Used Bike Price Prediction Project Report

Abstract

This project focuses on predicting the resale price of used bikes using machine learning techniques. The dataset was cleaned, explored, and transformed before building predictive models. Linear Regression was used as a baseline, followed by a Random Forest Regressor as an advanced model, which achieved better accuracy.

1. Introduction

The price of a used bike depends on various factors such as brand, engine capacity, ownership history, and usage. Manual estimation is unreliable, so this project aims to automate price prediction using data-driven approaches.

2. Dataset Description

The dataset contains information about used bikes including brand, model, kilometers driven, ownership, engine CC, and price. Raw data contained missing values, duplicates, and categorical variables that required preprocessing.

3. Data Preprocessing

Steps applied: - Removed null and duplicate values - Cleaned CC and kilometer fields - Converted categorical columns into numerical form - Encoded ownership levels - Created CC buckets

4. Exploratory Data Analysis (EDA)

Visualizations were created to understand: - Price distribution - Effect of engine capacity - Ownership vs resale value - Mileage influence

Log transformation of price was applied to handle skewed data.

5. Feature Engineering

New features created: - Owner numeric encoding - CC category buckets - Log-transformed price

6. Model Building

Two models were implemented:

Linear Regression

Used as baseline model.

Random Forest Regressor

Used as final model to improve accuracy and capture non-linear patterns.

7. Model Evaluation

Metrics used: - MAE - RMSE - R^2 Score

Results: - Linear Regression $R^2 \approx 0.52$ - Random Forest $R^2 \approx 0.66$

8. Conclusion

The Random Forest model significantly improved performance over Linear Regression. An R^2 score of 0.66 indicates strong predictive ability despite missing real-world factors like vehicle condition.

9. Future Work

- Use XGBoost or CatBoost
- Add condition-based features
- Region-based price modeling

10. Tools Used

- Google Colab
- Python
- Pandas, NumPy
- Matplotlib, Seaborn
- Scikit-learn