Laptop Price Estimation Using Advanced Machine Learning Algorithms

Dilip Adhikari

Abstract

Laptop prices vary widely due to differences in performance, brand reputation, hardware features, and changing consumer demand. Accurately estimating laptop prices is challenging because many of these factors interact in complex and non-linear ways. This project focuses on predicting laptop prices using supervised machine learning techniques based on key technical specifications. The study uses a dataset of 1,303 laptop records and applies systematic feature engineering, including the calculation of Pixels Per Inch (PPI), extraction of CPU and GPU brands, and decomposition of detailed memory configurations. Since the target variable, laptop price, exhibited a skewed distribution, a logarithmic transformation was applied. Multiple regression models, including linear models, tree-based ensembles, and a neural network, were trained and evaluated. Among the tested models, the XGBoost demonstrated strong generalization performance, achieving an R^2 value of 0.889, Mean Absolute Error (MAE) of 0.2024, and RMSE of 0.206 on the test data. These results confirm that sophisticated feature representation combined with ensemble techniques is highly effective for price forecasting in a heterogeneous technology market.

I. Introduction

Laptop pricing is influenced by a wide range of factors, including hardware components, performance capabilities, product design and others. As technology evolves rapidly, new models with different specifications enter the market frequently, making price estimation difficult for both consumers and sellers. As a result, determining whether a laptop is fairly priced based on its specifications is not straightforward. Machine learning provides a practical solution to this problem by modeling the relationship between laptop features and prices directly from data. Unlike simple pricing rules, machine learning models can capture non-linear patterns and interactions among variables such as processor type, memory configuration, display quality, and brand. The primary goal of this project is to develop, tune, and compare several supervised regression models to identify a reliable and accurate approach for predicting laptop prices.

II. Dataset Overview

The dataset selected for the project is Laptop Price Dataset from open-source Kaggle. The dataset consisted of about 1,303 unique laptop entries with 12 initial features as Company, TypeName, Inches (Screen Size), Screen Resolution, CPU, RAM, Memory, GPU, Operating System, Weight, and Price.

III. Feature Engineering

Effective feature engineering played a crucial role in improving the predictive performance of the regression models by transforming raw textual and categorical attributes into meaningful numerical representations.

A. Data Cleaning

The dataset was checked for missing values, inconsistencies, and outliers.

B. Exploratory Data Analysis (EDA)

Statistical summaries and visualizations (histograms, box plots, etc.) were used to understand variable distributions and identify trends.

*1) **Screen Resolution & Pixel Density (PPI):*** The ScreenResolution column was parsed to extract binary flags for Touchscreen and IPS panel presence, as well as the discrete X and Y resolutions. These were used to compute the crucial Pixels Per Inch (PPI) metric by using the relation:

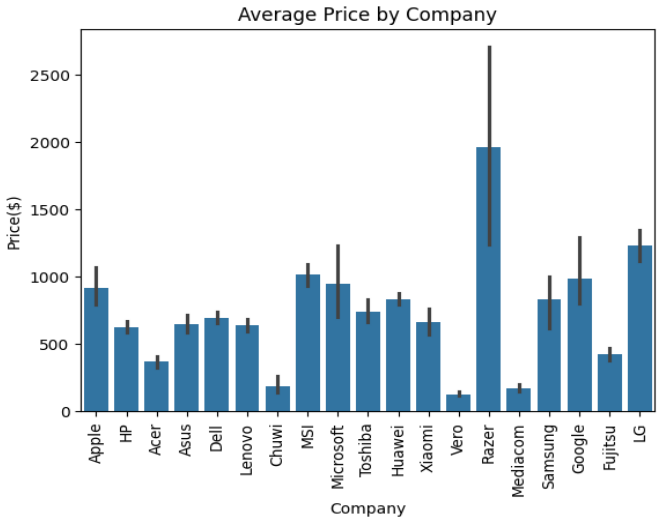

Fig. 1. Average Price by Company

$$\text{PPI} = \frac{\sqrt{X_{\text{res}}^2 + Y_{\text{res}}^2}}{\text{Screen Size (inches)}}$$

This transformation compresses display quantity into a single continuous variable that more accurately reflects visual sharpness than raw resolution values.

*2) **Component Brands**:* The highly verbose CPU and GPU columns were simplified by extracting the primary brand name like 'Intel Core i5, Intel Core i7' from CPU, and 'Nvidia, AMD, Intel' from GPU.

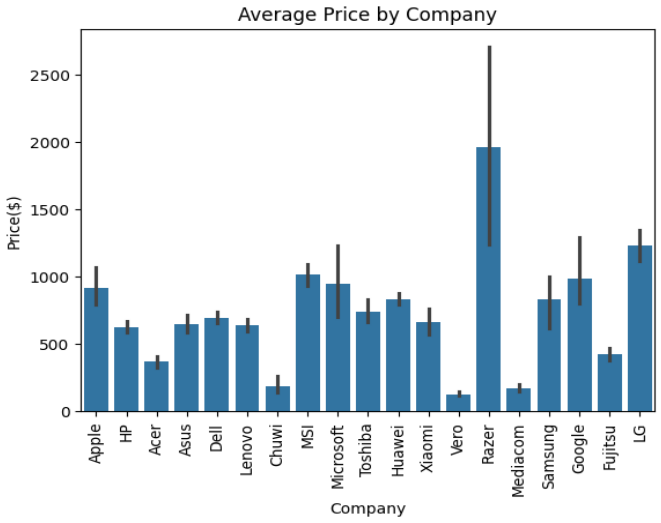

Fig. 2. Price by Company

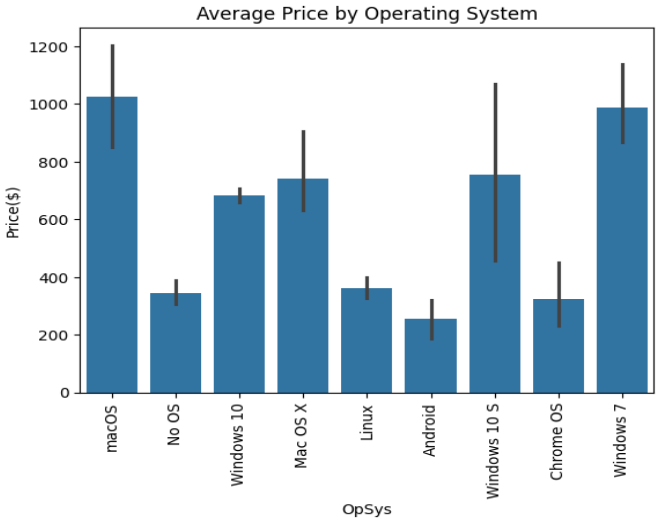

Fig. 3. Price by Operating System

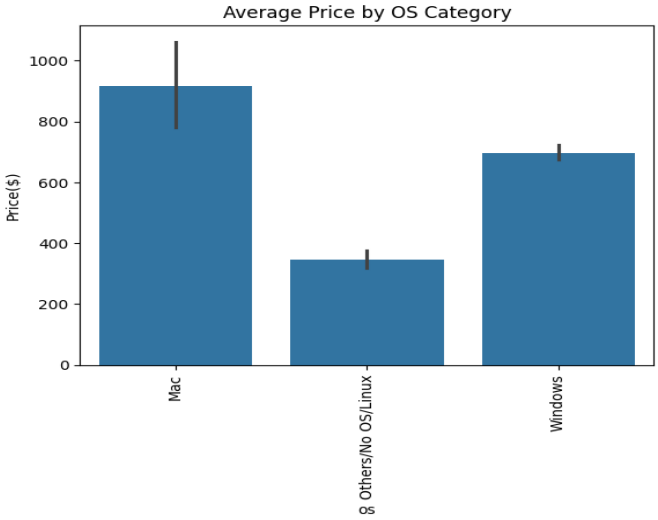

Fig. 4. price by OS Category

*3) **Memory Configuration**:* The complex Memory string feature was decomposed to quantify total storage across various types as HDD, SSD, Hybrid, and Flash Storage.

*4) **Operating System (OpSys)**:* The operating systems were categorized into three main classes as Windows, Mac, and Others/Linux/No Os. This categorization reduced noise from sparsely represented operating systems while preserving their pricing influence.

*5) **Target Variable Transformation**:* The distribution of the target variable, Price($) exhibited strong right skewness (non-normal). To improve

the convergence of non-linear models and satisfy the assumptions of linear models, the price was transformed using a log-transformation technique. This improved normality, stabilized variance, and enhanced model training behavior.

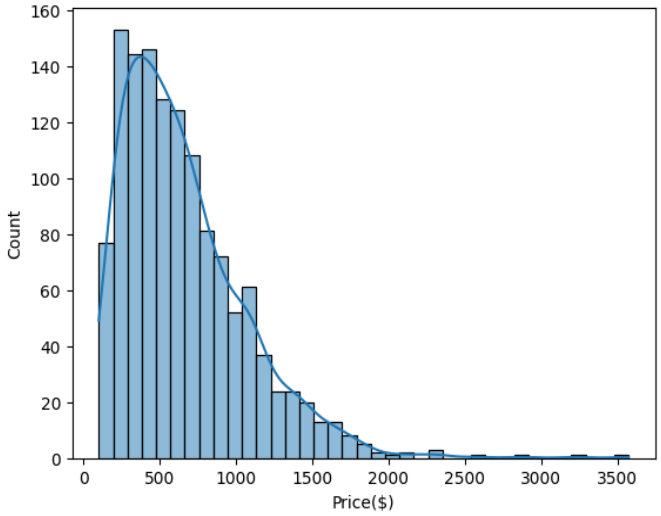

Fig. 5. Price Distribution

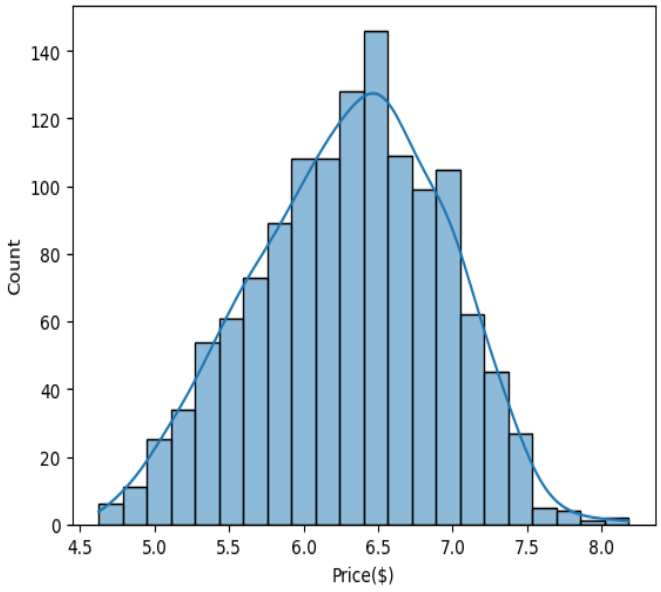

Fig. 6. Price Distribution After Log Transformation

6) *Preprocessing and Scaling:*

- **One-Hot Encoding**
  All relevant categorical features Company, TypeName, CPU Brand, GPU Brand, OpSys were encoded using OneHotEncoder with the drop='first' setting to mitigate multicollinearity.
- **Data Splitting**
  The dataset was split into 75% training and 25% testing subsets to validate model generalization.

C. *Correlation*

Correlation is an important statistical approach where if correlation between variables is -1, it means a strong negative correlation; 0 means no correlation, and 1 means strong positive correlation.

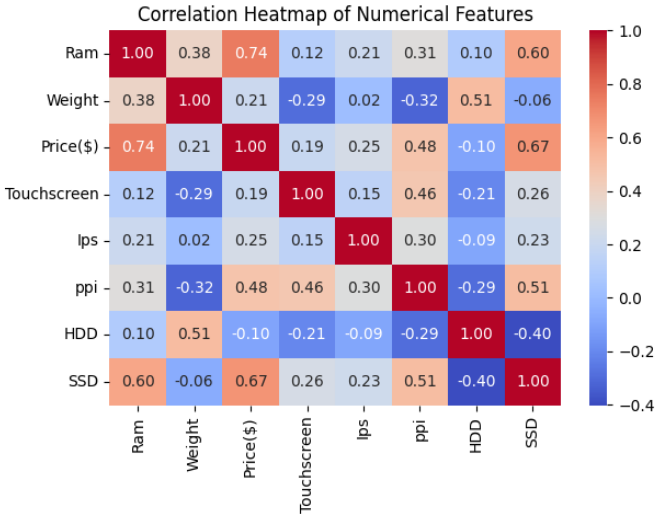

Fig. 7. Correlation Matrix

The correlation analysis shows that Price($) is primarily driven by hardware performance and storage type. Ram (0.74), SSD (0.67), and PPI (0.48) exhibit the strongest positive relationships with price, confirming that laptops with more memory, faster solid-state storage, and higher-resolution displays command significantly higher prices. Touchscreen and IPS display features show only moderate influence, while HDD storage and Weight provide minimal predictive value and can be considered secondary factors. Overall, the data clearly indicates that premium laptops are defined by high RAM, SSD storage, and superior display quality, while basic physical attributes like weight or HDD capacity contribute little to price determination.

IV. Statistical Method

A diverse selection of regression algorithms was implemented and evaluated to ensure a comprehensive comparison of model bias and variance characteristics.

A. *Linear Model*

These models are based on Linear Regression but include a penalty term (regularization) applied to the coefficients to prevent overfitting and improve generalization, especially with high-dimensional data.

- **Ridge Regression**
- **Lasso Regression**

B. *Kernel-Based Model*

A Support Vector Regressor (SVR) is a supervised machine learning algorithm used for its ability to classify with high dimensionality. It works by finding the optimal hyperplane that separates different classes of data points.

C. *Tree-Based Ensemble Methods*

- **Decision Tree Regressor**
- **Random Forest Regressor**
- **Gradient Boosting Regressor**
- **XGBoost Regressor**

These models are ensemble learning method that builds multiple decision trees and combines their results to improve prediction accuracy and reduce overfitting.

D. *Distance-Based Model*

A KNN is a supervised learning algorithm that works by finding the K-nearest data points to a new data point and using their class labels or values to make predictions

E. *Model Evaluation*

Models were evaluated using Model performance was rigorously assessed using three key regression metrics:

*1) **Mean Absolute Error (MAE)**:* MAE is the average of the absolute differences between the actual values and the predicted values. It tells, on average, how much predictions are ”off” from the actual values.

*2) **Mean Absolute Error (MAE)**:* Root Mean Squared Error (RMSE): RMSE is the square root of MSE. It brings the units back to the same scale as the original target variable, making it easier to interpret.

*3) **R-Squared(R^2)**:* Measures the proportion of the variance in the dependent variable that is predictable from the independent variables. The coefficient of determination (R^2) measures the proportion of variability in the response variable explained by the model, with values closer to 1 indicating better model fit, and negative values indicating performance inferior to the mean baseline.

V. Results and Interpretation

The analysis revealed that certain features significantly influence price of the laptop:

- **Brand :** Premium pricing is heavily driven by brand equity. Manufacturers like Apple, Razer, and LG consistently command higher price points. Similarly, laptops running macOS or high-end Windows configurations are priced significantly higher than those using entry-level operating systems.
- **RAM:**
- **SSD:**
- **PPI:**

TABLE I
Performance Comparison of Regression Models

Model	MAE	RMSE	R^2
Linear Regression	0.210	0.271	0.807
Ridge Regression	0.209	0.268	0.813
Lasso Regression	0.211	0.272	0.807
KNN	0.193	0.275	0.803
SVR	0.202	0.271	0.808
Decision Tree	0.186	0.254	0.832
Random Forest	0.159	0.208	0.887
AdaBoost	0.226	0.282	0.792
Gradient Boosting	0.159	0.212	0.883
XGBoost	0.152	0.206	0.889

Overall, the Random Forest model provided the best balance between sensitivity, specificity, accuracy, and AUC, making it the most suitable model for approving loan. Logistic Regression also showed reasonable performance and could serve as a simpler alternative if model interpretability is prioritized. Conversely, SVM and KNN demonstrated inadequate performance and are not recommended for this prediction task.

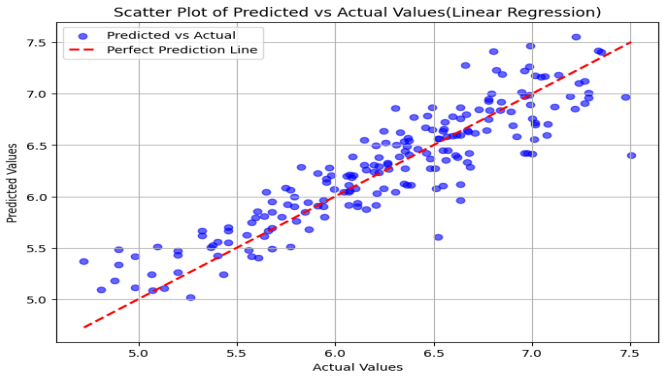

Fig. 8. Linear Regression

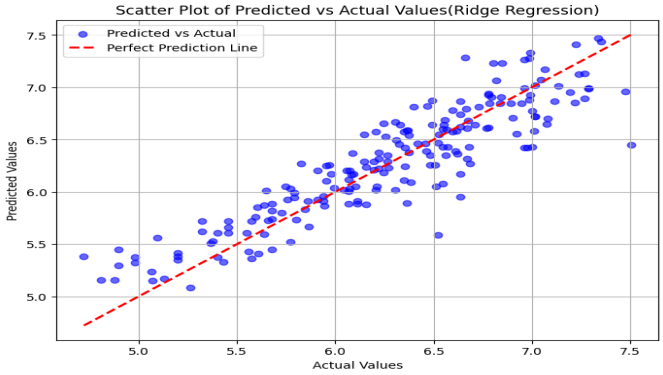

Fig. 9. Ridge Regression

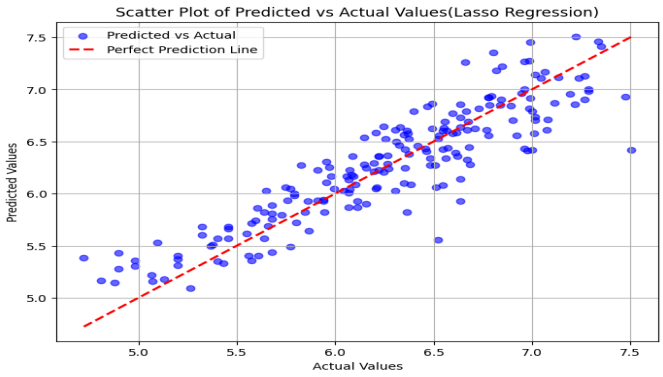

Fig. 10. Lasso Regression

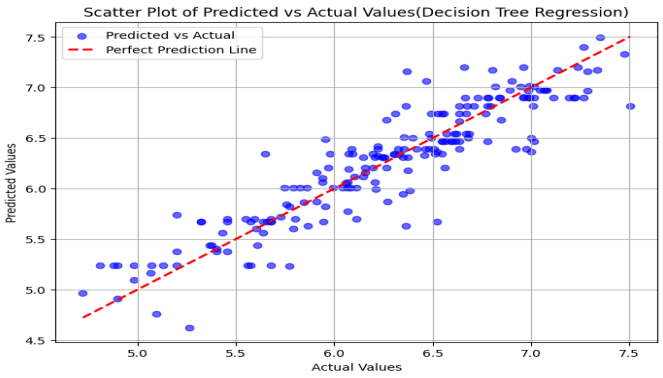

Fig. 11. Decision Tree

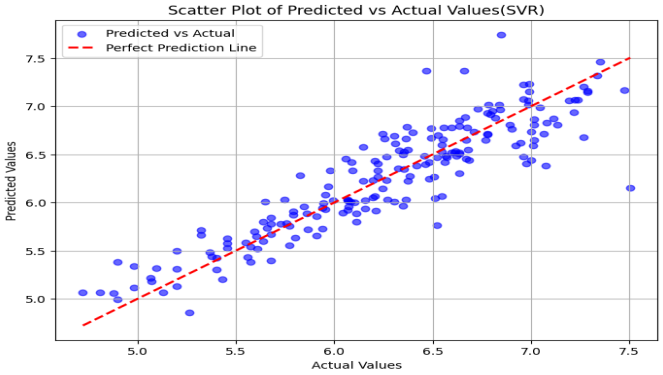

Fig. 12. Support Vector Regressor

TABLE II
PERFORMANCE COMPARISON OF REGRESSION MODELS

Model	R^2	**SD**
Linear Regression	0.8179	0.0356
Ridge Regression	0.8076	0.0323
Lasso Regression	0.8092	0.0345
KNN	0.7641	0.0354
SVR	0.8025	0.0428
Random Forest	0.8718	0.0155
AdaBoost	0.7820	0.0151
Gradient Boosting	0.8785	0.0258
XGBoost	0.8787	0.0182

VI. CROSS VALIDATION

Cross-validation is used to evaluate the performance of a machine learning model by dividing the dataset into multiple subsets (or folds) and training and testing the model on different portions of the data. It ensures that the model's performance is evaluated on unseen data and reduces the risk of overfitting.

Here, I used 5-fold cross validation (CV = 5) for all the above model.

The strong predictive performance of the XGBoost model (R^2 = 0.889, RMSE = 0.206, MAE = 0.152) can be attributed to its gradient-boosting framework, which sequentially combines weak learners to correct the errors of previous models.

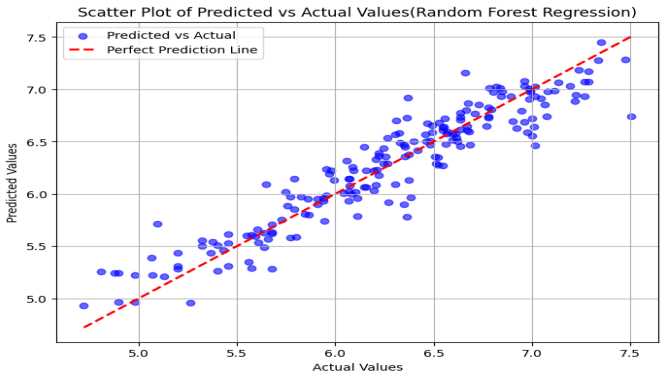

Fig. 13. Random Forest

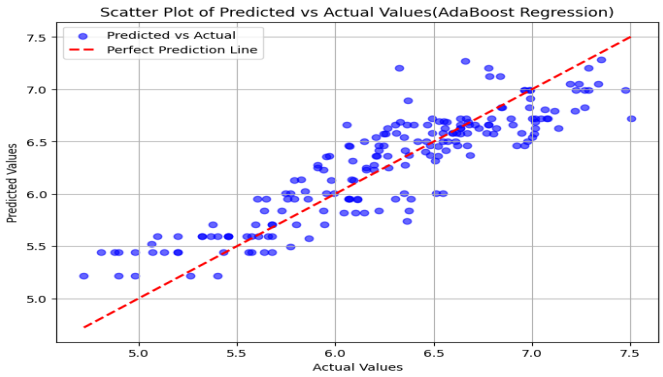

Fig. 14. AdaBoost Regression

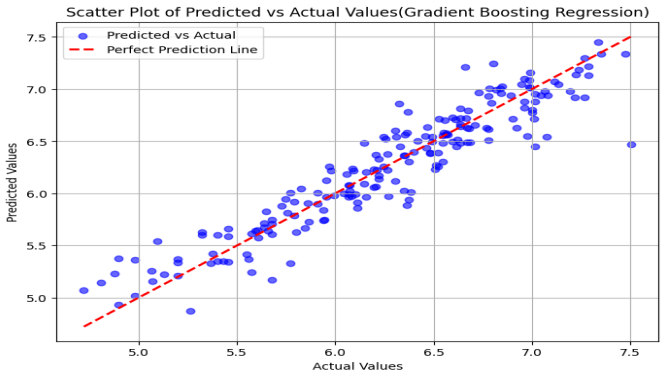

Fig. 15. Gradient Boosting Regression

VII. Conclusion

This study presented a comprehensive evaluation of multiple regression techniques for laptop price prediction, emphasizing both predictive accuracy and model stability. Across all experiments, ensemble tree-based model XGBoost ($R^2 = 0.889$) consistently outperformed linear, distance-based, kernel-based, and neural network approaches, demonstrating their effectiveness in capturing the complex nonlinear relationships and feature interactions inherent in laptop pricing data. Cross-validation results further reinforced these findings, with XGBoost achieving the highest mean predictive performance (Mean $R^2 = 0.8787$) while maintaining low variability across folds. Gradient Boosting exhibited comparable accuracy with slightly higher variance, whereas Random Forest emerged as a highly reliable and interpretable alternative, balancing strong performance with robustness and transparency. In contrast, linear and regularized regression models showed limited capacity to model nonlinear effects, and more complex approaches such as neural networks failed to provide performance gains commensurate with their added complexity. Overall, the results highlight the critical role of domain-informed feature engineering and validate ensemble learning as a robust framework for pricing problems in technology markets. The combination of high predictive accuracy, stability under cross-validation, and interpretability makes tree-based ensemble models particularly suitable for real-world deployment.

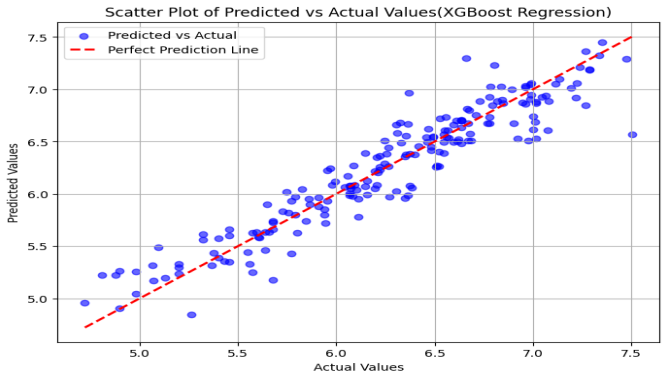

Fig. 16. XGBoost Regression

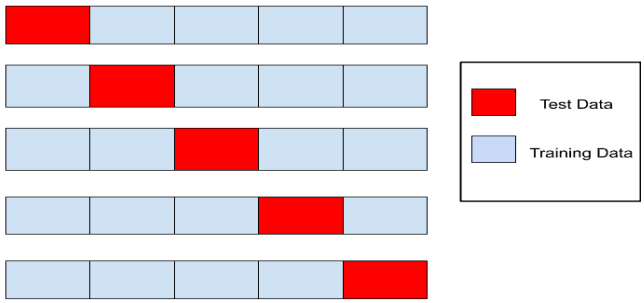

Fig. 17. Cross Validation