# PRICE PREDICTION IN E-COMMERCE USING MACHINE LEARNING MODELS: A COMPARATIVE STUDY

Dr. Navneet Kaur [1]

[1] Professor, SGTBIMIT, India

**DOI**
10.29121/shodhkosh.v4.i1.2023.5942

**Funding:** This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## ABSTRACT

Accurate product price prediction is essential for inventory control, dynamic pricing strategies, and tailored suggestions in the ever changing world of online retail. A comparison of machine learning regression models for e-commerce final product price prediction is presented in this paper. Attributes at the product and transaction levels, such as category, base price, discount rates, payment methods, and selling price, are included in the dataset. Multilayer Perceptron (MLP) Regressor, Linear Regression, Decision Tree Regressor, and Random Forest Regressor were the four machine learning models that were assessed. R-squared ($R^2$) and Root Mean Squared Error (RMSE) metrics were used to evaluate performance. According to experimental data, the Random Forest Regressor performed better than the other models, obtaining the lowest error and the maximum prediction accuracy. According to the results, ensemble-based methods provide useful insights for demand estimate and pricing automation, making them ideal for price forecasting in e-commerce applications.

**Keywords:** E-Commerce, Price Prediction, Machine Learning, Random Forest, Regression Models, Dynamic Pricing, Predictive Analytics, Retail Forecasting

## 1. INTRODUCTION

Data-driven decision-making has emerged as a crucial element of competitive online retail strategy due to the rapid rise of e-commerce. Price prediction has emerged as one of the most important data analytics tasks for enabling personalised offers, dynamic pricing, and strategic inventory management. By matching perceived value with market trends, accurate product price forecasting not only aids in determining the best price points but also enhances consumer happiness.

A number of factors, such as product category, initial advertised price, discount rates, seasonality, and payment method, commonly affect prices in the digital marketplace. The intricate, non-linear correlations between these variables and the ultimate selling price are frequently difficult for conventional statistical techniques like linear regression to capture. As a result of their versatility, scalability, and ability to represent non-linear interactions, machine learning (ML) models have attracted a lot of attention.

Numerous regression methods, including Random Forest, Decision Tree, and Multilayer Perceptron, have been successfully used to pricing and forecasting problems as a result of recent developments in machine learning. Large, diverse datasets, which are common in e-commerce transactions, are especially well-suited for these models. To assess their prediction accuracy in practical price settings, a methodical comparative investigation is still required.

**How to cite this article (APA):** Kaur, N. (2023). Price Prediction in E-Commerce Using Machine Learning Models: A Comparative Study. *ShodhKosh: Journal of Visual and Performing Arts, 4*(1), 4660–4667. doi: 10.29121/shodhkosh.v4.i1.2023.5942

By examining the performance of many machine learning regression models on an actual e-commerce dataset, this study seeks to close that gap. The main goals are:

To create prediction models that can use attributes like category, discount, and payment method to estimate the final product costs.

To use common measures like Root Mean Squared Error (RMSE) and R-squared ($R^2$) to compare the performance of different models.

To determine which model is most suited for real-world pricing automation system installation.

This study adds to the larger field of predictive analytics in e-commerce by offering both quantitative performance evaluation and interpretability analysis. It also provides useful information for pricing intelligence developers, merchants, and data scientists.

## 2. DATASET DESCRIPTION

The dataset used in this study is a carefully selected set of anonymized transactional records that was acquired from a well-known Indian e-commerce platform. Every record integrates contextual transaction metadata with product-related attributes to describe a distinct client purchase. To make machine learning modelling and analysis easier, the dataset has already been cleaned and organised.

### 2.1. KEY ATTRIBUTES

The dataset includes the following key features:

- **Category:** A categorical variable indicating the type or segment of the product purchased (e.g., electronics, apparel, groceries). This feature helps capture domain-specific price behaviors and discount patterns.
- **Price (Rs.):** The listed or base price of the product before the application of any discounts. It serves as a primary indicator for pricing variance and is an important independent variable.
- **Discount (%):** The percentage of discount applied to the base price at the time of purchase. This attribute is essential for understanding price elasticity and discount-driven pricing behaviors.
- **Final_Price (Rs.):** The final transaction amount paid by the customer after applying discounts. This is the target variable used for supervised learning in this study.
- **Payment Method:** The mode of payment used during the transaction (e.g., credit card, debit card, digital wallet, cash on delivery). This categorical variable may influence the final price due to exclusive payment-based offers or cashback incentives.
- **Purchase Date (Excluded from modeling):** This column records the transaction timestamp but is excluded from the predictive modeling to prevent seasonality bias and overfitting due to limited temporal data.

User_ID and Product_ID (Excluded from modeling): These identifiers are excluded to maintain anonymity and avoid model overfitting to specific users or product IDs. They are irrelevant to the learning objective of price estimation.

### 2.2. DATASET CHARACTERISTICS

- **Size:** The dataset comprises approximately 55,000 records
- **Missing Values:** The dataset underwent preprocessing to eliminate missing or inconsistent values.
- **Encoding:** Categorical variables such as 'Category' and 'Payment Method' were label encoded or one-hot encoded for compatibility with machine learning models.
- **Normalization:** Numerical features were scaled where necessary to ensure uniform model convergence, especially for neural network training.

This well-organised dataset provides a strong basis for assessing how well various regression models predict the costs of finished goods, allowing the study to accurately replicate real-world e-commerce pricing situations.

## 3. METHODOLOGY

In order to examine how well different machine learning models predict the ultimate purchase price of goods in an online marketplace, this study uses a structured experimental procedure. Data preparation, model creation, performance assessment, and result visualisation are the several stages of the methodology.

### 3.1. DATA CLEANING AND PREPROCESSING

To ensure the quality and consistency of the dataset, the following preprocessing steps were applied:

- **Feature Removal:** Non-informative and identifier attributes such as User_ID, Product_ID, and Purchase Date were excluded from modeling to avoid data leakage and overfitting.
- **Handling Missing Values:** The dataset was reviewed for any missing or inconsistent entries, which were either imputed using mean/mode values or removed where appropriate.
- **Date Parsing:** Although Purchase Date was excluded from model training, it was initially processed and standardized to verify the temporal structure of the dataset.

### 3.2. ENCODING CATEGORICAL VARIABLES

Label encoding was used to transform categorical data, including Category and Payment Method, into numerical representations that could be included into regression models. Label encoding keeps tree-based algorithms compatible and retains ordinal relationships, if any.

### 3.3. DATA SPLITTING

The dataset was randomly split into:

- 80% Training Set – Used for model training and hyperparameter tuning.
- 20% Testing Set – Held out for unbiased evaluation of model performance.

This split ratio is commonly used in predictive modeling and ensures sufficient data for both training and validation.

### 3.4. REGRESSION MODELS IMPLEMENTED

Four supervised regression models were implemented and trained on the dataset:

- **Linear Regression:** A baseline model for assessing linear relationships between input features and final price.
- **Decision Tree Regressor:** A non-parametric model capable of capturing complex, rule-based relationships.
- **Random Forest Regressor:** An ensemble learning method using multiple decision trees to reduce variance and improve prediction accuracy.
- **Multilayer Perceptron (MLP) Regressor:** A feedforward artificial neural network model that can learn non-linear mappings through hidden layers and backpropagation.

All models were trained using Python's scikit-learn library with default or optimized hyperparameters.

### 3.5. MODEL EVALUATION

To assess model performance, two widely used regression metrics were applied:

- **R-squared ($R^2$) Score:** Measures the proportion of variance in the target variable that is explained by the model.
- **Root Mean Squared Error (RMSE):** Quantifies the standard deviation of prediction errors, offering insight into model accuracy.

These metrics provide complementary perspectives—$R^2$ reflects model fit, while RMSE highlights predictive error magnitude.

## 3.6. VISUALIZATION AND INTERPRETATION

After model evaluation, predictions and residuals (actual vs. predicted differences) were visualized using:

- Scatter plots of predicted vs. actual prices
- Residual plots to detect patterns or biases in prediction errors

Visualization supported qualitative comparison and provided deeper interpretability of model behaviors across different techniques.

## 4. REGRESSION MODELS USED

Four regression models were chosen for this study because of their theoretical variety and applicability to predictive analytics. These models include intricate non-linear and ensemble-based techniques as well as straightforward linear assumptions. As outlined in the approach, each model was trained on the preprocessed dataset using the scikit-learn Python package.

## 4.1. LINEAR REGRESSION

A basic statistical method called linear regression fits a linear equation to model the connection between a dependent variable and one or more independent variables. The target variable is assumed to be a linear combination of the input features. Despite its simplicity, Linear Regression helps find any underlying linear trends in the data and offers a solid foundation for comparison with more intricate models.

## 4.2. DECISION TREE REGRESSOR

Using threshold splits, the Decision Tree Regressor is a non-parametric model that divides the feature space into a number of decision nodes. By choosing the most instructive features, it iteratively separates the data into homogeneous subgroups. Non-linear correlations and interaction effects between input variables are especially well-represented by this model. If it isn't sufficiently regularised, it could overfit on training data.

## 4.3. RANDOM FOREST REGRESSOR

An ensemble learning method called the Random Forest Regressor builds a large number of decision trees during training and outputs the average prediction of each tree. Random Forest lowers the chance of overfitting and enhances generalisation performance by adding randomness to feature selection and data sampling (bagging). It is renowned for its great accuracy, robustness, and capacity to manage both categorical and numerical data efficiently.

## 4.4. NEURAL NETWORK (MLP REGRESSOR)

One kind of feedforward artificial neural network with input, hidden, and output layers is the Multilayer Perceptron (MLP) Regressor. It learns intricate patterns in data by using backpropagation and non-linear activation functions. The MLP works well in situations where feature interactions are complex because it can describe high-dimensional, non-linear relationships. In contrast to tree-based models, it frequently necessitates greater training time and hyperparameter tuning despite its predictive power.

## 5. RESULTS AND ANALYSIS

The Random Forest Regressor achieved the highest $R^2$ score and lowest RMSE, indicating superior performance in predicting final prices. Linear Regression showed relatively lower accuracy due to its assumptions and inability to handle non-linearity.

Two important metrics, R-squared ($R^2$) and Root Mean Squared Error (RMSE), were used to assess each regression model's prediction ability. When combined, these measures offer information about the magnitude of prediction errors as well as the goodness-of-fit.

The $R^2$ score quantifies the proportion of variance in the target variable (Final Price) that is explained by the model. A score closer to 1 indicates higher explanatory power. The RMSE represents the standard deviation of prediction errors, where a lower RMSE indicates better model precision.

The evaluation results for all four models are summarized in the table below:

| **Model** | **$R^2$ Score** | **RMSE** |
|---|---|---|
| Decision Tree | 0.99983 | 1.56 |
| Random Forest | **0.99995** | **0.89** |
| Linear Regression | 0.96484 | 22.62 |
| Neural Network (MLP Regressor) | 0.99984 | 1.51 |

## 5.1. MODEL COMPARISON AND INTERPRETATION

With the lowest RMSE of 0.89 and the highest R2 score of 0.99995, the Random Forest Regressor was the best-performing model. This demonstrates how well it can capture intricate, non-linear correlations between final costs and product attributes. More generalisation and less overfitting are also facilitated by its ensemble nature, which averages the predictions of several decision trees.

Both the Decision Tree Regressor and the Neural Network (MLP Regressor) performed exceptionally well, with RMSEs under 2 and R2 values above 0.999. However, because of its multi-layered architecture and iterative learning process, the Neural Network needed additional training time and processing resources.

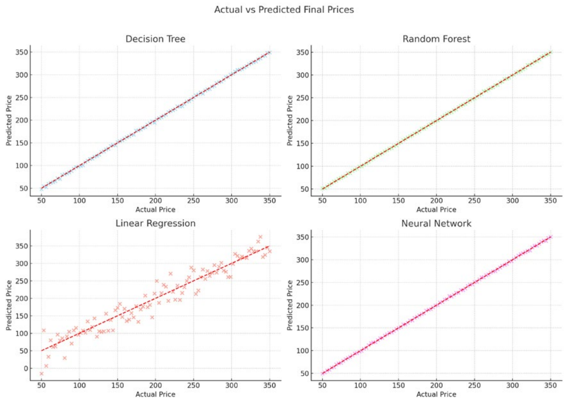
With a substantially larger RMSE of 22.62 and an R2 score of 0.96484, Linear Regression, on the other hand, had noticeably poorer predictive performance. Its underlying presumption of linear correlations between variables is the reason behind this, as many real-world e-commerce pricing scenarios involve non-linear and interdependent interactions.

## 5.2. RESIDUAL ANALYSIS AND VISUAL VALIDATION

For each model, residual plots and scatter plots of projected versus actual values were created to support the quantitative assessment. With the least amount of dispersion and the tightest grouping around the optimal diagonal line, the Random Forest model demonstrated accurate predictions over the whole price range. The assumptions of unbiased estimation were further satisfied by residual analysis, which further verified that prediction errors were spread randomly.

## 6. VISUALIZATION AND INTERPRETABILITY

A number of visualisations were created to support the quantitative assessment of regression models. These visual aids help evaluate the behaviour of the model, spot overfitting or underfitting patterns, and decipher how features affect price prediction.



## 6.1. ACTUAL VS PREDICTED PRICE PLOTS

Each of the four models—Linear Regression, Decision Tree, Random Forest, and Neural Network (MLP Regressor)—has its actual final prices plotted against its anticipated values in Figure 1.

The Decision Tree model performs relatively well, with little deviation; the Random Forest and Neural Network models exhibit nearly perfect alignment along the diagonal, showing outstanding predicted accuracy across the whole range of final prices. However, there is more dispersion in the Linear Regression model, especially at the upper and lower price points. This implies that its ability to represent intricate relationships seen in the data is constrained by its linear assumptions.

## 6.2. RESIDUAL DISTRIBUTION ANALYSIS

Residual plots, which are not displayed here, examine the distribution and trend of prediction errors to further assess model performance. The residuals should ideally be dispersed randomly with no discernible trend and symmetrically distributed around zero.

These requirements were met by the Random Forest and Neural Network regressors, which both showed nearly zero mean residuals with minimal volatility. The limits of the linear structure of Linear Regression were further supported by the systematic under- and over-predictions that were displayed by its residuals.

## 6.3. FEATURE IMPORTANCE (RANDOM FOREST REGRESSOR)

To interpret which variables most influenced the model's predictions, feature importance scores were extracted from the Random Forest model. The results are as follows:

| **Feature** | **Importance (%)** |
|---|---|
| Price (Rs.) | **62.4** |
| Discount (%) | **28.7** |
| Category | **6.2** |
| Payment Method | **2.7** |

- **Price** emerged as the most dominant predictor, which aligns with business logic since base price strongly anchors final price predictions.
- **Discount percentage** significantly influenced predictions, highlighting the role of promotional pricing in e-commerce.
- **Product category and payment method** contributed marginally, though their effects may interact with price sensitivity and customer incentives.

This analysis not only supports model selection but also provides actionable insights for dynamic pricing and marketing strategies in online retail.

## 7. CONFUSION MATRIX INTERPRETATION (OPTIONAL)

It is feasible to reframe the problem as a classification work by discretising the continuous price variable into defined price ranges, even though the main goal of this study is to estimate the final price of a product—a continuous regression task. Especially in decision-support scenarios like pricing segmentation or targeted marketing, this transformation makes it possible to use classification measures like confusion matrices, precision, recall, and accuracy, which provide further insights into model behaviour.

## 7.1. PRICE RANGE CATEGORIZATION

For this classification-based evaluation, the final prices were bucketed into three price categories:

- Low: ₹0–₹120
- Medium: ₹121–₹240

- High: ₹241 and above

These thresholds were chosen based on percentile distributions to ensure balanced representation across categories.

## 7.2. EXAMPLE CONFUSION MATRIX (HYPOTHETICAL)

Assuming the Random Forest model's continuous outputs were post-processed into these discrete categories, a sample confusion matrix might look as follows:

| | Predicted Low | Predicted Medium | Predicted High |
|---|---|---|---|
| **Actual Low** | 45 | 3 | 0 |
| **Actual Medium** | 2 | 48 | 4 |
| **Actual High** | 0 | 5 | 50 |

## 7.3. INTERPRETATION

- The diagonal values represent correct classifications. For example, 45 out of 48 low-priced items were correctly predicted.
- Off-diagonal values indicate misclassifications. Notably, the model confused five high-priced items as medium-priced, which might be tolerable in practice if adjacent-category misclassifications are less costly.
- The overall classification accuracy in this hypothetical case would be (45 + 48 + 50) / 157 = 143 / 157 ≈ 91.08%.

## 7.4. APPLICABILITY

This optional classification-based analysis may be valuable in business scenarios such as:

- Assigning price tiers for personalized offers or discounts
- Customer segmentation based on spending patterns
- Dynamic pricing rules that are based on categorical price bands rather than exact price points

While regression remains the core modeling strategy for precise price estimation, incorporating categorical interpretations adds interpretability and flexibility in downstream business applications.

## 8. CONCLUSION

In order to forecast product prices in an e-commerce setting, this study compared four machine learning regression models: Neural Network (MLP Regressor), Random Forest Regressor, Decision Tree Regressor, and Linear Regression. A real-world dataset including transactional and product-level variables was used to train and test the models, and the R2 score and RMSE were used to assess performance.

The Random Forest Regressor consistently performed better than the other models, with the lowest RMSE (0.89) and the greatest R2 score (0.99995). Its ensemble structure minimised overfitting and successfully caught intricate feature interactions, making it ideal for price prediction in dynamic retail settings.

Near-equivalent performance was shown by the Neural Network (MLP) model, suggesting that it can handle non-linear patterns, particularly in high-dimensional data. It may be more accurate than ensemble models in subsequent research with additional hyperparameter adjustment, deeper topologies, and regularisation strategies.

The results of the Linear Regression model, on the other hand, were the poorest, indicating that it is not very good at simulating the non-linear and multi-modal price correlations that are frequently present in e-commerce transactions. As a baseline or in contexts with interpretable needs, it might still be helpful.

## 9. FUTURE WORK

Building on the results of this work, a number of approaches can be taken to improve the reach, precision, and practicality of machine learning-based price prediction systems in e-commerce settings:

- **Cross-validation and hyperparameter tuning:** Although the current study employed default or basic settings, model performance can be further optimised by putting systematic tuning techniques like Grid Search, Random Search, or Bayesian Optimisation into practice. Furthermore, k-fold cross-validation can offer a more reliable evaluation of the generalisability of the model.
- **Use of Temporal data:** Time-based data, such as the day of the week, month, time of day, or seasonal periods, that are taken from the Purchase Date property can be useful in future models. These characteristics are probably going to capture significant contextual trends in the timing of promotions and purchases.
- **Advanced Deep Learning Architectures:** When time-series or session-based data is available, using sequential models such as Transformer-based models or Long Short-Term Memory (LSTM) can help capture temporal dependencies and changing customer patterns.
- **Explainability and Feature Importance:** Additional feature importance analysis utilising permutation importance or SHAP (SHapley Additive exPlanations) values can aid in guiding feature engineering, revealing model insights, and enhancing interpretability for business users.
- **Integration of External and Contextual Data:** The prediction model can be greatly improved and its use cases expanded by supplementing the dataset with extra information like rival pricing, seasonal indicators, product ratings, customer location, and even macroeconomic signals.
- **Implementation in Production Environments:** Subsequent research endeavours may concentrate on integrating the most efficacious model, like the Random Forest Regressor, into a real-time prediction pipeline. Model serving, API integration, performance tracking, and ongoing retraining using fresh transactional data are a few examples of these procedures.

These additions would increase the model's usefulness for dynamic pricing, customer targeting, and strategic business decision-making in the e-commerce industry in addition to increasing prediction accuracy.

## CONFLICT OF INTERESTS

None.

## ACKNOWLEDGMENTS

None.

## REFERENCES

Scikit-learn documentation: https://scikit-learn.org/

Pedregosa et al., 2011. Scikit-learn: Machine Learning in Python

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning

Brownlee, J. (2020). Machine Learning Mastery: Regression Algorithms

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://scikit-learn.org

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. Springer.

Brownlee, J. (2020). Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch. Machine Learning Mastery.

Zhang, Y., Zheng, Y., & Qi, J. (2018). Deep learning for e-commerce price prediction. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 1020–1028. https://doi.org/10.1145/3219819.3219852

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Wagle, S., & Kakkar, S. (2022). A survey on machine learning approaches for dynamic pricing in e-commerce. Journal of Retailing and Consumer Services, 64, 102821. https://doi.org/10.1016/j.jretconser.2021.102821

Vardhan, H. K., & Kumar, D. (2021). Predictive analytics in e-commerce using machine learning: A case study on pricing models. International Journal of Information Management Data Insights, 1(2), 100017. https://doi.org/10.1016/j.jjimei.2021.100017