AI-Based Sri Lankan Used Car Price Prediction

Suhira Balarajan – 214206G

1. Introduction

The objective of this project is to develop a machine learning model capable of predicting used car prices in Sri Lanka using structured listing data. Vehicle prices in Sri Lanka vary significantly depending on brand, model, year of manufacture, mileage, transmission type, fuel type, location, and available features. Current pricing practices are largely subjective and inconsistent.

This project aims to:

- Build an accurate regression model for price prediction
- Apply systematic preprocessing and feature engineering
- Evaluate performance using proper validation methods
- Integrate explainability using SHAP
- Deploy the model using Streamlit and Docker

The task is formulated as a **regression problem**, where the target variable is Price.

2. Dataset Description

2.1 Data Source and Collection

The dataset was compiled by scraping car listings from ikman.lk and Riyasewana.com, which are the two most prominent online vehicle marketplaces in Sri Lanka. This multi-source approach ensures a comprehensive representation of the local market, capturing a wide variety of vehicle types and pricing behaviors across different regions of the country.

2.2 Dataset Scale

The dataset consists of 9,788 rows and 16 columns. This scale provides a robust foundation for training the XGBoost regressor, allowing the model to learn complex patterns between vehicle features and their market value while maintaining high statistical significance.

2.3 Features of the dataset

The dataset contains structured used vehicle listings with the following features:

- Brand
- Model
- YOM (Year of Manufacture)
- Engine (cc)
- Gear
- Fuel Type
- Millage(KM)
- Town
- Leasing
- Condition
- AIR CONDITION
- POWER STEERING
- POWER MIRROR
- POWER WINDOW
- Date (used for feature extraction)
- Price (Target Variable)

2.4 Feature Engineering

The following features were derived:

- ListingYear (from Date)
- ListingMonth (from Date)
- VehicleAge = 2025 − YOM

3. Data Preprocessing

The following preprocessing steps were applied:

3.1 Data Cleaning

- Removal of duplicate rows
- Dropping rows with missing target values
- Conversion of numeric columns to proper numeric format

3.2 Outlier Removal

Extreme price outliers were removed using the **IQR method**:

$$Q1 - 1.5IQR \leq Price \leq Q3 + 1.5IQR$$

Equation 1: IQR method

This helped reduce distortion in model training.

3.3 Encoding and Scaling

Preprocessing was implemented using a Scikit-Learn Pipeline and ColumnTransformer:

- Numerical features:
    - Median imputation
    - Standard scaling
- Categorical features:
    - Most frequent imputation

- One-hot encoding

All transformations were integrated into the model pipeline to ensure reproducibility and deployment compatibility.

4. Exploratory Data Analysis (EDA)

4.1 Target Variable Distribution

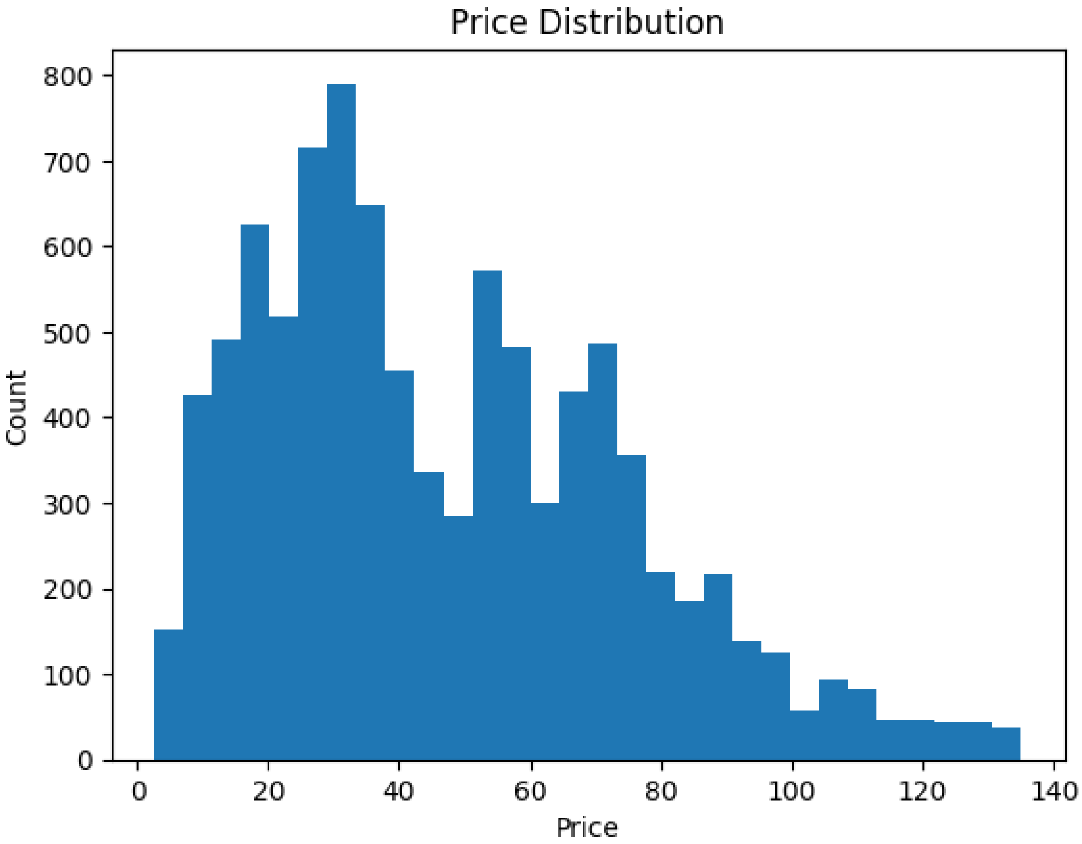

Figure 1: Price Distribution Histogram

Observation:

- The distribution shows variability across price ranges.
- Outlier removal reduced extreme distortions.

4.2 Numerical Feature Distributions

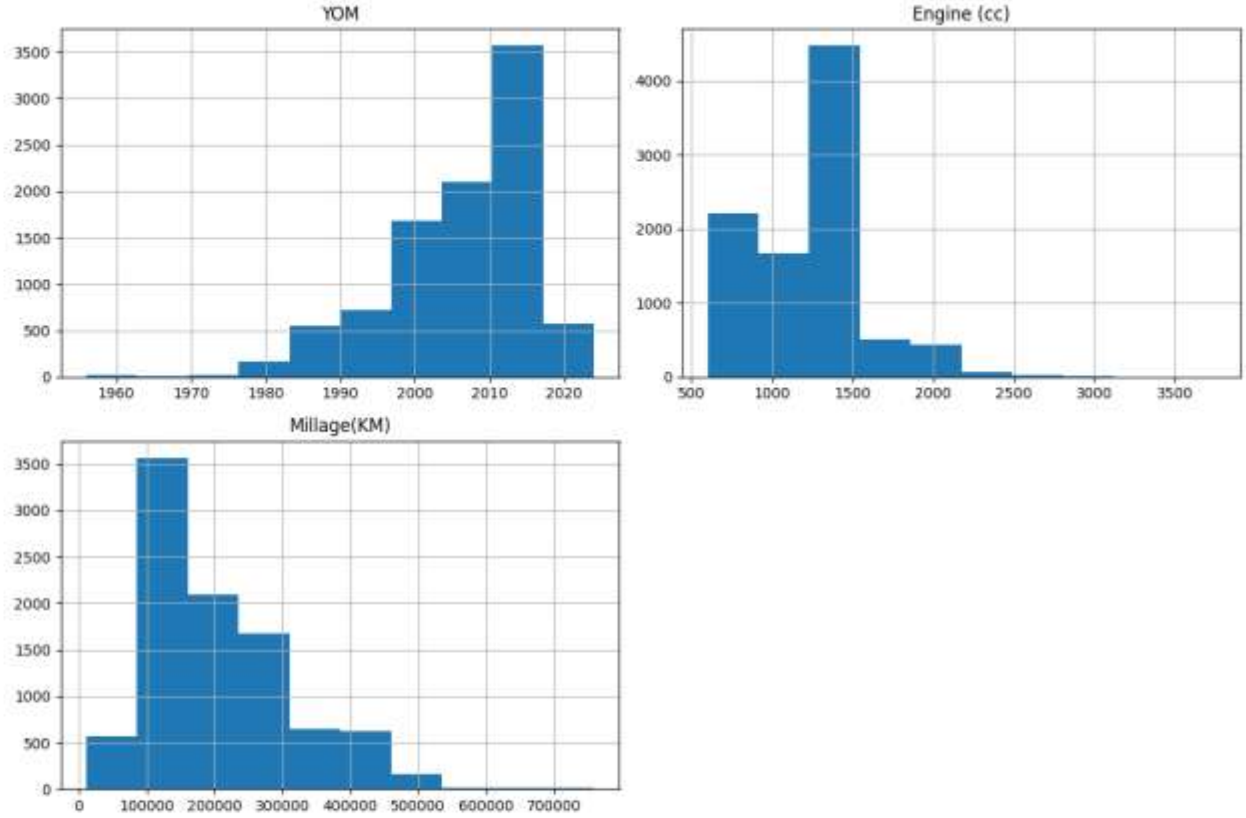

Figure 2:Numeric Feature Histograms

Observation:

- Mileage and engine capacity show wide variation.
- YOM shows clustering around recent years.

4.3 Correlation Analysis

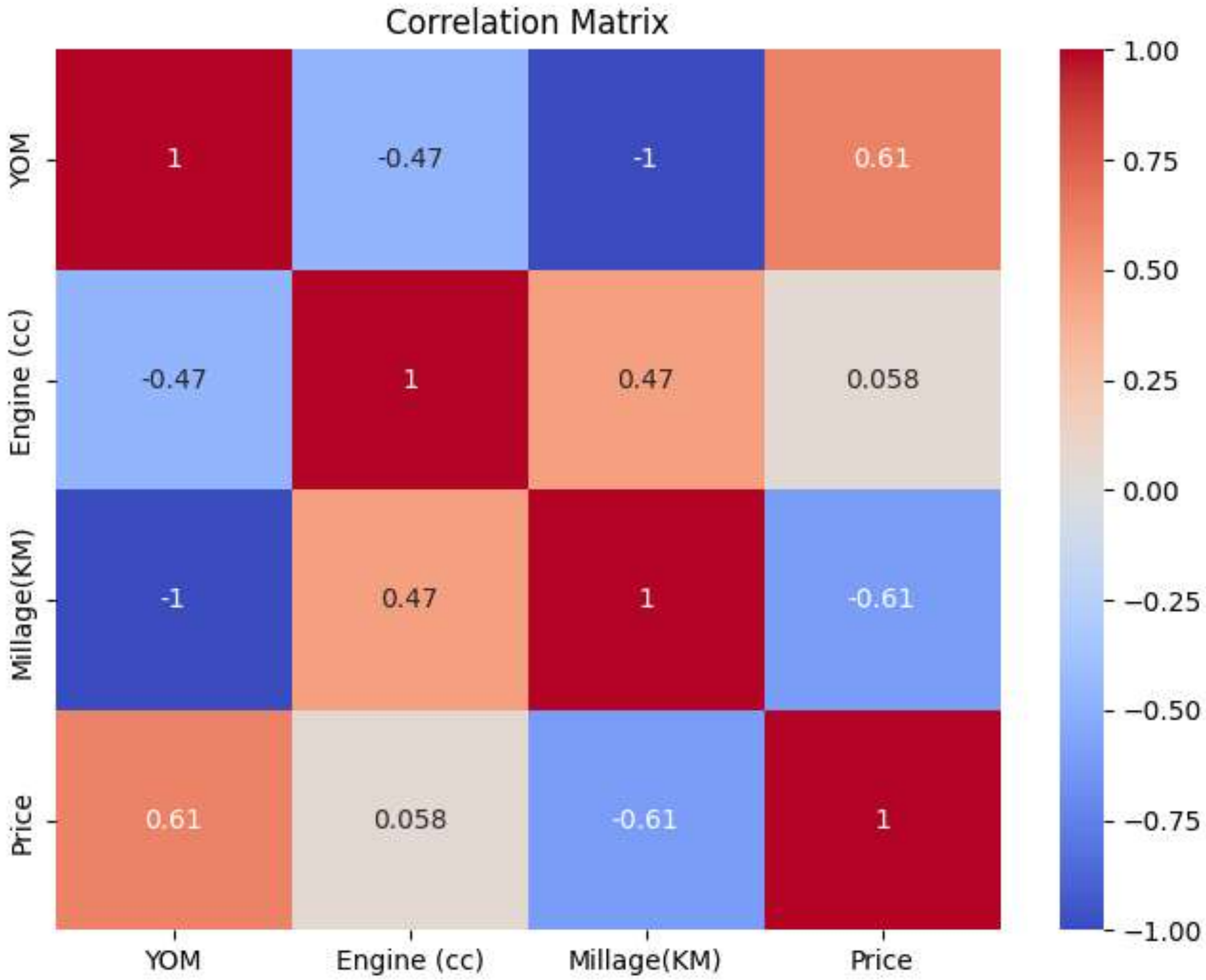

Figure 3:Correlation Heatmap

Observation:

- VehicleAge and YOM show meaningful relationship with Price.
- Mileage shows moderate negative correlation with Price.

4.4 Price Relationships

Price vs YOM

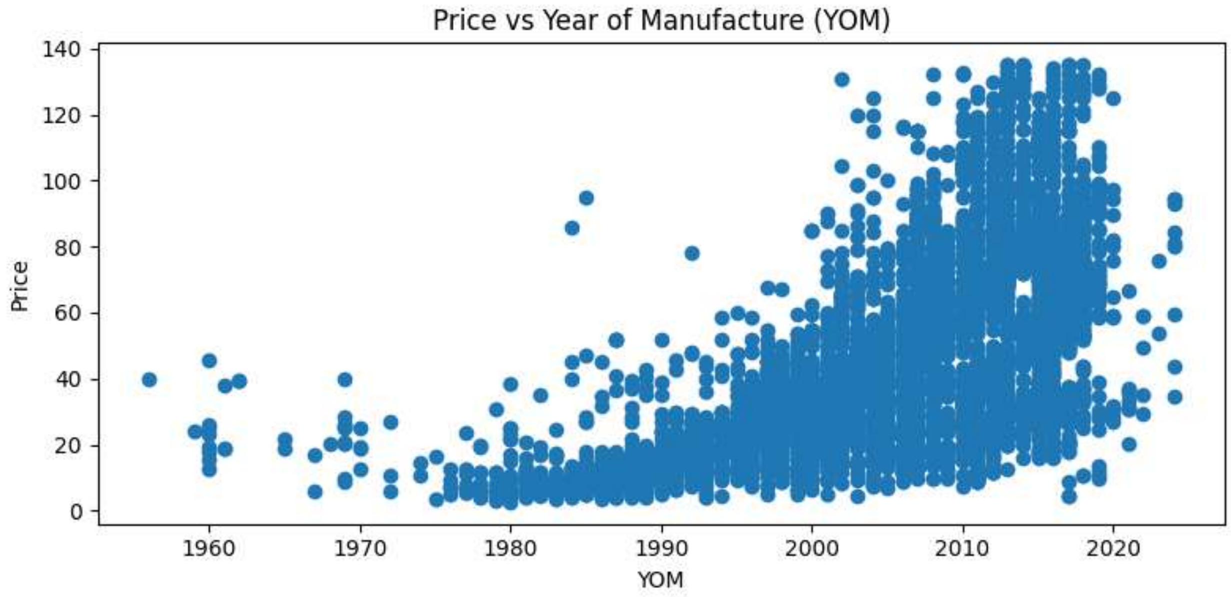

Figure 4:Price vs YOM Scatter Plot

Trend: Newer vehicles tend to have higher prices.

Price vs Mileage

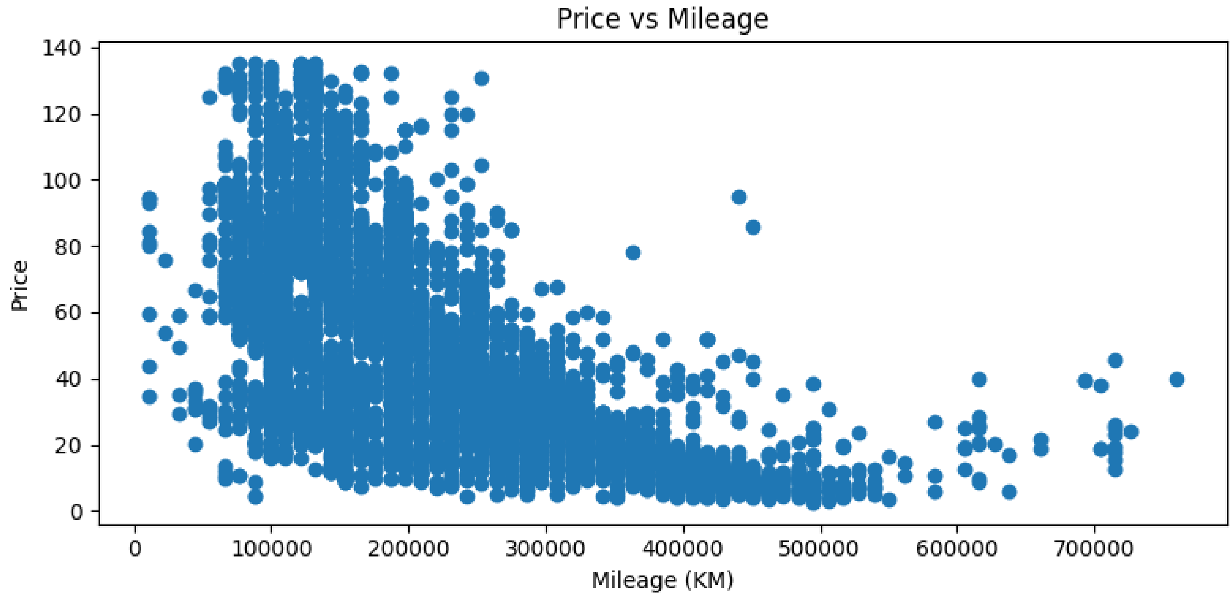

Figure 5:Price vs Mileage Scatter Plot

Trend: Higher mileage vehicles generally have lower prices.

5. Model Selection

The **XGBoost Regressor** was selected because:

- It effectively models non-linear relationships.
- It performs well on structured tabular data.
- It includes built-in regularization.
- It captures feature interactions automatically.

No deep learning models were used, as required by the assignment guidelines.

6. Experimental Setup

The dataset was split as:

- 70% Training
- 15% Validation
- 15% Testing

Evaluation metrics used:

- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- R^2 Score

7. Baseline Model Performance

A baseline XGBoost model was trained with:

- n_estimators = 300
- learning_rate = 0.1
- max_depth = 6

Test Results (Baseline XGBoost)

- RMSE: **7.7813**

- MAE: **4.6436**
- R²: **0.9268**

The baseline model demonstrates strong predictive capability.

7.1 Actual vs Predicted

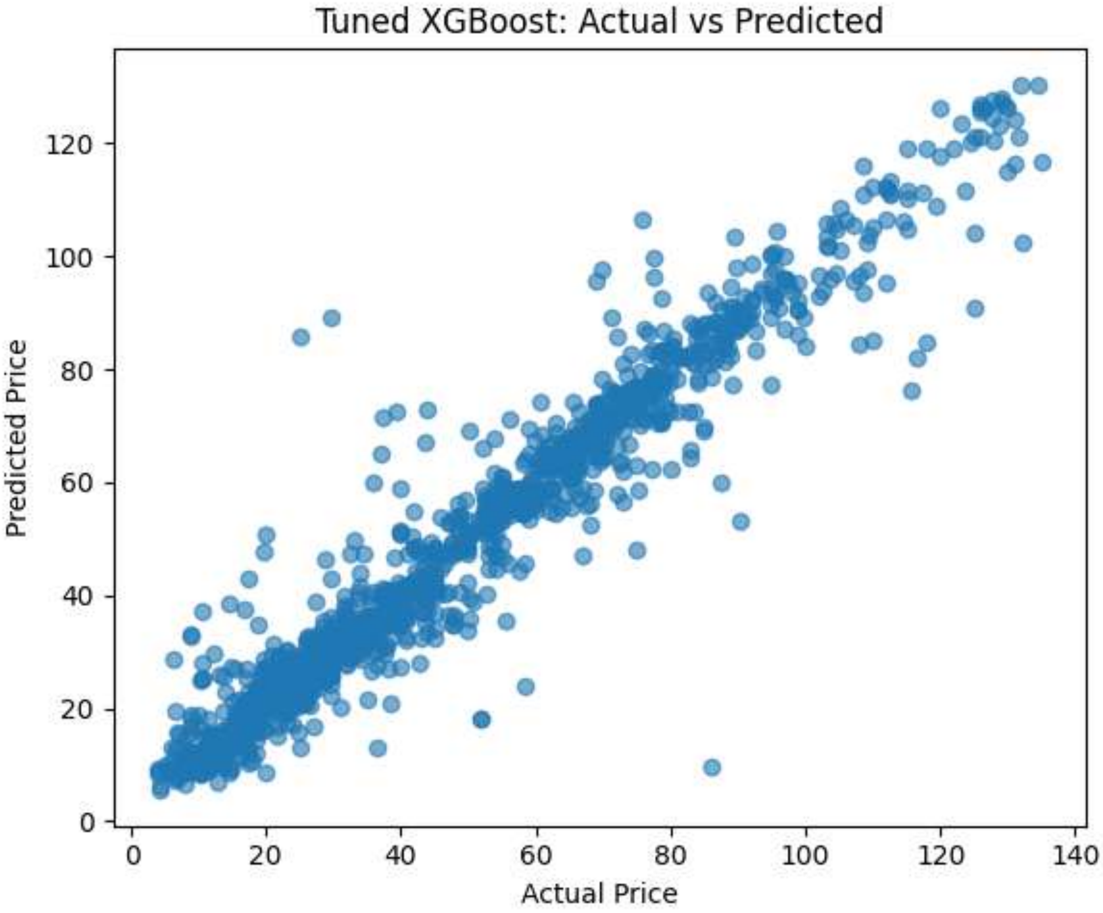

Figure 6:Actual vs Predicted Plot

Observation:

- Predictions cluster near the diagonal.
- Some dispersion exists at higher price values.

7.2 Residual Distribution

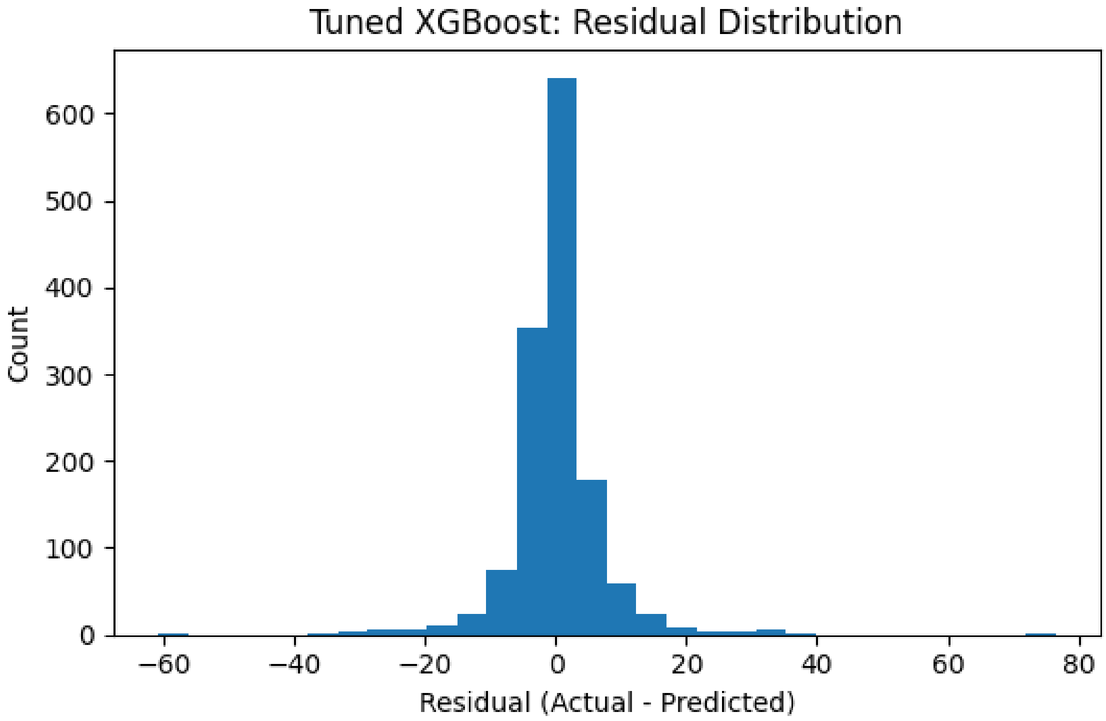

Figure 7:Residual Histogram]

Observation:

- Residuals are centered near zero.
- No extreme skew observed after outlier removal.

8. Hyperparameter Tuning

Hyperparameter tuning was performed using **RandomizedSearchCV** with 3-fold cross-validation.

Parameters tuned included:

- n_estimators
- learning_rate
- max_depth
- subsample

- colsample_bytree
- min_child_weight
- reg_alpha
- reg_lambda

9. Tuned Model Performance

Test Results (Tuned XGBoost)

- RMSE: **7.3387**
- MAE: **4.2128**
- R^2: **0.9349**

Improvement Over Baseline

Metric	Baseline	Tuned
RMSE	7.7813	7.3387
MAE	4.6436	4.2128
R^2	0.9268	0.9349

Table 1:Comparison table for base model and tune model

The tuned model shows:

- Reduced RMSE
- Reduced MAE
- Improved R^2

This confirms effective hyperparameter optimization.

10. Explainability (SHAP Analysis)

SHAP was applied to interpret model predictions.

10.1 Global Feature Importance

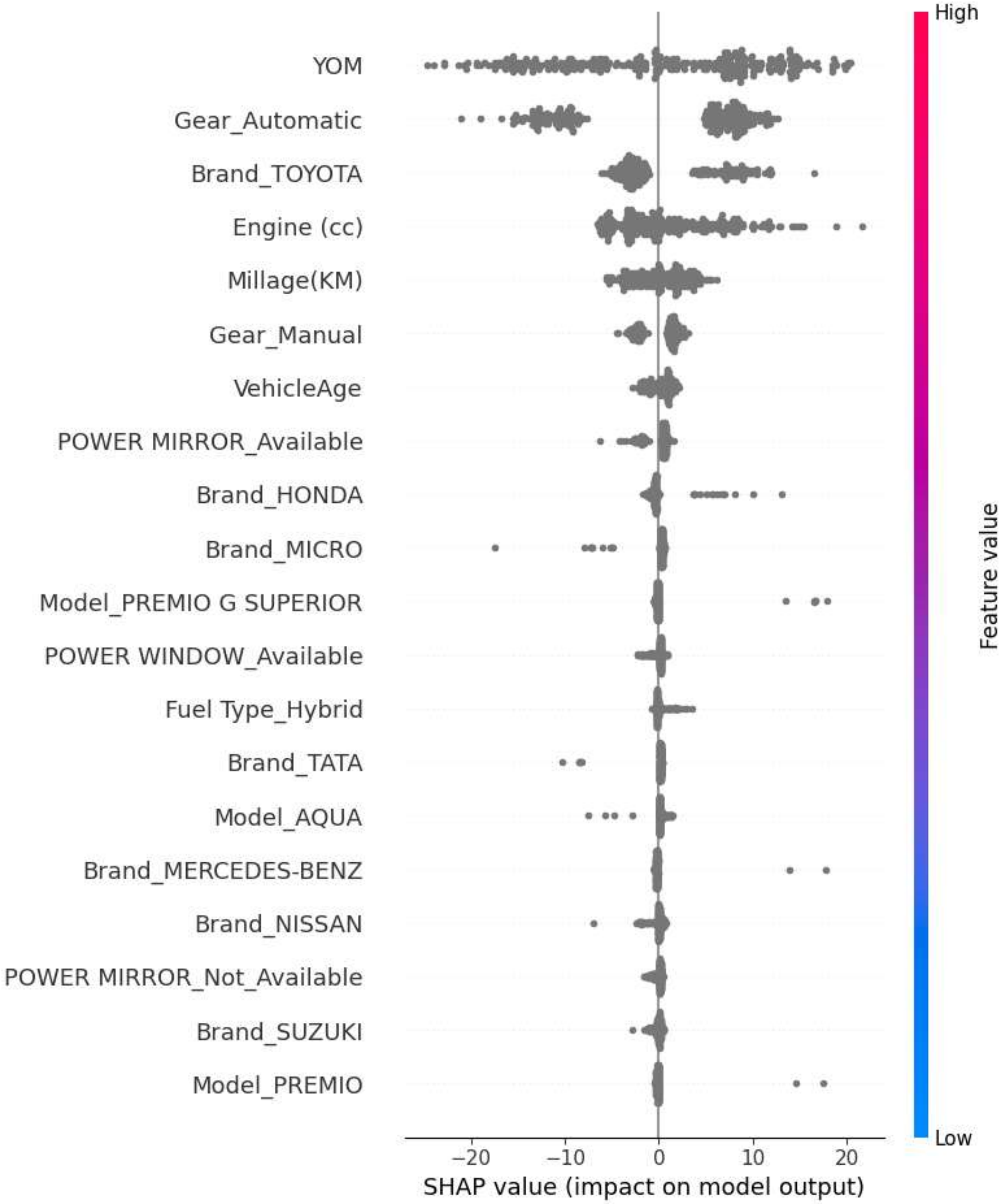

Figure 8: SHAP Summary Plot

Observation:

- VehicleAge strongly influences price.
- Brand significantly affects valuation.
- Mileage contributes negatively to price.
- Location (Town) impacts market valuation.

10.2 Individual Prediction Explanation

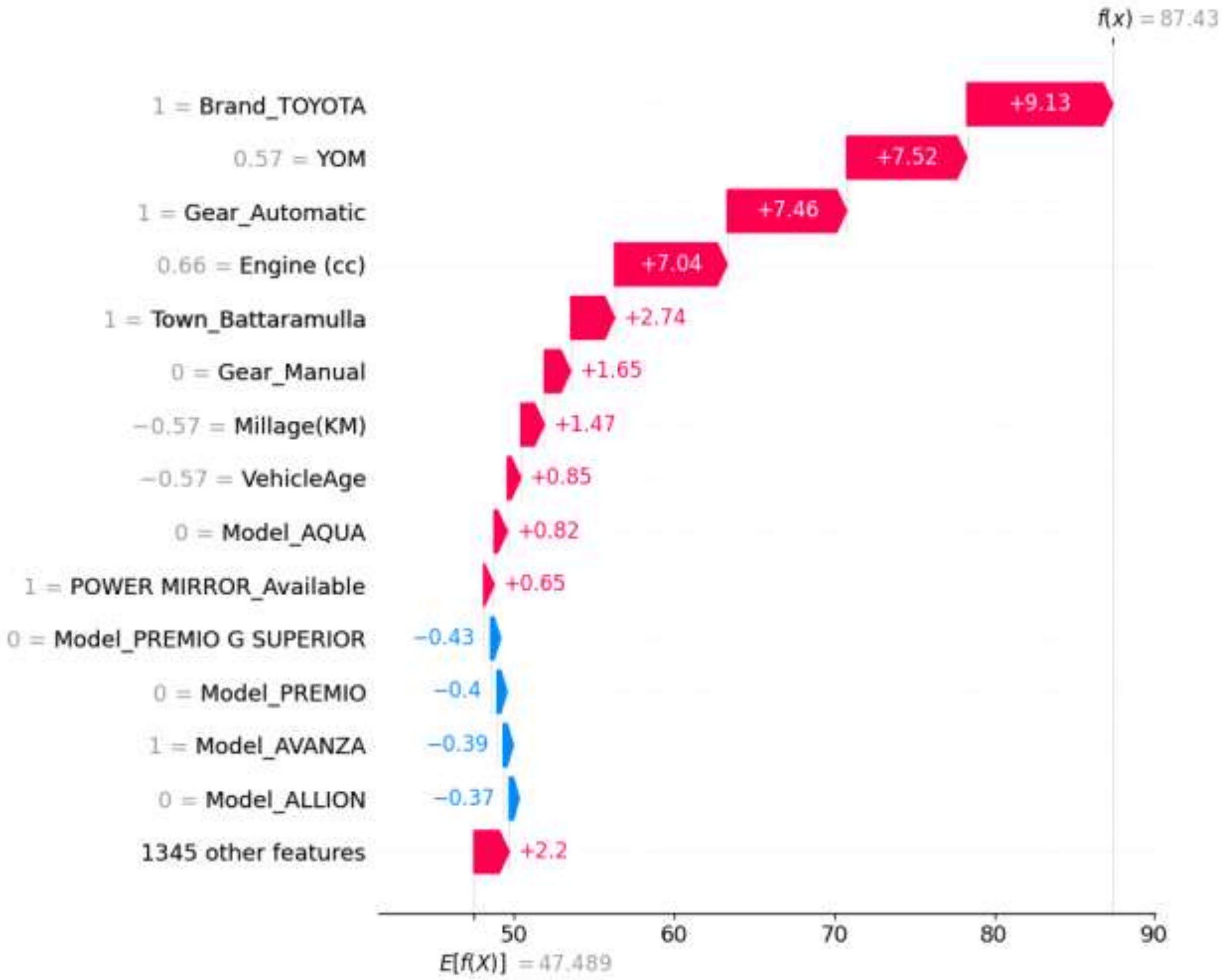

Figure 9:SHAP Waterfall Plot]

Interpretation:

- Features contributing positively to price are shown in red.

- Features reducing predicted value are shown in blue.
- The model explanation aligns with logical pricing behavior.

11. Front-End Integration

A Streamlit application was developed to:

- Accept vehicle details from users
- Generate real-time predictions
- Display SHAP explanation for each prediction

The trained pipeline was serialized using Pickle and loaded within the Streamlit application.

12. Docker Deployment

The application was containerized using Docker to ensure:

- Dependency isolation
- Environment consistency
- Portable deployment

The Docker configuration includes:

- Python 3.10
- XGBoost
- SHAP
- Streamlit

13. Critical Discussion

13.1 Limitations

- Dataset limited to available listings.
- No accident history or service record data.
- Market trends and economic factors not included.

- Possible inconsistencies in listing quality.

13.2 Data Quality Issues

- Outliers in price distribution (mitigated using IQR).
- Categorical inconsistencies across listings.
- Possible duplicate advertisements.

13.3 Bias Considerations

- Certain brands may dominate dataset.
- Urban towns may be overrepresented.
- Model predictions reflect historical listing patterns.

13.4 Real-World Impact

The system provides estimation support and should not replace professional valuation or market negotiation.

14. Conclusion

This project demonstrates that XGBoost regression combined with SHAP explainability can effectively model used vehicle prices in Sri Lanka. The tuned model achieved an R^2 score of 0.9349 with improved RMSE and MAE compared to the baseline model.

The integration of explainable AI and deployment through Streamlit and Docker makes the system both practical and interpretable.