- MSE: 33,757,930,000 – Highest MSE, reflecting the highest variance in errors and sensitivity to outliers.

- R2 Score: 0.825 – Lowest R2 Score, indicating that this model explains the least amount of variance in the target variable.

**Advantages:**

- Simplicity: Simple to understand and implement. Fast to train and predict.

- Baseline Model: Good as a baseline model to compare with more complex models.

**Disadvantages:**

- Limited Complexity: May not capture complex relationships in the data, leading to less accurate predictions.

**Conclusion**

The Gradient Boosting Regressor (Tuned) stands out as the best model for predicting car prices due to its superior accuracy, minimal error, and strong ability to capture complex patterns in the data. Random Forest Regressor (Tuned) also performs very well, though slightly less accurately than Gradient Boosting. Decision Tree Regressor and Linear Regression offer less accuracy and performance but may be useful in simpler scenarios or as baseline models.

This analysis helps in understanding the relative strengths and weaknesses of each model, guiding you towards choosing the best model for your specific prediction needs.

**Interactive Application**

1**. Streamlit Application:**

- An interactive Streamlit app was developed to allow users to input car features and receive real-time price predictions based on the trained Gradient Boosting model.

- The app includes sliders for input features, providing an intuitive interface for users.

**Results and Visualization**

**1. Feature Importance:**

- A bar plot of feature importances highlighted which features most significantly impact car prices.

**2. Correlation Heatmap:**

- A heatmap was used to visualize correlations between features, helping to understand relationships in the data.

**3. Hypothesis Testing**:

- Statistical tests were conducted to examine relationships between features and their significance in predicting car prices.

This structured approach ensures that the model is robust, the data is clean, and the predictions are accurate and interpretable.

