MIT Vishwaprayag University, Solapur

**School of Computing**
**B.Tech (Computer Science and Engineering)**

PROJECT REPORT

# Used Motorcycle Price Prediction

**A Machine Learning Project**

Sohan Vasekar

**SCFU123016**

## Abstract

This project presents a machine learning-based system for predicting the resale price of used motorcycles. A real-world dataset containing motorcycle listings was analyzed to identify factors influencing resale value. The workflow includes exploratory data analysis, preprocessing, outlier handling, feature encoding, model development, and performance evaluation. Multiple regression models were implemented and compared, among which Random Forest regression achieved the best performance. The final model was deployed using a Streamlit web application to provide interactive price predictions.

# Contents

# 1 Introduction

The resale price of a motorcycle depends on factors such as manufacturing year, usage, ownership history, seller type, and original showroom price. Without structured analysis, resale pricing often becomes subjective and inconsistent.

Machine learning enables data-driven price estimation by learning patterns from historical resale data. This project develops a regression-based system to estimate motorcycle resale prices using real-world data.

# 2 Problem Statement

The objective of this project is to build a machine learning system that accurately predicts the resale price of a used motorcycle. The system should:

- Identify key features influencing resale value
- Handle missing values and outliers
- Compare multiple regression models
- Provide predictions through a deployable interface

# 3 Dataset Description

The dataset used in this project (`BIKEDETAILS.csv`) contains information about used motorcycles listed for resale. It consists of 1061 records with the following attributes:

- **name**: Motorcycle model name
- **year**: Manufacturing year
- **seller_type**: Individual or dealer
- **owner**: Number of previous owners
- **km_driven**: Distance travelled
- **ex_showroom_price**: Original showroom price
- **selling_price**: Resale price (target variable)

### Dataset Source

The dataset was obtained from Kaggle.

# 4 Exploratory Data Analysis

EDA was performed to understand dataset characteristics and identify data quality issues. Key activities included:

- Checking data types, missing values, and duplicates
- Analyzing numerical feature distributions
- Detecting outliers using distribution plots and the IQR method
- Examining correlations after encoding
- Identifying relationships such as selling price with manufacturing year, kilometers driven, and ex-showroom price

## 4.1 Missing Value Handling

Missing values were observed only in the `ex_showroom_price` feature. Since this feature exhibited skewness, missing values were imputed using the median to avoid the influence of extreme values.

$$\text{Imputed Value} = \text{Median}(\text{ex\_showroom\_price})$$

## 4.2 Outlier Analysis

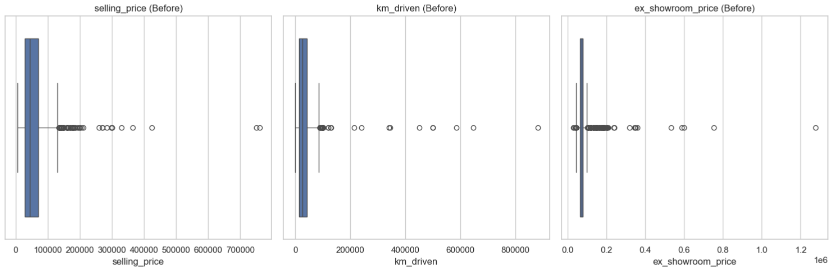

Figure 1: Outliers Before Treatment

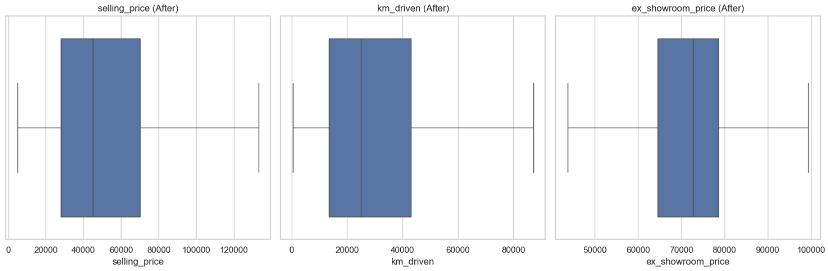

Figure 2: Outliers After Treatment

Outliers in numerical features were identified using the Interquartile Range (IQR) method:

$$IQR = Q3 - Q1$$

$$\text{Lower Bound} = Q1 - 1.5 \times IQR, \quad \text{Upper Bound} = Q3 + 1.5 \times IQR$$

Extreme values were capped to stabilize model behavior and reduce distortion during training.

### 4.3 Distribution Analysis Before and After Outlier Handling

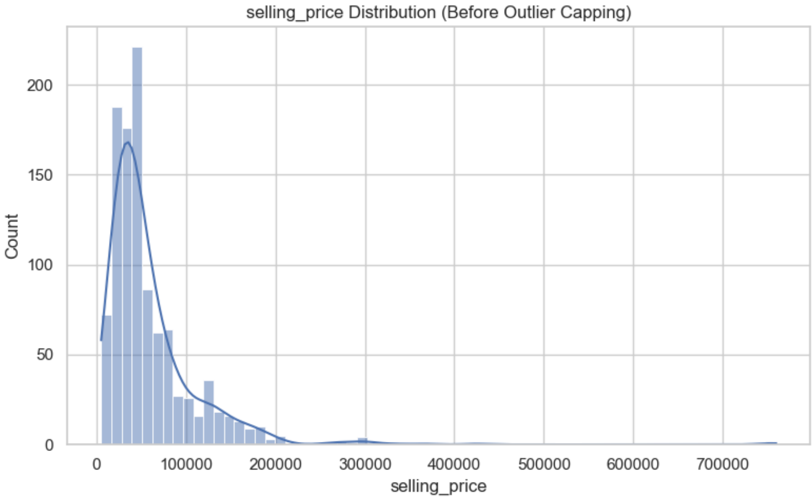

Figure 3: Selling Price Distribution Before Outlier Handling

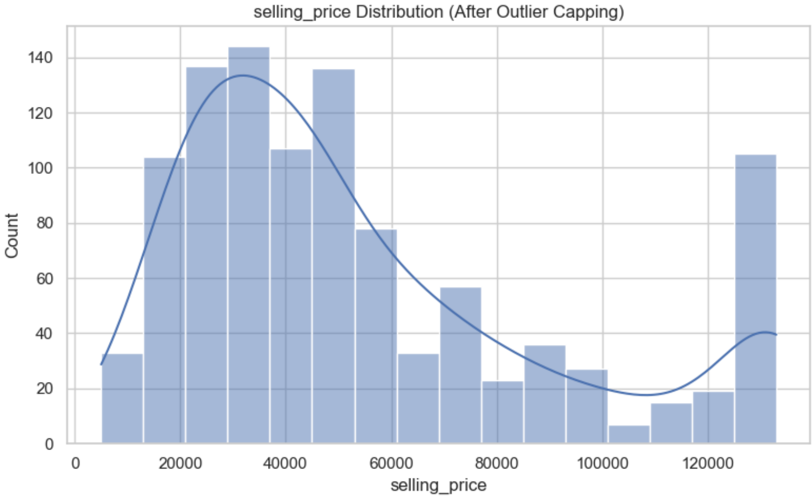

Figure 4: Selling Price Distribution After Outlier Handling

**Explanation:** Before preprocessing, the selling price distribution exhibits strong right skew due to a small number of high-priced motorcycles. These extreme values can disproportionately influence regression models. After outlier handling, the distribution becomes more compact while preserving overall pricing trends.

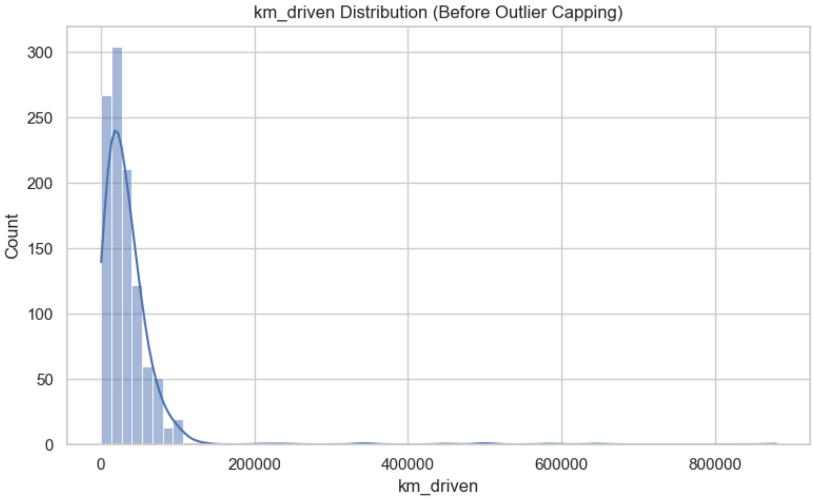

Figure 5: Kilometers Driven Distribution Before Outlier Handling

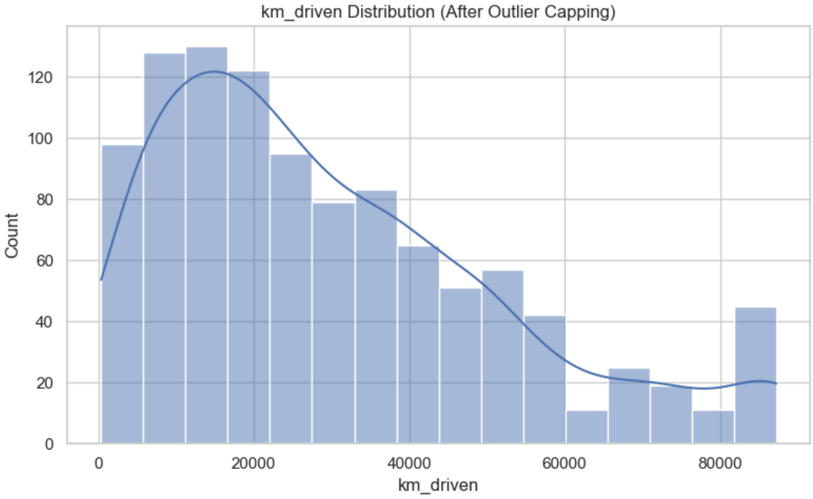

Figure 6: Kilometers Driven Distribution After Outlier Handling

**Explanation:** The kilometers driven feature initially contains extreme usage values. After preprocessing, the distribution reflects more realistic riding patterns, reducing noise during model training.

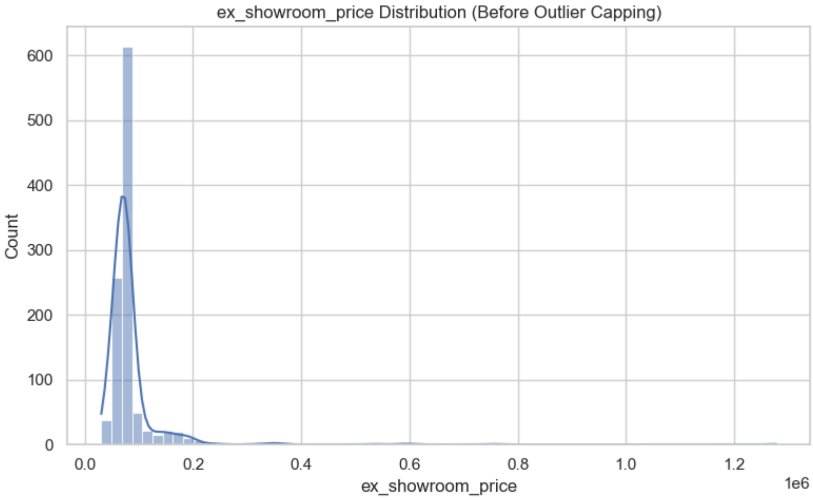

Figure 7: Ex-Showroom Price Distribution Before Outlier Handling

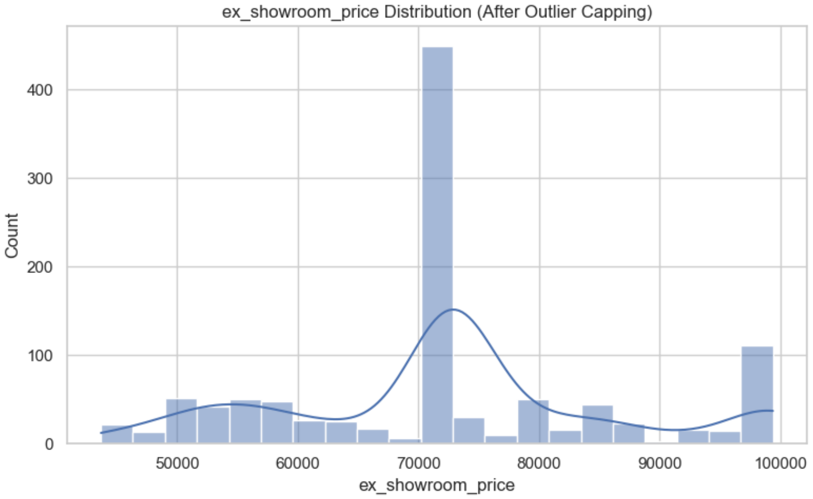

Figure 8: Ex-Showroom Price Distribution After Outlier Handling

**Explanation:** The ex-showroom price feature shows a highly skewed distribution before preprocessing. After outlier handling, the distribution becomes more stable, improving its suitability for regression modeling.

## 4.4 Correlation Analysis

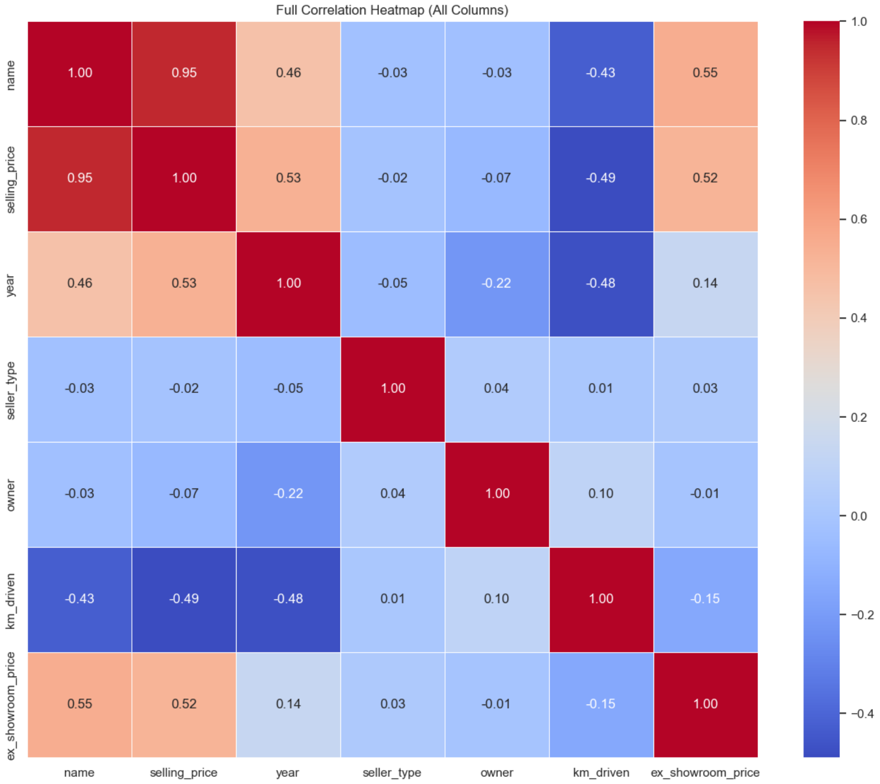

Figure 9: Correlation Heatmap

**Explanation:** Strong positive correlations are observed between selling price, ex-showroom price, and manufacturing year, confirming their predictive importance.

## 4.5 Feature-wise Price Insights

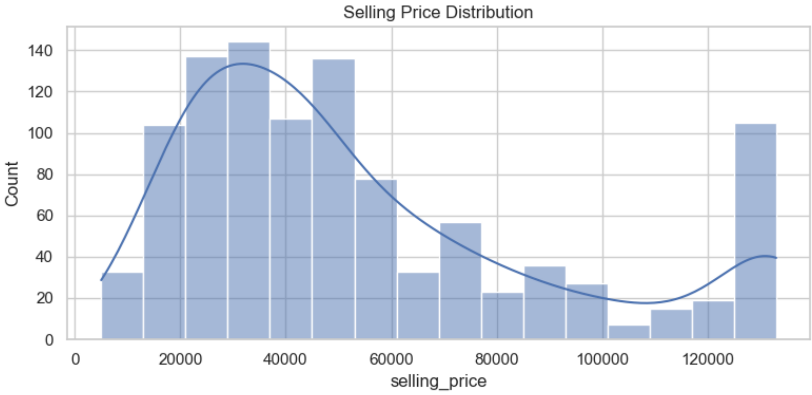

Figure 10: Distribution of Selling Price

**Explanation:** The selling price distribution is right-skewed, indicating that most motorcycles fall into lower to mid-price ranges, while a smaller number of premium motorcycles are priced significantly higher.

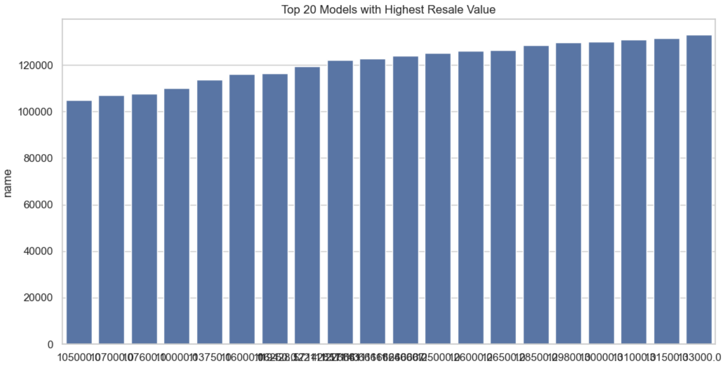

Figure 11: Brand-wise Selling Price Comparison

**Explanation:** Certain brands consistently command higher resale prices due to factors such as brand reputation, engine reliability, and market demand.

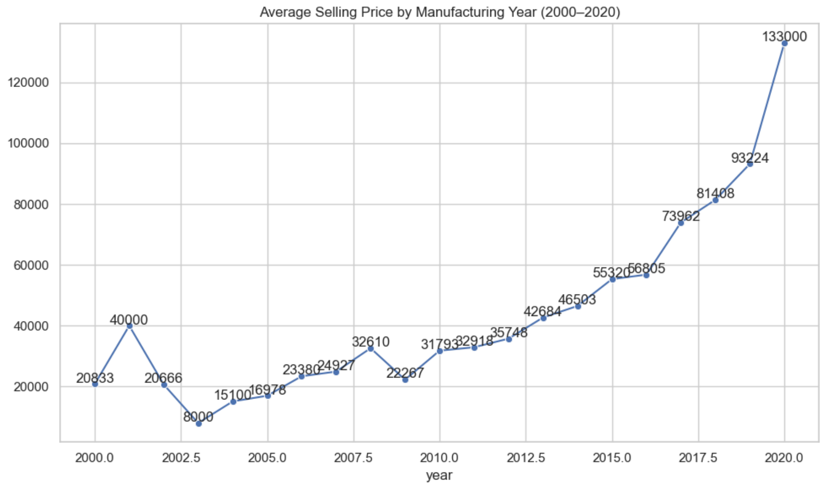

Figure 12: Manufacturing Year vs Selling Price

**Explanation:** Newer motorcycles generally retain higher resale value, while older models experience price depreciation over time.

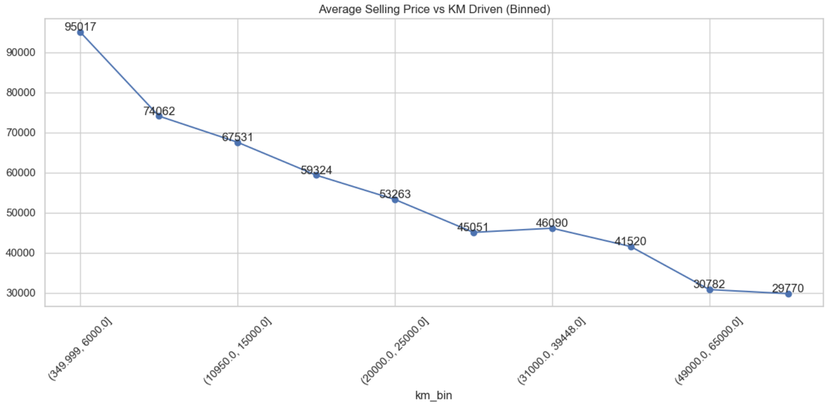

Figure 13: Kilometers Driven vs Selling Price

**Explanation:** As kilometers driven increase, the resale price tends to decrease, reflecting wear and usage impact on vehicle value.

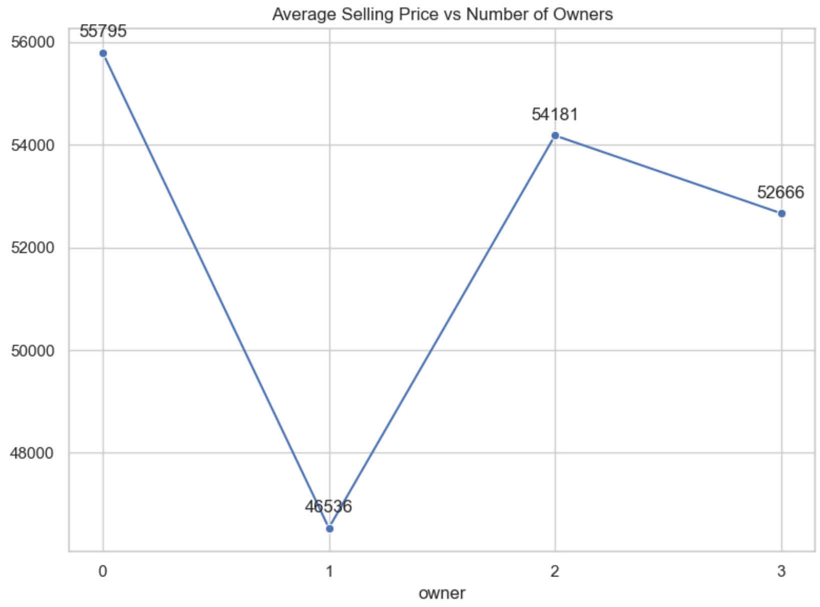

Figure 14: Number of Owners vs Selling Price

**Explanation:** Motorcycles with fewer previous owners generally have higher resale value, as multiple ownership often indicates increased usage or maintenance uncertainty.

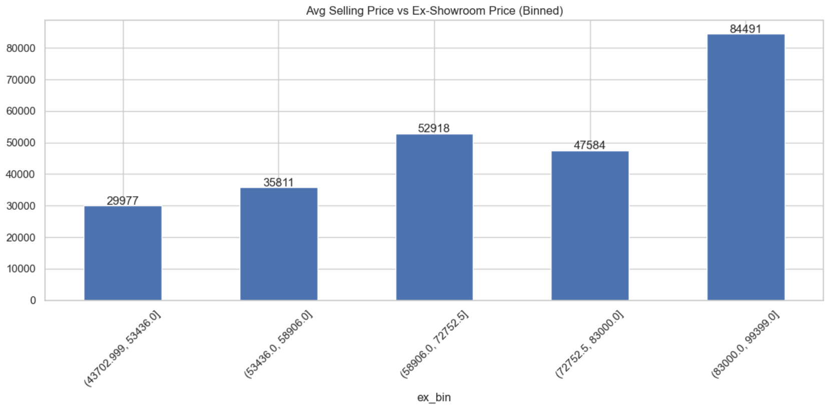

Figure 15: Ex-Showroom Price vs Resale Price

**Explanation:** A strong positive relationship exists between ex-showroom price and resale price, confirming that the original value of a motorcycle significantly influences its resale value.

# 5 Model Development

## 5.1 K-Nearest Neighbors Regression

KNN regression predicts price by averaging the prices of the nearest motorcycles in feature space. It assumes that similar motorcycles have similar resale values.

## 5.2 Decision Tree Regression

Decision Tree regression partitions the dataset using feature-based rules. It captures non-linear relationships but may overfit without depth constraints.

## 5.3 Random Forest Regression

Random Forest combines multiple decision trees trained on random subsets of data and features. This ensemble approach reduces overfitting and improves generalization, making it effective for complex pricing patterns.

## 5.4 Gradient Boosting Regression

Gradient Boosting builds trees sequentially, where each tree corrects the errors of the previous one. It emphasizes difficult samples and often achieves strong predictive performance.

# 6 Model Evaluation

Model performance was evaluated using regression-specific diagnostic plots and the coefficient of determination ($R^2$).

## 6.1 $R^2$ Score

| **Model** | $R^2$ **Score (%)** |
|---|---|
| KNN | 92.40 |
| Decision Tree | 91.84 |
| Random Forest | 93.05 |
| Gradient Boosting | 91.83 |

### 6.2 Actual vs Predicted Analysis

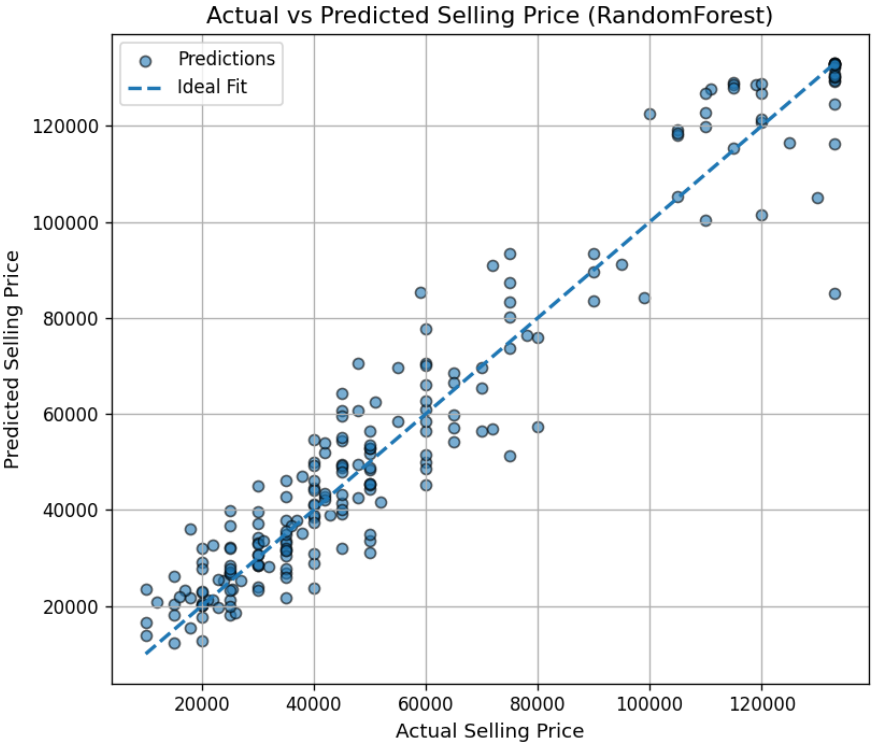

Figure 16: Actual vs Predicted Selling Price

**Explanation:** Predictions closely align with actual values, indicating strong overall model accuracy.

### 6.3 Residual Error Analysis

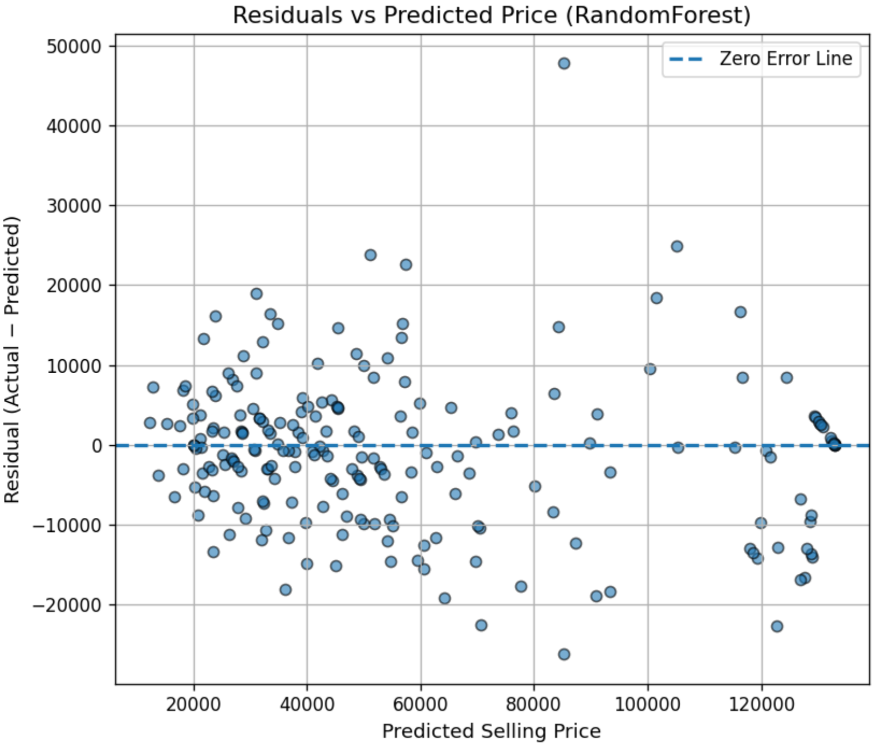

Figure 17: Residuals vs Predicted Price

**Explanation:** Residuals are centered around zero, showing low bias. Error variance increases at higher prices, indicating reduced accuracy for premium motorcycles.

## 7 Project Repository

The complete source code, dataset preprocessing scripts, trained models, evaluation plots, and deployment files for this project are available on GitHub at the following link: `github.com/vasekarsohan/Used-Bike-Price-Prediction`

## 8 Conclusion

This project presents a complete machine learning pipeline for used motorcycle price prediction. Detailed EDA, robust preprocessing, and model evaluation demonstrate that Random Forest regression provides reliable performance for this task.