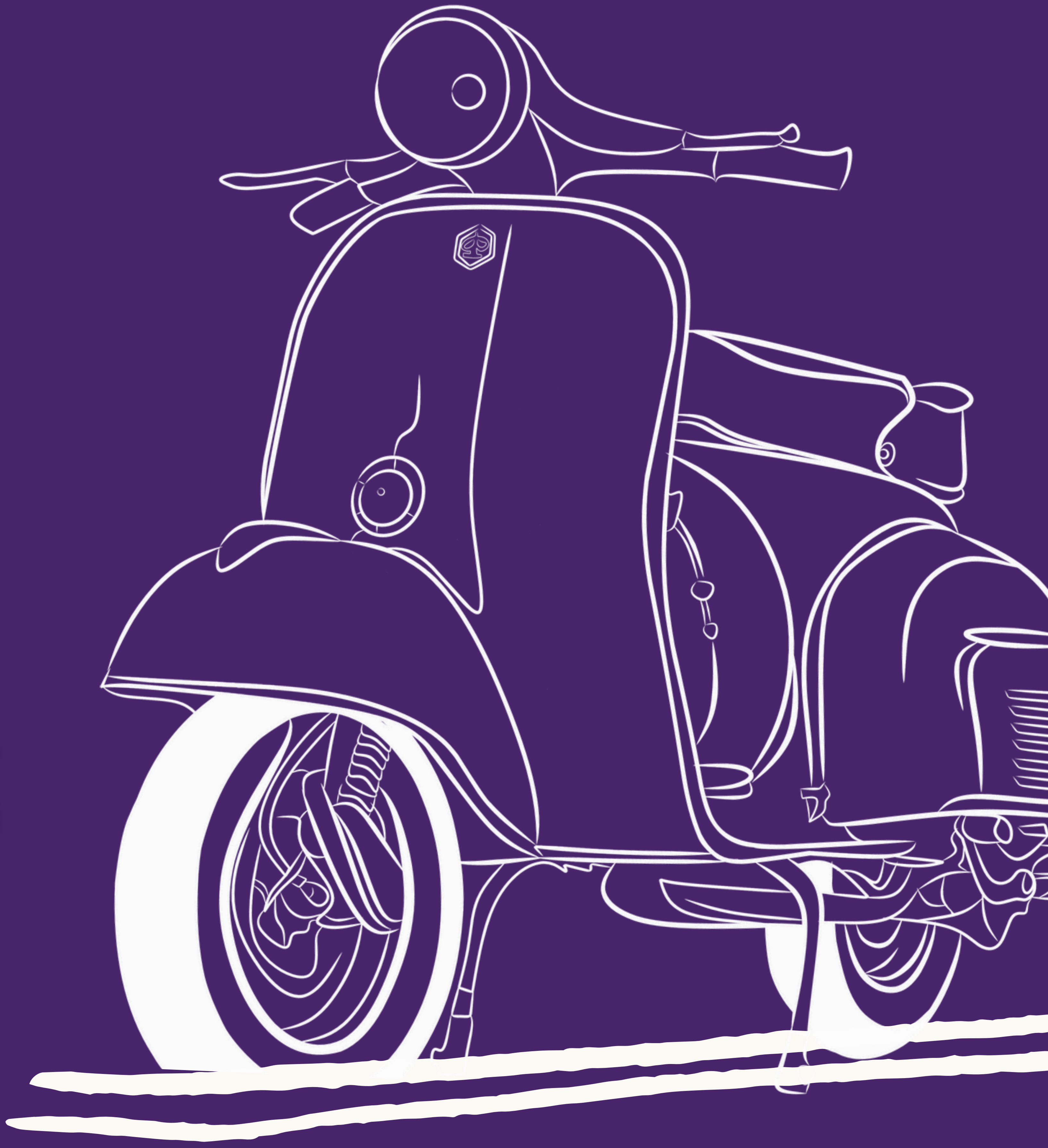
UNIFIED MENTOR

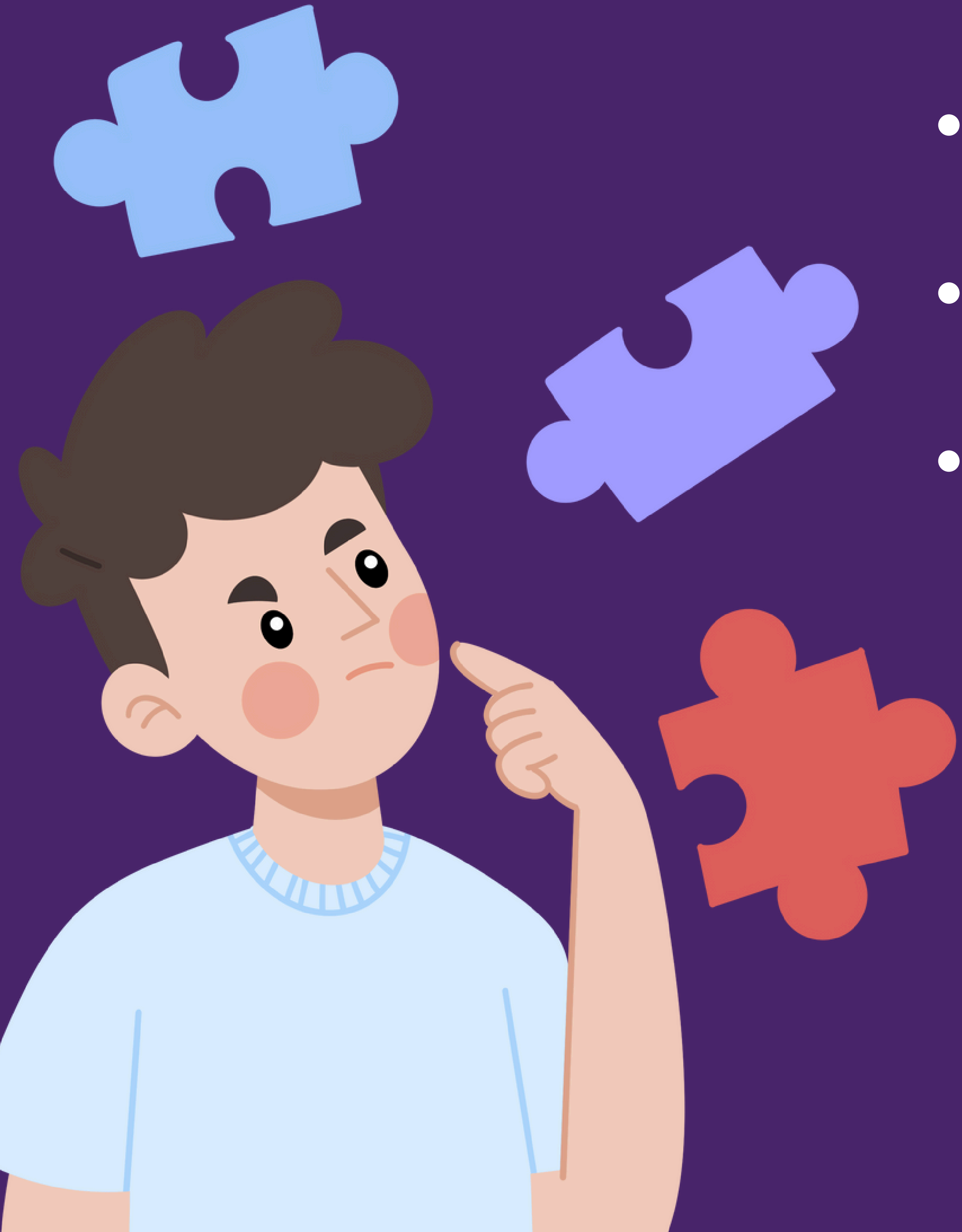
DATA SCIENCE INTERNSHIP PROJECT 1



USED BIKE PRICE

FEATURE ENGINEERING, EDA & MACHINE LEARNING

"IMAGINE YOU WANT TO SELL YOUR BIKE OR YOU WANNA BUY. HOW DO YOU KNOW THE RIGHT PRICE?"



- The used bike market in India is huge but pricing is inconsistent.
- Buyers risk overpaying, sellers risk undervaluing.
- Online platforms lack transparent, data-driven pricing tools.

A Challenge

Unstructured Pricing

No standard method; prices depend on seller expectations rather than actual bike condition.

Lack of Transparency

Buyers and sellers have information asymmetry, leading to mistrust.

Market Inefficiency

Platforms like OLX/Droom show wide price gaps for similar bikes, making deals slower and frustrating.

Multiple Influencing Factors

Year, mileage, power, location, and ownership history interact in complex ways → hard to estimate manually.

Dataset Overview

01 Source: Used Bikes Dataset (CSV)

02 Rows: 7858 | Columns: 7+ features

03 Features: model_year, kms_driven, mileage, power, location, owner, price

Data Preprocessing

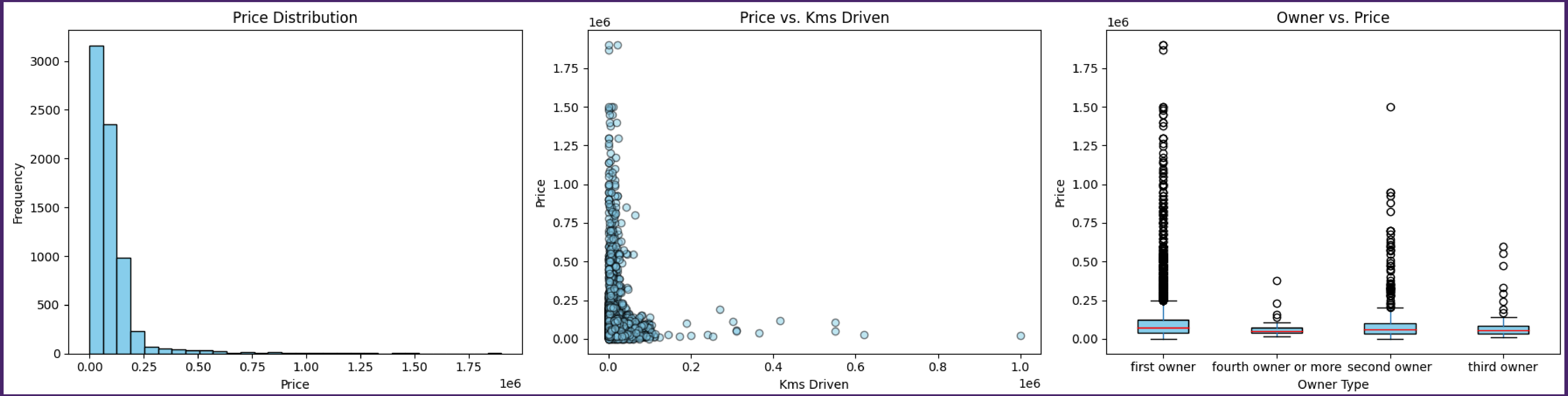
- Removed missing & inconsistent values.
- Cleaned text units: “20,000 Km” → 20000.
- Outliers removed with IQR method.
- Standardized formats for mileage, power.

EDA Insights

Price Distribution: Skewed, most bikes under ₹1 lakh.

Scatter: Price vs. Kms Driven → Negative correlation.

Conclusion: Higher kms = Lower resale value n.



EDA Insights visualization

- Boxplot: Owner Type vs. Price → First owner bikes fetch higher resale value.
- Location impact: Prices vary across cities.
- Mileage is a strong positive influencer.

Feature Engineering

Created new features

bike_age = Current Year - model_year
power_to_cc ratio (if CC present)

Scaling & Normalization

for numeric columns (mileage, power, cc, kms_driven) to bring features to similar ranges.

One-Hot Encoding

Brand, Owner Type, Location

Log Transformation

on price and kms_driven to handle skewed distributions.

(feature creation + encoding + transformation + scaling)

Model Training

Models Tested:

- Linear Regression – Simple baseline, interpretable but limited for complex data.
- Decision Tree Regressor – Handles non-linearities, but prone to overfitting.
- Random Forest Regressor (Best Performer) – Robust ensemble method, reduces variance, captures feature interactions.

Train-Test Split:

- Dataset split into 80% training, 20% testing for unbiased evaluation.

Cross Validation:

- Applied k-Fold Cross Validation (k=5) to ensure stable performance across different subsets.

Feature Importance:

- Random Forest helped identify top predictors: Bike Age, Kms Driven, Mileage, Power.

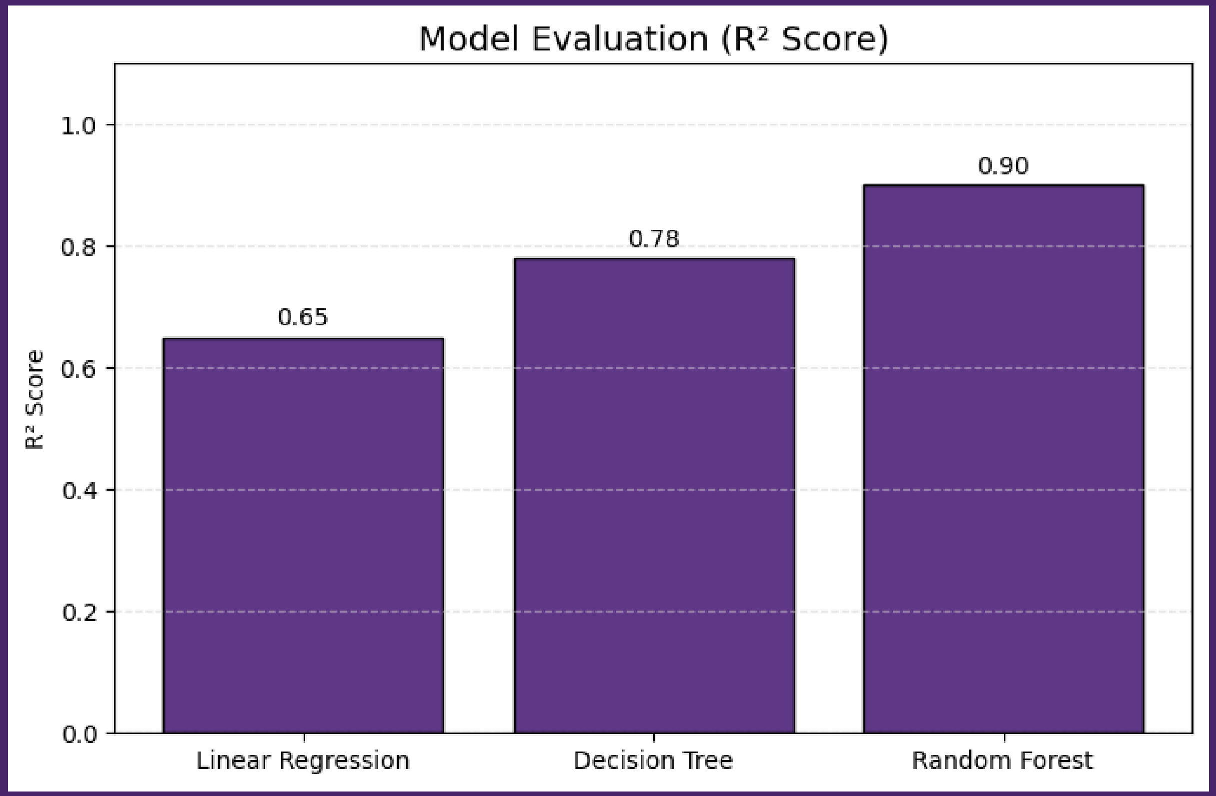
Model Evaluation

Metrics

- R^2 (Coefficient of Determination)
- Root Mean Squared Error
- Mean Absolute Error

Random Forest

- achieved highest R^2 score.



Robust against outliers & feature interactions.

Results

Random Forest gave the best accuracy

Example Prediction

- Input: 2020 Bajaj | 5000 km | 20 mileage | power 20 | engine 150 | owner type
- Predicted Price: ₹234,536

Deployment

Web App Built with Streamlit

- Lightweight, interactive, and easy to deploy.
- Clean UI with customized background & theme.

Workflow:

- User Input → Preprocessing (Scaling + Encoding) → ML Model → Predicted Price

Features:

- Real-time predictions within seconds.
- Handles both numeric & categorical inputs.
- Error-free deployment by saving artifacts

Accessibility:

- Can be hosted on Streamlit Cloud / Heroku for public use.

Business Use Cases

- **Buyers:** Know fair market value.
- **Sellers:** Set competitive prices.
- **Dealers/Showrooms:** Build trust with transparent pricing.
- Platforms (OLX, Droom, CredR): Improve customer experience.

Future Scope

Deploy app on Cloud(Heroku/AWS/GCP).

Expand dataset with more bike brands & specs.

Add Recommendation System: “Best bikes under ₹70K.”

Integrate with dealer platforms for real use.

ML can predict resale bike prices accurately

CONCLUSION

This project successfully demonstrates the application of Machine Learning in predicting used bike prices.

Through data cleaning, feature engineering, and robust modeling, we built an interactive Streamlit app that provides instant price predictions.
The solution not only showcases technical depth but also emphasizes practical usability in real-world scenarios.