AI-Based Sri Lankan Used Bike Price Prediction System

Adshaya Balarajah – 214024V

1. Introduction

The used motorcycle market in Sri Lanka is highly dynamic, with prices influenced by brand reputation, engine capacity, mileage, vehicle age, and geographic location. Due to inconsistent pricing practices and subjective valuation, buyers and sellers often face uncertainty.

This project aims to build a machine learning regression model capable of predicting used bike prices using structured listing data.

The task is formulated as a regression problem, where:

- Target Variable → Price

XGBoost regression was selected due to its ability to model non-linear relationships and interactions effectively.

2. Dataset Description

2.2 Data Source and Collection Context

The objective of this study is to predict used motorcycle prices based on their technical and contextual attributes. The dataset was collected by scraping listings from ikman.lk, a popular online marketplace in Sri Lanka.

The data represents used bike advertisements posted between January 2022 and March 2022, covering a wide range of motorcycle brands, models, engine capacities, usage levels, and geographical locations across Sri Lanka. Each record corresponds to a unique bike listing and includes seller-provided details relevant to price estimation.

2.3 Dataset Scale

The dataset consists of:

- Total records: 5,016 used bike listings

- Total features: 14 columns
- Geographical scope: Sri Lanka
- Time period: January 2022 – March 2022

2.3 Attributes of the dataset

The dataset (used-bikes.csv) contains used motorcycle listings with the following attributes:

- Brand
- Model
- Bike Type
- Trim/Edition
- Year
- Mileage
- Capacity (cc)
- District
- Price
- Other listing details

2.4 Data Cleaning

The following preprocessing steps were performed:

- Removed "Rs" and commas from Price
- Removed " km" from Mileage
- Removed " cc" from Capacity
- Converted all numeric columns to float
- Extracted District from post details
- Removed irrelevant columns:

- Summary
- URL
- Post_Details
- Seller
- Title

3. Feature Engineering

Several engineered features were introduced:

3.1 Vehicle Age

$$Vehicle\_Age = 2024 - Year$$

Equation 1:Vehicle Age calculation in Feature Engineering

Entries with unrealistic age (> 50 years) were removed.

3.2 Interaction Features

- Mileage_per_Year = Mileage / (Vehicle_Age + 1)

Equation 2:Mileage per year calculation in Interaction Feature

- CC_Age_Interaction = Capacity × Vehicle_Age

Equation 3:CC Age Interaction

- CC_Mileage_Interaction = Capacity × Mileage

Equation 4:CC Mileage Interaction

These features capture:

- Usage intensity
- Engine depreciation effects
- Capacity-related wear patterns

4. Target Transformation

The price distribution was right-skewed.

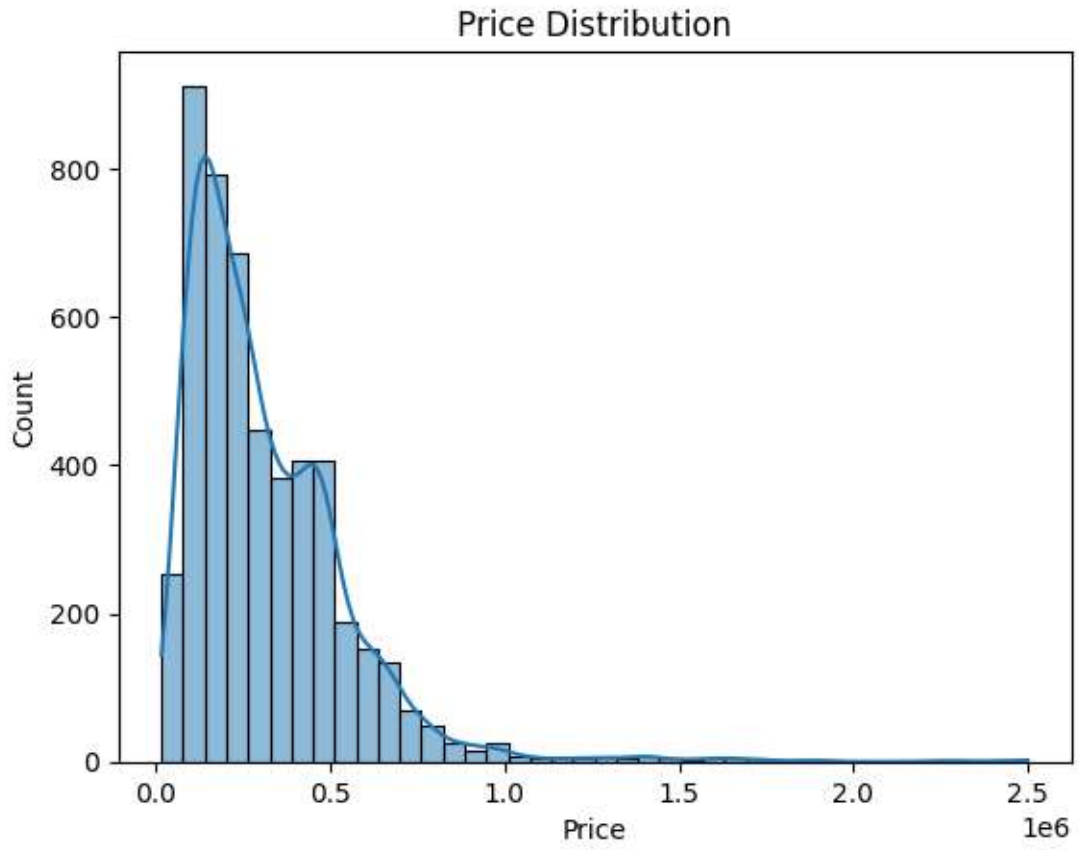

Figure 1:Price Distribution Plot

To stabilize variance and improve regression performance, a logarithmic transformation was applied:

$$Log\_Price = log(1 + Price)$$

Equation 5:logarithmic transformation

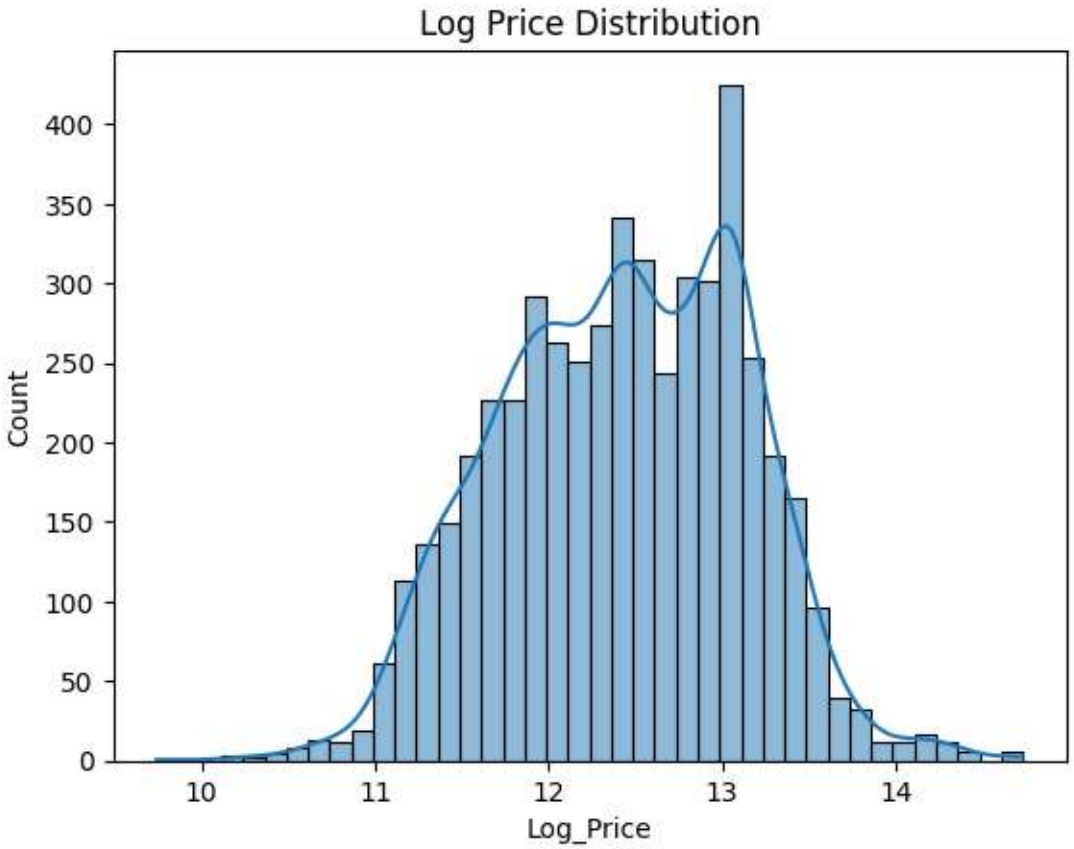

Figure 2:Log Price Distribution Plot

This improves model learning by reducing the impact of extreme values.

5. Exploratory Data Analysis

5.1 Correlation Analysis

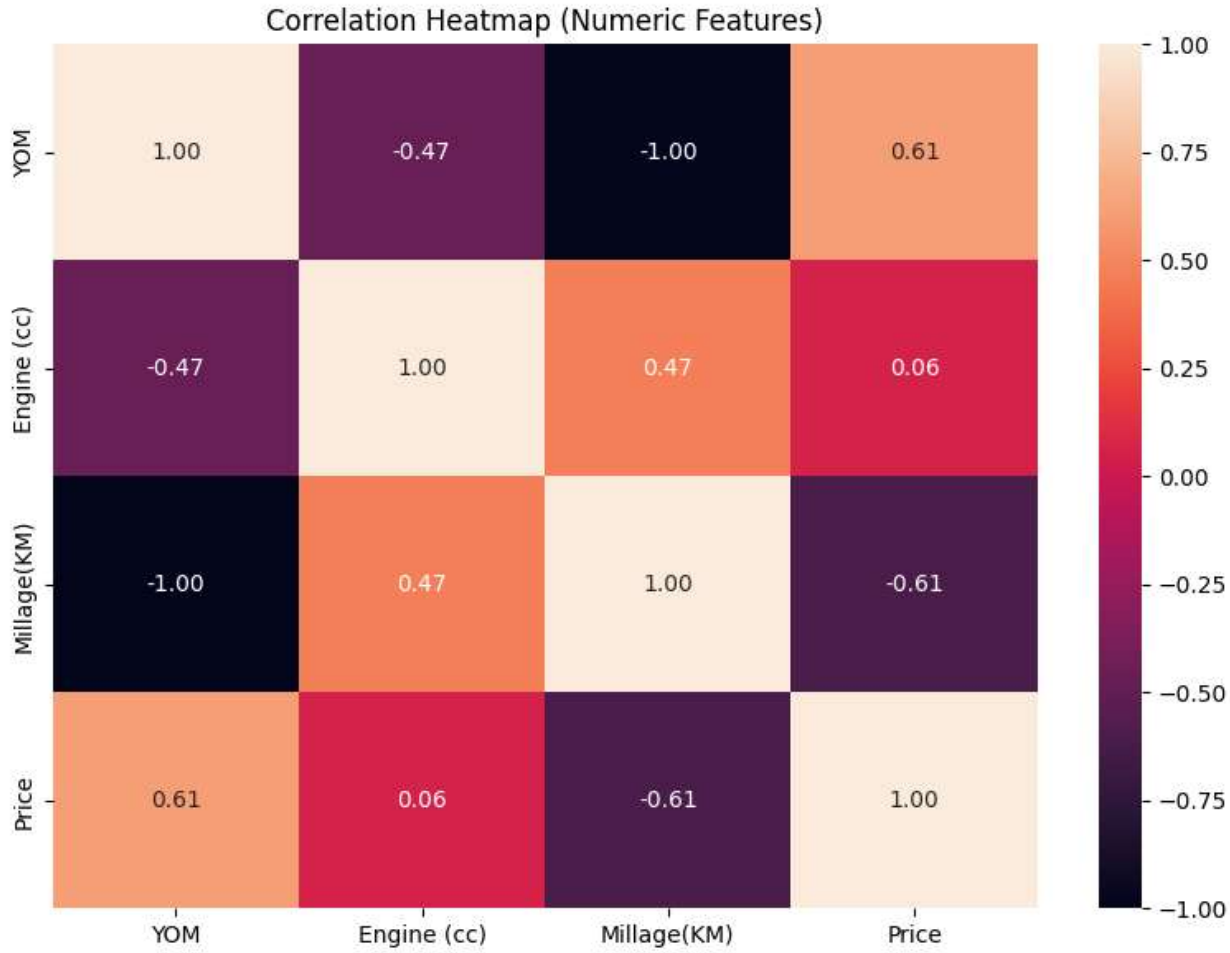

Figure 3:Correlation Heatmap

Observations:

- Vehicle age negatively correlates with price.
- Capacity shows moderate positive correlation.
- Mileage negatively impacts valuation.

6. Model Development

6.1 Data Split

- 80% Training

- 20% Testing

Random split with fixed random state for reproducibility.

7. Baseline Model

An XGBoost Regressor was trained with:

- n_estimators = 500
- learning_rate = 0.05
- max_depth = 6
- subsample = 0.8
- colsample_bytree = 0.8

Baseline Performance

- RMSE: 133,939 LKR
- MAE: 92,606 LKR
- R^2: 0.5781

The model explains approximately 57.8% of the variance in bike prices.

8. Hyperparameter Tuning

Hyperparameter optimization was performed using:

- RandomizedSearchCV
- 3-fold Cross Validation
- R^2 as scoring metric

Parameters tuned:

- n_estimators
- max_depth
- learning_rate

- subsample
- colsample_bytree
- min_child_weight
- gamma

9. Tuned Model Performance

Final Tuned Results

- RMSE: 132,760 LKR
- MAE: 91,901 LKR
- R^2: 0.5855

Performance Comparison

Metric	**Baseline**	**Tuned**
RMSE	133,939	132,760
MAE	92,606	91,901
R^2	0.5781	0.5855

Table 1:Comparison table for base model and tune mode

The tuned model shows slight but consistent improvement.

10. Model Interpretation (SHAP Analysis)

SHAP was applied to interpret feature contributions.

10.1 Global Feature Importance

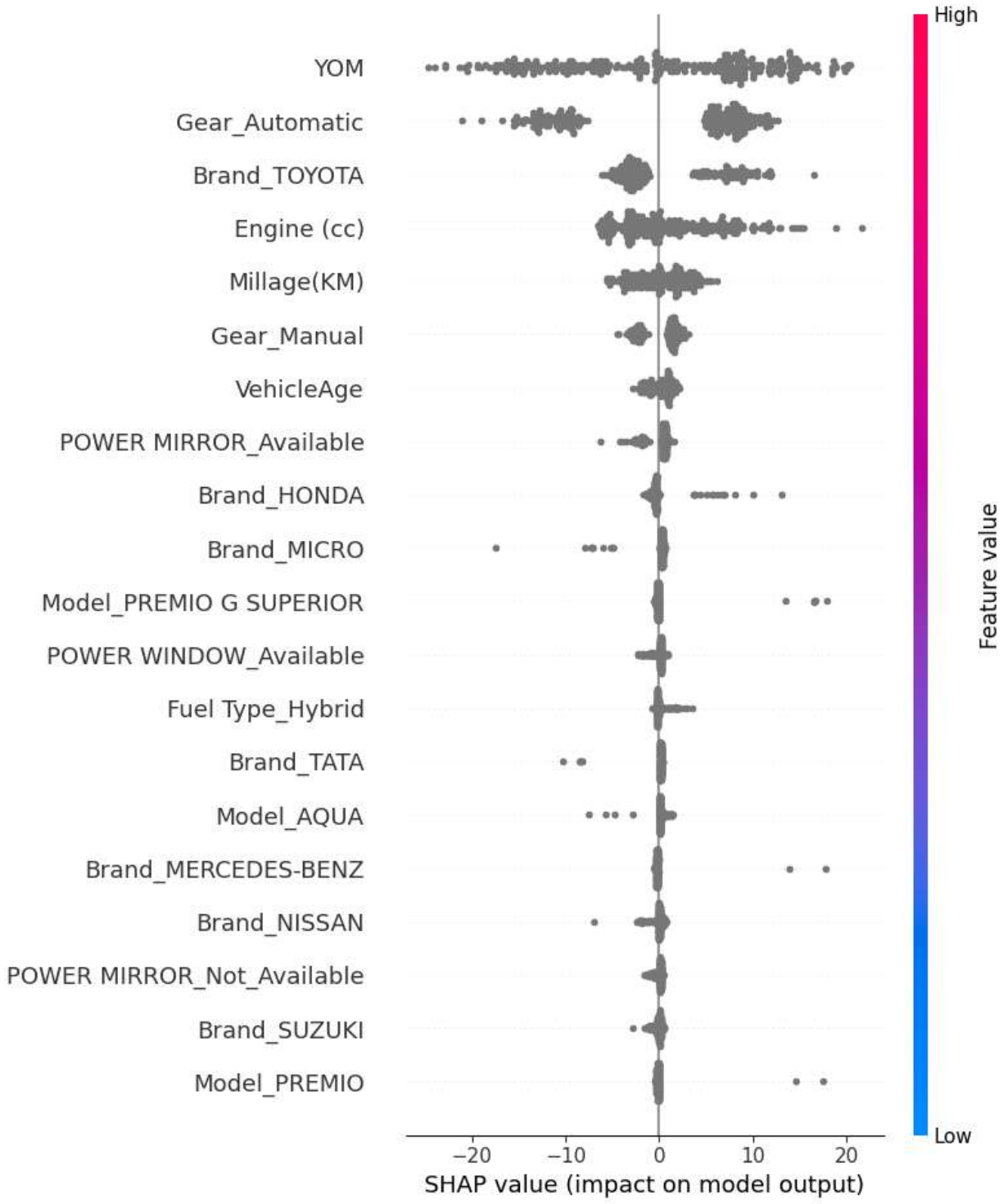

Figure 4:Summary Plot

Top influential features include:

- Vehicle_Age
- Capacity
- Mileage_per_Year
- District
- Brand

These align with expected depreciation and engine power effects.

10.2 Feature Dependence

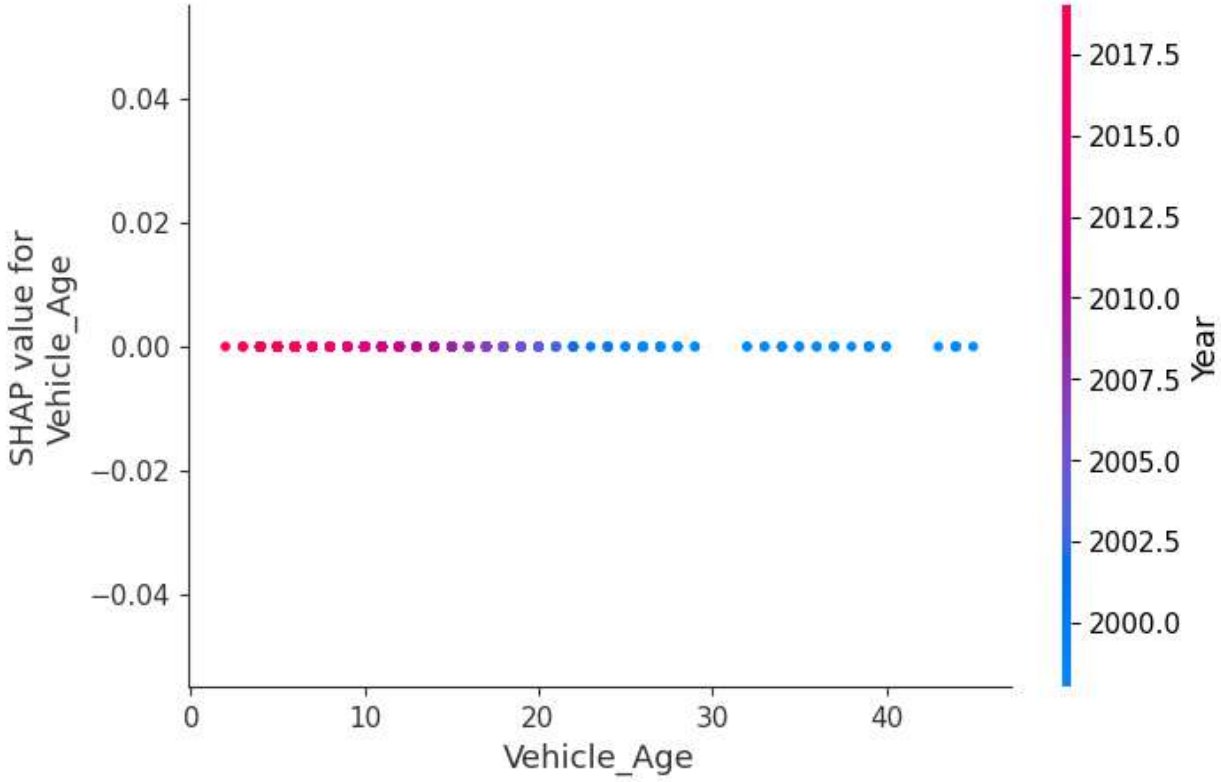

Figure 5:SHAP Dependence Plot for Vehicle_Age

Interpretation:

- Older bikes significantly reduce predicted price.
- Non-linear depreciation patterns are captured.

10.3 Individual Prediction Explanation

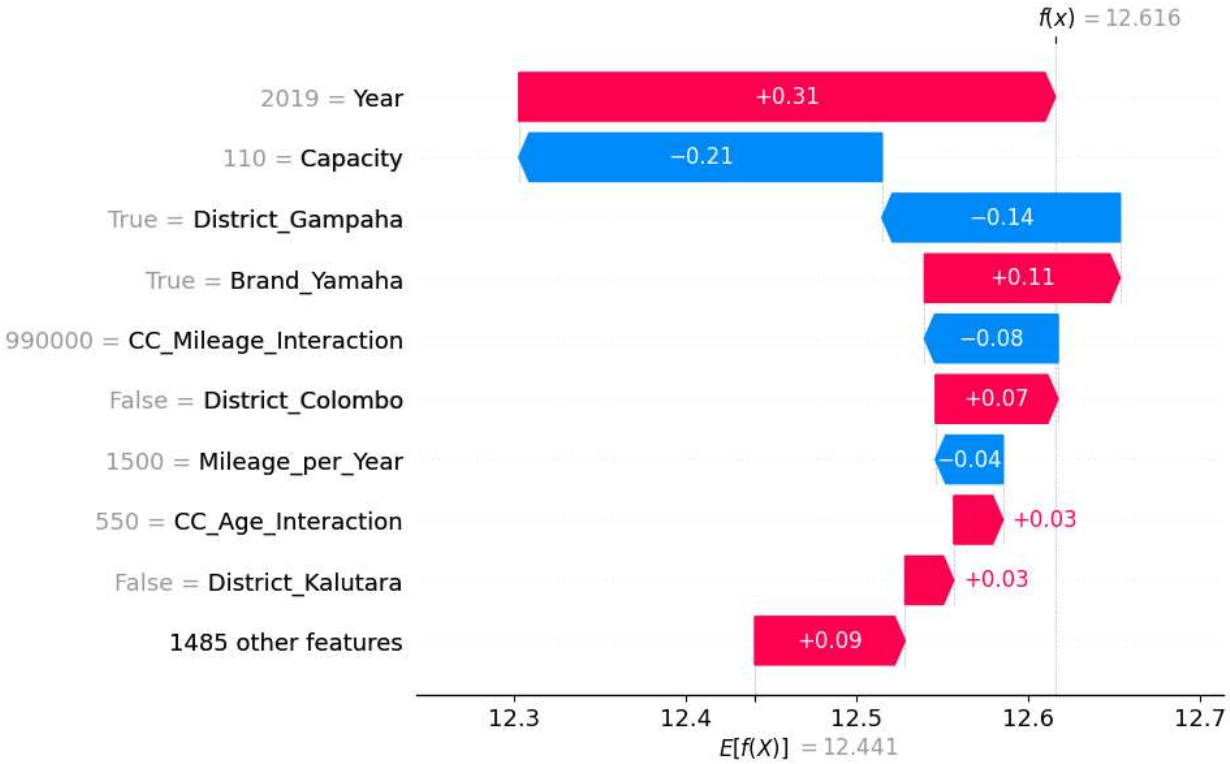

Figure 6: SHAP Waterfall Plot

This visualization explains why a specific bike received its predicted valuation.

11. Model Export

The trained model was exported using:

- bike_price_xgb.json
- model_columns.pkl
- shap_background.pkl

This allows:

- Deployment
- SHAP-based explanation in production
- Reproducible predictions

12. Critical Discussion

12.1 Limitations

- Moderate predictive power ($R^2 \approx 0.58$)
- No service history or accident data
- No market demand indicators
- Random split may not reflect temporal market trends

12.2 Data Quality Issues

- Listing inconsistencies
- Possible duplicate advertisements
- Manual price entries may contain noise

12.3 Practical Use

The model provides a reference price estimation tool, but should not replace professional inspection or negotiation.

13. Conclusion

This project demonstrates that XGBoost regression combined with log transformation and interaction feature engineering can effectively model used bike prices in Sri Lanka.

The tuned model achieved:

- R^2: 0.5855
- RMSE: 132,760 LKR

While the predictive strength is moderate, the model successfully captures depreciation patterns and engine-related price effects.

The integration of SHAP ensures transparency and interpretability in pricing decisions.