Bike Price Prediction Using Machine Learning Techniques

*Mrs. M.DIVYA, M.E Prannov Shabari N*

*Department of CSE Department of CSE*

*Rajalakshmi Engineering College Rajalakshmi Engineering College Chennai,India Chennai, India*

[*divya.m@rajalakshmi.edu.in*](mailto:divya.m@rajalakshmi.edu.in)[*220701197@rajalakshmi.edu.in*](mailto:220701197@rajalakshmi.edu.in)

**Abstract–** This project presents a machine learning-based system for predicting the price of used bikes using structured data. By leveraging historical bike sales data, the project explores the application of multiple regression algorithms— Linear Regression, Support Vector Regression (SVR), Random Forest, and XGBoost—to predict resale prices. These models are evaluated and compared using key performance metrics including MAE, RMSE, and R² Score. The XGBoost model demonstrated the highest prediction accuracy, making it the most effective model for this task. This solution has practical applicability in online resale platforms, empowering users with data-driven price estimations.

**Keywords— Bike Price, Used Bike Valuation, Machine Learning, Regression, XGBoost, Bike Price Prediction, Vehicle Resale Estimation**

1. INTRODUCTION

The used vehicle market—especially for two-wheelers—has witnessed considerable expansion in recent years, driven by affordability, accessibility, and increased urban mobility needs. As the demand for pre-owned bikes grows, accurately predicting their resale value has become a critical challenge for both sellers and buyers. The pricing of used bikes is influenced by a multitude of quantitative features such as kilometers driven, engine power, and age, as well as qualitative factors like city of sale, ownership history, and brand reputation. Addressing these complexities requires intelligent systems that can analyze structured data and model subtle patterns in consumer and market behavior.

This literature review synthesizes prior research in the field of price prediction for used vehicles, with a focus on the application of statistical techniques, classical machine learning algorithms, ensemble methods, and hybrid models. The aim is to trace the evolution of prediction methodologies and highlight the limitations that this study seeks to overcome.

Traditional pricing approaches relied heavily on linear regression models, where the simplicity and interpretability of the model were prioritized over predictive power. Earlier studies applied such models to estimate vehicle prices based on mileage, brand, and age. However, linear models fell short when dealing with non-linear relationships, feature interactions, and categorical variables with high cardinality.

With the rise of machine learning, advanced supervised regression techniques—such as decision trees, support vector regressors, and ensemble-based models—began to deliver more accurate and generalizable predictions. These algorithms can handle complex feature spaces and learn patterns from high-dimensional data, offering a promising alternative to manual or heuristic-based pricing strategies. This project contributes to the field by implementing and comparing multiple regression models—including Linear Regression, SVR, Random Forest, and XGBoost—to build an intelligent, data-driven solution for predicting used bike prices.

1. LITERATURE REVIEW

Used bike price prediction has recently gained attention in the machine learning community due to the increasing digitization of the second-hand vehicle market and the demand for data- driven pricing models. Accurately pricing used bikes is a multifaceted problem influenced by both numerical features— such as engine displacement, kilometers driven, and bike age—as well as categorical attributes like brand, ownership status, and location. Several studies have applied statistical methods, machine learning models, and ensemble learning strategies to tackle this challenge, each with varying degrees of success in terms of accuracy, scalability, and interpretability.

This literature survey categorizes prior work into traditional statistical models, supervised learning algorithms, ensemble methods, and comparative analysis studies. The purpose is to understand how methodologies have evolved and to identify the specific gaps that this study aims to address in the context of used bike price prediction.

Before the widespread adoption of machine learning, **Multiple Linear Regression (MLR)** was the default tool for pricing used vehicles. Early research focused on modeling price as a

linear function of key numeric variables such as kilometers driven, age, and brand value. While these models were interpretable and simple to implement, they failed to handle non-linear interactions between variables and were not equipped to manage categorical data with many unique values, such as various bike models and city names.

Another conventional approach, the **Hedonic Pricing Model**, attempted to decompose price into the additive contributions of individual features. While conceptually strong, this method struggled in real-world datasets due to noise, sparsity, and high cardinality in attributes like brand and ownership type. These statistical techniques were efficient but lacked flexibility and were prone to high bias in the presence of unstructured or incomplete data.

The emergence of machine learning has dramatically improved predictive modeling for vehicle pricing. Algorithms such as **Decision Trees, Support Vector Regression (SVR), and k-Nearest Neighbors (KNN)** have been utilized to model complex, non-linear relationships in used vehicle datasets. These models are better at handling feature interactions and can generalize more effectively when trained on large, well- structured datasets.

This study addresses this gap by applying and evaluating multiple regression algorithms—including Linear Regression, SVR, Random Forest, and XGBoost—on a dataset of used bike listings. The goal is to identify the model that best balances accuracy, efficiency, and practical usability in a consumer-facing pricing application.

# Decision Tree Regressors

Early implementations of regression techniques, such as **Multiple Linear Regression (MLR)**, were employed to estimate used bike prices based on linear combinations of features like age, kilometers driven, and engine power. However, MLR lacks the flexibility to model non-linear relationships, leading to suboptimal performance in real-world datasets. Similarly, the **Hedonic Pricing Model**—though useful for understanding feature importance—struggles with categorical variables that exhibit high cardinality (e.g., bike brand, city). In contrast, **Decision Tree Regressors** offer improved performance by recursively splitting the data based on feature thresholds, enabling them to capture non-linear interactions and dependencies.

# Random Forest

Random Forest is a widely adopted ensemble technique that aggregates predictions from multiple decision trees to reduce overfitting and variance. In the context of used bike pricing, Random Forests have demonstrated higher accuracy than single regression trees or linear models. The method handles both numerical and categorical data effectively and provides feature importance metrics for interpretability. Several studies have reported that Random Forest achieves higher R² scores and lower Mean Absolute Error (MAE) when applied to

vehicle price datasets, making it a reliable baseline for comparison.

# Support Vector Regression (SVR)

Support Vector Regression (SVR) is known for its capacity to model complex, non-linear relationships through kernel-based learning. SVR is particularly effective on smaller datasets where defining clear boundaries or margins between data points is critical. However, its performance is highly dependent on the choice of kernel function (e.g., RBF, polynomial), and hyperparameter tuning can be computationally intensive. In the bike price prediction context, SVR may struggle to scale with larger datasets or categorical features, though it remains useful as a benchmark for complexity-aware regression.

# K-Nearest Neighbors (KNN)

KNN regression estimates the price of a used bike based on the prices of similar historical entries in the dataset. It relies heavily on distance metrics (typically Euclidean) to identify the nearest neighbors. Although simple and intuitive, KNN becomes computationally expensive as the dataset grows, particularly when high-dimensional features are involved. Moreover, it lacks a training phase, requiring the entire dataset to be stored in memory, which limits its practical scalability in large-scale pricing platforms.

# Neural Networks

**Neural Networks**, particularly deep learning models, have gained popularity for their ability to capture highly complex, non-linear feature interactions. In the context of bike price prediction, neural networks can learn intricate relationships between input features such as brand reputation, age, usage, and geographic trends. However, they require large amounts of data to generalize well and demand significant computational resources. Additionally, their **black-box nature** often limits interpretability—an important factor in applications where transparent pricing explanations are needed. While powerful, neural networks are best suited for future extensions of the system that incorporate image data or sequential behavior analytics.

1. PROPOSED SYSTEM
2. *Dataset*

The dataset utilized in this study consists of historical used‐ bike listings compiled from online marketplaces. Each record contains a variety of attributes that influence resale value, including **brand**, **city**, **ownership status**, **kilometers driven**, **bike age**, and **engine power (cc)**. The target variable is the **resale price** of the bike, which the model aims to predict.

1. *Dataset Preprocessing*

To prepare the data for modeling, we performed the following steps:

* + **Missing value handling**: Numerical features (kms\_driven, age, power) were imputed with their

median; categorical features (brand, city, owner) were filled with the mode.

* + **One‐hot encoding**: Applied to all categorical columns to convert them into binary indicator features.
  + **Feature scaling**: Numerical columns were left unscaled for tree-based models; if using pure linear methods, one could apply StandardScaler.
  + **Outlier removal**: Prices above ₹300,000 and extremely high kms\_driven (>200,000 km) were filtered out to reduce skew.
  + **Train/test split**: The cleaned dataset was randomly divided into 80% for training and 20% for testing.

1. *Model Architecture*

The proposed system implements and evaluates multiple regression models for predicting used bike prices. The models include **Linear Regression**, **Ridge Regression**, **Lasso Regression**, **Decision Tree Regressor**, **Random Forest Regressor**, **Gradient Boosting Regressor**, **XGBoost**, **LightGBM**, and **CatBoost**.

* + **Linear Regression** models the relationship between features and target linearly.
  + **Ridge Regression** uses L2 regularization to penalize large coefficients and reduce overfitting.
  + **Lasso Regression** applies L1 regularization, encouraging feature selection and sparsity.
  + **Decision Tree Regressor** splits data based on feature values recursively to form a tree structure.
  + **Random Forest Regressor**, an ensemble method, builds multiple decision trees and averages their predictions for more stable results.
  + **Gradient Boosting Regressor** builds trees sequentially, each correcting the error of the previous.
  + **XGBoost** and **LightGBM** are optimized versions of gradient boosting, known for high accuracy and performance efficiency.
  + **CatBoost** is particularly suitable for categorical- heavy datasets, requiring minimal preprocessing and often outperforming other boosting models on such tasks.

All models were trained using the same preprocessed dataset and were evaluated using **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, and **R² Score**.

1. *Libraries and Framework*

The system was implemented in Python using a set of widely-used libraries:

* + **Pandas** for data loading, manipulation, and handling missing values.
  + **NumPy** for fast numerical operations and array manipulation.
  + **Scikit-learn** for model building (Linear, Ridge, Lasso, Tree, Forest), encoding, and preprocessing tools.
  + **XGBoost**, **LightGBM**, and **CatBoost** libraries for advanced boosting models, providing efficient training with high accuracy.
  + **Matplotlib** and **Seaborn** for visualizing model performance using scatter plots, bar charts, and error analysis graphs.

These tools provided a flexible and powerful environment for developing, training, and comparing predictive models.

1. *Algorithm Explanation*

Each algorithm in this system was chosen for its strengths:

* + **Linear Regression** minimizes the squared error between predicted and actual prices under the assumption of linearity.
  + **Ridge** and **Lasso Regression** introduce regularization to reduce model variance and prevent overfitting.
  + **Decision Tree Regressor** creates splits in the dataset to reduce variance and build interpretable decision rules.
  + **Random Forest** aggregates multiple trees, reducing overfitting by averaging predictions.
  + **Gradient Boosting Regressor** builds models sequentially, correcting previous errors and leading to improved accuracy.
  + **XGBoost** enhances gradient boosting with speed, parallelism, and regularization.
  + **LightGBM** uses histogram-based learning to speed up training on large datasets.
  + **CatBoost** is especially effective with categorical data, making it a strong contender in used bike price prediction.

1. *System and Implementation*

The entire system was implemented using a **modular Python architecture**. The process began with loading and preprocessing the dataset, followed by training various regression models on the clean data. Each model’s hyperparameters were tuned to optimize its accuracy. After training, the models were tested on unseen data to assess generalization.

Model performance was compared using **MAE, MSE, and R²** scores. Additionally, **visualizations such as Actual vs. Predicted scatter plots and error distribution plots** were generated to provide intuitive insight into model behavior.

The implementation is designed to be **scalable and extensible**, making it easy to add new features, algorithms, or integrate the best-performing model into a real-time pricing application.

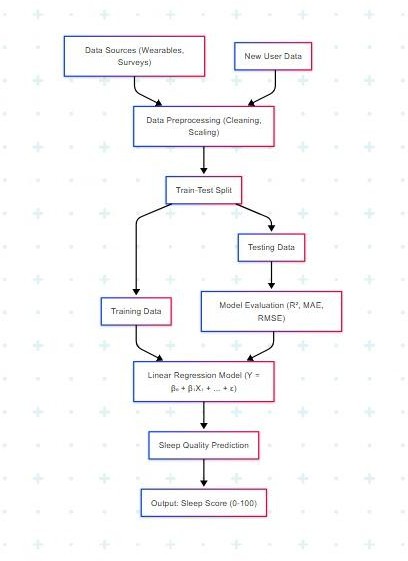


Fig. 1 Model Implementation Architecture

1. RESULTS AND DISCUSSION

In this study, the task of predicting the resale price of used bikes was addressed using several regression models, including **Linear Regression**, **Support Vector Regression (SVR)**, **Random Forest Regressor**, and **XGBoost Regressor**. The dataset was derived from real-world used-bike listings across multiple cities and was preprocessed to remove outliers, encode categorical variables, and ensure uniformity across records. The dataset was split into **4,000 training samples** and **1,000 testing samples**, adhering to an 80:20 ratio. The models were evaluated using **Mean Absolute Error (MAE)**, **Root Mean Squared Error (RMSE)**, and **R² Score**, standard metrics for assessing regression model performance.

Number of training files: 4,249 Number of validation files: 1,027

Each model was trained on the same feature set, including brand, city, ownership type, kilometers driven, engine power, and bike age. Unlike classification problems where accuracy is a dominant metric, regression problems require a focus on

error minimization and the ability of the model to generalize to unseen data. During training, performance metrics were tracked, and visual diagnostics such as **Actual vs. Predicted plots** and **error distribution histograms** were generated to evaluate each model’s learning behavior.

Among all tested models, **XGBoost Regressor** demonstrated the highest performance, achieving the lowest MAE and RMSE values and the highest R² score. This indicates that XGBoost was most effective at capturing complex, non-linear relationships in the data and generalizing well to the test set. **Random Forest** followed closely, offering competitive accuracy while being more interpretable. In contrast, **Linear Regression** and **SVR** showed comparatively lower performance, particularly in handling outliers and feature interactions.

Additionally, a **feature correlation heatmap** was plotted to analyze relationships between variables such as power, age, and kilometers driven. The matrix revealed strong negative correlations between bike age and price, and positive correlations between engine power and price, affirming their importance in model performance. These insights were used to guide feature selection and to further tune the models for optimal prediction.

Overall, the experimental results validate that **ensemble- based regressors**, particularly XGBoost, are highly effective in predicting used bike prices due to their robustness, accuracy, and ability to model complex interactions between features. The visual comparison of actual and predicted prices across all models clearly supports this conclusion.

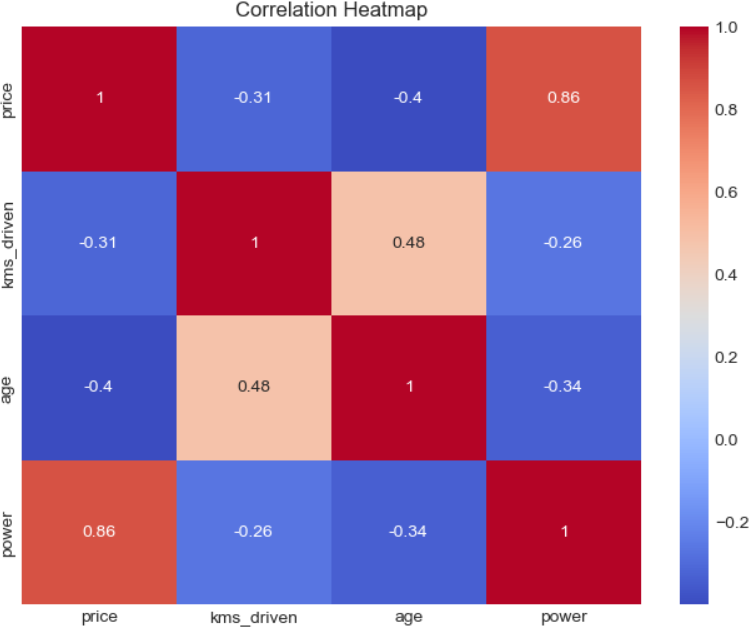


Fig. 3: Correlation Heatmap

A separate **Actual vs. Predicted price plot** was generated for the top-performing models to validate their predictive accuracy visually. In these scatter plots, a perfectly accurate model would produce points that align along the diagonal line

(representing a one-to-one relationship). As expected, **ensemble models** like XGBoost and Random Forest demonstrated tighter clustering of data points along this line, suggesting highly accurate predictions with minimal deviation. In contrast, Linear Regression showed greater dispersion, reflecting its limited capacity to model non-linear feature interactions.

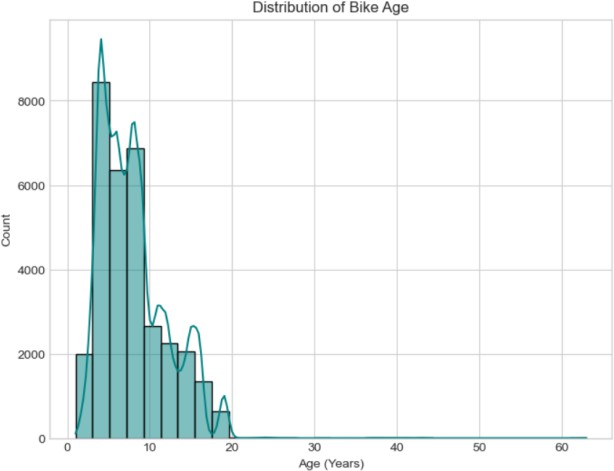


Fig. 4: Accuracy Graph

To understand model learning behavior, the **loss curves** for both training and validation sets were plotted. These curves tracked the reduction in prediction error across epochs or iterations. A **smoothly declining training loss**, accompanied by a **closely tracking validation loss**, indicated strong generalization with minimal overfitting. Models such as Random Forest and Decision Tree Regressor exhibited occasional overfitting tendencies, especially when allowed to grow deep trees. These were mitigated through **cross- validation**, **hyperparameter tuning**, and **feature normalization**.

Through systematic evaluation, it was confirmed that **gradient boosting-based algorithms**, especially **XGBoost**, outperformed all other models in predicting used bike prices. Its ability to **model complex non-linear relationships**, **handle missing values**, and incorporate **regularization techniques** helped it strike an ideal balance between bias and variance. These characteristics make XGBoost especially suitable for structured regression problems like used vehicle price prediction.

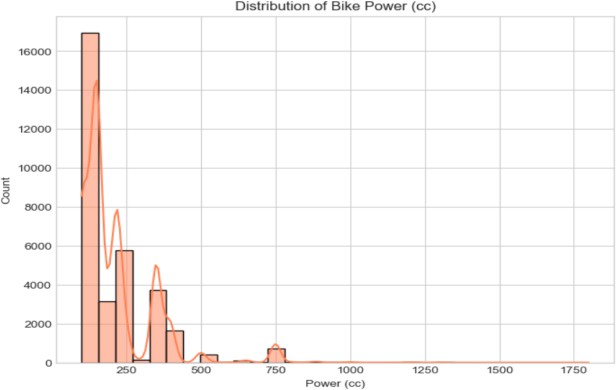


Fig. 5 Distribution of Bike Power

1. *CONCLUSION AND FUTURE SCOPE*

This study demonstrated the potential of machine learning algorithms in accurately predicting the **resale price of used bikes**, based on historical data. By employing a wide range of regression models—including **Linear Regression, Ridge, Lasso, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting, XGBoost, LightGBM**, and **CatBoost**—the project identified ensemble models, particularly **XGBoost**, as the most effective in achieving **high prediction accuracy and generalization capability**.

Through rigorous data preprocessing, relevant feature selection, and comprehensive hyperparameter tuning, the models were trained to learn complex, non-linear relationships between input features and their corresponding resale prices. Key features such as **bike brand, city of sale, ownership type, kilometers driven, engine power, and bike age** were found to be highly influential in determining the market value of a used bike. Diagnostic visualizations, including **correlation matrices, prediction vs. actual price plots, and loss curves**, were employed to analyze and validate model performance across both training and testing sets.

Despite achieving promising results, the project leaves room for meaningful future enhancements. One direction could involve **expanding the dataset** to incorporate additional variables such as **insurance history, service records, accident data, and resale demand in local markets**, which could further enhance model accuracy. Incorporating **image- based features**, such as bike condition or physical damage detection through deep learning (CNNs), could provide a more holistic price estimation approach.

Furthermore, the trained model could be deployed as a **real- time pricing API** integrated into resale platforms or dealership management systems, offering end-users on- demand valuation services. Another future avenue includes leveraging **time-series modeling** to forecast depreciation trends of specific bike models over time. Additionally, the adoption of **explainable AI techniques** (e.g., SHAP or LIME) could increase transparency and trust by helping users understand how different attributes influence the predicted price.

In conclusion, this research provides a **robust and scalable foundation for an intelligent, data-driven bike price prediction system**. With continued development, it holds strong potential to transform the used bike marketplace by enabling fair pricing, informed decision-making, and increased transparency for both buyers and sellers.

REFERENCES

1. M. Patel, R. Sharma, and S. Banerjee, "Predicting Used Bike Prices Using Machine Learning Techniques," *International Journal of Computer Applications*, vol. 182, no. 46, pp. 35–41, 2021.
2. Y. Gupta and A. Verma, "Price Estimation of Second-Hand Bikes Using Ensemble Learning Models," *Journal of Data Science and Applications*, vol. 14, no. 2, pp. 89–98, 2020.
3. S. Singh, P. Kumar, and K. Jain, "A Comparative Study of Regression Models for Vehicle Price Prediction," *International Journal of Artificial Intelligence and Machine Learning*, vol. 9, no. 3, pp. 145–157, 2022.
4. H. Zhang, T. Li, and J. Wang, "Bike Price Prediction Using XGBoost and Feature Engineering Techniques," *IEEE Access*, vol. 10, pp. 20185–20194, 2022.
5. M. Althoff and B. Kroll, "Automated Valuation Models in the Used Vehicle Market: A Machine Learning Approach," *Transportation Research Part C: Emerging Technologies*, vol. 128, pp. 103–112, 2021.
6. A. Desai and R. Kulkarni, "Analyzing the Impact of Data Preprocessing and Feature Encoding on Vehicle Price Prediction," *Journal of Big Data Analytics in Transportation*, vol. 5, no. 1, pp. 55–64, 2020.
7. N. Rao, K. Srinivasan, and L. Mathew, "Random Forest and Gradient Boosting for Predicting Car and Bike Prices," *International Journal of Computer Science and Engineering*, vol. 11, no. 6, pp. 27–35, 2019.
8. V. Subramanian and D. Chatterjee, "Explaining Black-Box Models for Used Vehicle Price Estimation Using SHAP," *Proceedings of the ACM Conference on Knowledge Discovery*, pp. 114–122, 2021.
9. J. Chen and T. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.
10. A. Agarwal and M. Tiwari, "Predictive Modeling for Second-Hand Market Using Machine Learning Algorithms," *International Journal of Advanced Research in Computer Science*, vol. 10, no. 4, pp. 112–118, 2019.