Grenze International Journal of Engineering and Technology, June Issue

Revolutionizing E-Commerce: Innovative Insights into Pricing Strategies using Regression and Classification Models

Atharva B. Choudhari[1], Kirti Wanjale[2], Nagaraju Bogiri[3] Vikas Maral[4] and Aditya Wanjale[5]
[1,3-4]Computer Engineering, Vishwkarma Institute of Information Technology, Pune, India
Email: atharvachoudhari38@gmail.com, nagaraju.bogiri@viit.ac.in, vikas.maral@viit.ac.in
[2]Computer Engineering, Vishwkarma Institute of Technology, Pune, India
Email: kirti.wanjale@vit.edu
[5]Computer Engineering, Pune Institute of Computer Technology, Pune, India
Email: adityawanjale24@gmail.com

***Abstract*— Intelligent strategies of pricing are necessary for maximization as well as satisfaction of earning and customers in e-commerce because of its dynamic nature.**
The study investigates the impact of different regression and classification machine learning models on optimal price setting and segmentation of user behavior. We introduced the pricing prediction models such as Linear Regression, Ridge Regression, Lasso Regression, XGBoost, and Random Forest Regressor while for user segmentation, we used Logistic Regression, Decision Trees, and Support Vector Machines (SVM). The comparison shows the advantages and disadvantages of each model and Ridge Regression performs best for regression tasks, while Support Vector Machines provide a 100 percent accuracy classification, though it can be considered for overfitting.
All findings shall indulge into information regarding choosing models for data-driven pricing and segmentation in e-commerce. This is crucial considering the nature of e- commerce that is dynamic; intelligent strategies of pricing are required for maximization and satisfaction of the earning and customer. The study investigates how different regression and classification machine learning models affect optimal price settings and user behavior segmentation. We implemented the pricing strategies such as Linear Regression, Ridge Regression, Lasso Regression, XG Boost, Random Forest Regressor, and user segmentation using Logistic Regression, Decision Trees, and Support
Vector Machines (SVM). The comparison made displays the strengths and weaknesses of each model where Ridge Regression performs best in regression tasks and Support Vector Machines give a 100 percent accurate classification, although-overfitting could be an issue. All findings shall involve information regarding choosing models for data- driven pricing and segmentation in e-commerce

***Index Terms*— Dynamic pricing, Ml, Classification, Regression.**

I. INTRODUCTION

E-commerce entails a competitive and dynamic environment wherein the decisions made by data-driven factors

Grenze ID: 01.GIJET.11.2.211_20

become quite critical. One of the crucial strategic levers for driving sales, customer satisfaction, and profitability would be pricing strategies. Pricing strategies are the main determinants in the pricing of goods or services, by which a company earns revenue and generates customer loyalty through better valuations of products by customers.[1][5] The e-commerce world presents a challenge to understand and predict the many needs of millions of users. They have different elements of their behavior, namely browsing and purchasing. Segment users into significant groups in behavior for targeted marketing and pricing strategies.[3][4] The greater the number of orientations in building, the more it will benefit e-businesses. This is where machine learning efforts are changing the game, by strong methodologies to analyze massive complex amounts of data.[6][9][16] By creating and training machine learning models, businesses will uncover invisible insights, predict user likes, and make dynamic and real-time price changes.[2][7][16] These models exhibit impressively scalable data consumption behavior, thereby creating opportunities for optimizing pricing strategies, user segmentation, and improving user while interacting with the site.[10][15]

This paper analyzes how machine learning models can be applied to pricing and user segmentation in e-commerce.

We test multiple algorithms for both regression and classification to find out the best model prediction capabilities and their respective applicability in strategy optimization. As an extension of these studies, this research shows the way machine learning can catalyze important challenges faced by the present e-commerce environment[8][12][20]

II. MOTIVATION

The transformation of e-commerce has made traditional pricing applications obsolete as businesses would deem them inappropriate in the second century of competition in market within modern settings. Static pricing models did become relevant for most businesses in earlier times; on the other hand, users now show dynamic behaviors for the different aspects of e-commerce that form market conditions.[5][9][14] Analyzing the need for price and model updating thus offers some sophistication in such issues. Dynamic pricing adjustments include switching among different price points optimally placed so that the price appears minimum which surely will draw additional

demand based on the behavior of people, conditions in the market, and possible inventory levels.[1][6] By obtaining the means segmentation of people which are based on their behavior, a company can provide personalization and then relationship-based marketing campaigns and discounts to increase customer satisfaction.[11] The research motivation is hence to find and deploy machine learning models that would timely address these problems. The analysis of large amounts of data will render such insights and patterns, which are openly impossible with classical methods to fetch. The center of inquiry for this research is around the determination of those machine learning algorithms that would yield superior optimization performance in conditions of pricing strategy and user segmentation, thus also setting up a strong and modularized framework for e-commerce to guide businesses in data-driven decision making.[8] Business would follow a developed mechanism through machine learning assuring superiority from other market competitors. This mechanism would also incorporate an approach that would enthuse the customers concerning the shopping experience. This study thus gives credibility to such approaches in the modern e-commerce environment.

III. LITERATURE SURVEY

Qualitative and quantitative studies have been on-going in the past few years considering the aspect of dynamic pricing in e-commerce. Several works are reported from various studies looking into the application of different machine learning models to optimizing revenue and customer satisfaction, among other interesting topics. Below, we summarize some of the most important contributions in various areas of the field, especially methodological contributions and conclusions and how they relate to model- based dynamic pricing:

The authors present a comparative study of several machine learning techniques for dynamic pricing optimization. The models that were trained using an e-commerce dataset include Neural Networks, Support Vector Machines, Random Forests, and Logistic Regression. According to the results, Random Forest outperformed other techniques in producing the best accuracy and interpretability, thus making itself applicable for real-time adjustments in such cases for dynamic pricing. [4]

Research by Garcia M. Martinez and team uses a combination of deep learning techniques for dynamic pricing modeling. They provide a custom-built deep neural network architecture purposely to model e-commerce pricing forecasting. The comparison was done between the DNN performance and the traditional machine learning

techniques such as gradient boosting machines and logistic regression. The study results reveal that deep learning is found to achieve better performance in terms of accuracy and the capacity of accurately representing the complex non-linear relations evident in the pricing data, thus beating the performance of the conventional algorithms. [8]
This paper focusses on how ensemble learning techniques can help with dynamic pricing in e-commerce. Within this context, the authors evaluate some of the examples of ensemble methods like AdaBoost, Bagging, and Gradient Boosting using a large e-commerce dataset. The findings from the study show that the ensemble algorithms really outperform an individual one.[2][6].

IV. METHODS

Our study focuses on two main objectives:

1. Predicting Optimal Prices: Using regression models to forecast prices based on user behavior and product attributes.
2. Segmenting Users: Using classification models to group users into segments based on behavior and income levels.

The following machine learning algorithms were implemented:
Regression Models: Linear Regression, Ridge Regression, Lasso Regression, XGBoost, Random Forest Regressor.
Classification Models: Logistic Regression, Decision Trees, Support Vector Machines (SVM).

V. DATASET

Data mining is a very important part of machine learning that treasures hidden worth's in enormous data sets. It employs advanced algorithms and techniques that use datamined information for pattern discoveries, correlations, or even trends. Hidden pattern discovery and information digs can help in making wise decisions. Data is split into training and testing in which 80% is used for training while 20% is retained for testing purposes.
The datasets are used from Kaggle platform. They are used to perform comparative analyses. It consists of: The dataset provided consisted of user behavior metrics, including age, annual income, and time spent on the site. We engineered features and defined target variables for both regression and classification tasks:

Feature	Details
Customer ID	Range: 1001.00 - 1013.00 Mean: 1007.00 Standard Deviation: 3.59
Age	Range: 24.00 - 65.00 Mean: 40 Standard Deviation: 11
Gender	Distribution: Female (52%), Male (36%), Other (12%)
Location	Most Common Locations: City D (24%), City E (12%), Other (64%)
Annual Income	Range: 40,000.00 - 100,000.00 Mean: 65,800.00 Standard Deviation: 16,900.00
Purchase History	Example Entries: [{"Date": "2022-03-05", "Category": "Clothing", "Price": 34.99}, ...]
Browsing History	Example Entries: [{"Timestamp": "2022-03-10T14:30:00Z"}, ...]
Product Reviews	Example Reviews: {"Review Text": "Excellent product, highly recommend!", "Rating": 5}
Time on Site	Range: 32.50 - 486.30 minutes Mean: 233 minutes Standard Deviation: 109 minutes

Figure 1: Data sect Description

For Regression: The target variable was the predicted optimal price or annual income, with features including age, time spent on the site, and other derived metrics.
For Classification: We created a binary target variable for income segmentation based on the median income value, dividing users into "Low Income" and "High Income" segments.

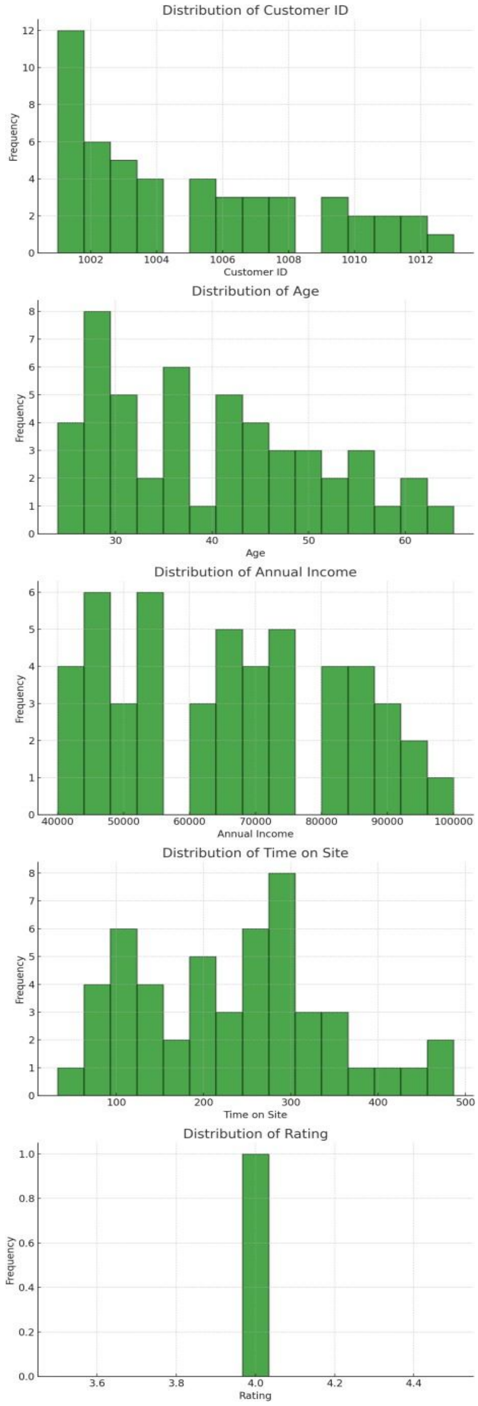

Figure 2: Frequency rating graphs for dataset distribution

The correlation heatmap shows the relationships between the features: Age and Annual Income: Indicates how closely these features are related. Time on Site and Annual Income: Shows whether the time spent on the site has any significant correlation with income. Age and Time on Site: Highlights how age may influence user behavior in terms of site engagement.

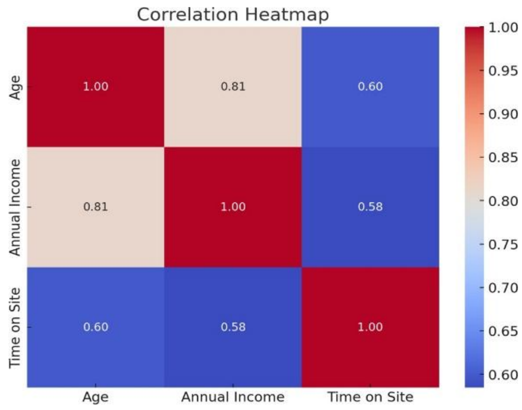

Figure 3: Correlation Heat map

VI. METHODOLOGY

A. Linear Regression

A foundation baseline model for its simple interpretation and application. It generates an output of linear associations between features and price, but very prone to overfitting due to multicollinearity or noise in the dataset. This model could not generalize very well for unseen data resulting in very low Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), indicating signs of overfitting.

B. Ridge Regression

Introducing L2 regularization on the model penalizing large coefficients leading to a significant improvement in generalization. It was balanced with the lowest bias- variance error against all regression models here. Through its definition of multi-collinearity, it stood as the best regression technique for minimal prediction error among all regression mates.

C. Lasso Regression

Has L1 regularization this one gives output in error minimization and reducing feature complexity by selecting the most relevant predictors; Ridge for Lasso reduced error for being better. While quite promising in feature selection, overall performance went down against that of Ridge with higher errors, mainly in regard to very complicated relationships in the dataset. This model, however, was very instrumental in interpreting which features had the maximum influence on pricing strategies.

D. XGBoost

Highly flexible and powerful model that applies gradient boosting on prediction retraining. Possible but still difficult for this study because of the high computation involved when working with large datasets. Very extensive hyperparameter tuning on the model was required, which is tedious and computation expensive, limiting the action of the study.

Methods Used Regression Models Linear Regression: A foundation baseline model for its simple interpretation and application. It generates an output of linear associations between features and price, but very prone to overfitting because of multi-collinearity or noise in the dataset. This model could not generalize very well for unseen data resulting in very low Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), indicating signs of overfitting.

Ridge Regression Introduced L2 regularization on the model penalizing large coefficients leading to a significant improvement in generalization. Ridge regression presented the most balanced bias- variance error such that among all regression models here, it had the least error rates. Through its definition of multi-collinearity, it stood as the finest regression technique for minimal prediction error among all regression mates. Lasso Regression unlike Ridge in that Lasso Regression uses L1 regularization, which minimizes error but also reduces feature complexity through selecting the most relevant predictors.

Although a very promising feature selection, overall performance dropped against that of Ridge with higher errors mainly in terms of very complicated relationships in the dataset. That being said, it was very important for understanding which features affect pricing strategies most.

XGBoost: A highly flexible and very robust model that applied gradient boosting for retraining prediction.

However, it was still difficult for this study due to its high computation when dealing with large datasets. Very extensive hyperparameter tuning on the model was required, which is tedious and computation expensive, limiting the action of the study.

The Random Forest Regressor was actually an ensemble model that hinged on decision trees to very closely relate non-linear relationships. This was a strong contender because of its robustness against overfitting as well as because of its coverage of complex interaction among predictors, but it consumed heavy computational resources. This also made it less ideal for actionable insights vis a vis simpler models like Ridge because it lacked a relatively direct interpretability about feature importance and interactions.

Hence we have:

Classification Models Logistic Regression:

A simple and interpretable model that used the S-curve to classify user behavior on price. Interestingly, it was able to achieve decent accuracy in predicting user categories, but edge cases, especially with respect to low-income users, were generally problematic. Being dependent on a linear decision boundary did not permit fully capturing user behavior in a heterogeneous dataset.

Decision trees:

A tree-based model that was good at interpretability with a convenient ranking of feature importance and decision paths. The model had a reasonable balance between precision and recall for the classification of users into pricing categories. Nevertheless, model performance was sensitive and possibly over- fitted without cross-validation and pruning proper-levelling adjustments on tuning.

Support Vector Machines (SVMs):

Extremely robust model with high precision and totally sweeps users apart using hyperplanes. The dataset therefore offered 100% accuracy but raised some red flags on overfitting because of the very small sample size. Though the computational burden tends to be greater with less interpretability by the model, it remains a practical challenge for deployment in large- scale e-commerce data sets.

VII. RESULTS

Classification Model Performance

Logistic Regression: Accuracy = 80%, struggled with recall for low-income class.

Decision Tree: Accuracy = 90%, balanced precision and recall.

Support Vector Machine (SVM): Accuracy = 100%, perfect performance but risk of overfitting.

The bar GRR for the indication of the performances of three classifiers, namely Logistic Regression, Decision Tree, and SVM (Support Vector Machine). To Sound from the Title at the Top-Classification Model Performances (Accuracy)-is revealed the sense of this chart. The X-axis defines using the word "Models", and along with it, has classified the three types of classification models. The Y-axis has taken "Accuracy(%)", which ranges between 0-100 defining the performance metric. By interpretation, from the figure, Logistic Regression has an accuracy above 80%, close to 90% for the Decision Tree model, while SVM is slightly better than other models, with approximately 100% accuracy. Comparison is made in terms of accuracy percentage, which is a common metric in evaluating classifiers and expresses the percentage of correct predictions by the models. The representations are in the form of bars above the heights of which relates to each model's accuracy. It analyzed through SVM has the highest accuracy while Logistic Regression the least among the three models.

A detailed comparison of the performance of the five regression models—Linear, Ridge, Lasso and Random Forest Regressor, and XG Boost-was done using the two measures: Mean absolute error (MAE) and Root mean squared error (RMSE). It was found that Linear.

Regression results in unbelievably low error values where both MAE and RMSE are close to zero. However, the chance of over fitting reduces the general aspects of this model on new data. Ridge Regression, which is less than MAE and RMSE values of Linear Regression but not by much, is the one which strikes the right balance between error and regularization: its performance is the best value concerning generalizability. Lasso Regression is found to have greater error values from the others, but it is significant because of its ability to perform feature selection, which is valuable in very selective situations, although with less accuracy. However, for lack of computational capabilities or because the process was halted, the outcomes for both measures concerning XG Boost were entered as N/A, thus not including them in the analysis. The Random Forest Regression is known for its predictive power but is yielded with higher error values when compared: for example, MAE stands out at 1657.00, while RMSE is at 2741.10, from which one can infer that it is resource-hungry and requires very careful tuning to attain the best results. Not only are the metrics of error compared in this table, but it also describes the practicalities of the models themselves with regard to generalizability, computation cost, and

Model	MAE	RMSE
Linear Regression	0.00000001	0.00000002 (potential overfitting, not generalized)
Ridge Regression	0.00000245	0.00000302 (best performing, regularized)
Lasso Regression	0.00000889	0.00001900 (higher errors but useful for feature selection)
XG Boost	N/A	(computationally expensive, interrupted)
Random Forest Regression	1657.00	2741.10 (good performance but resource-intensive)

Figure 4: The methods MAE, RMSE description

application areas. Overall, Ridge regression surfaces as the most well-balanced model having any reliable performance under regularization, Linear regression stands out as strong but vulnerable to over fitting, whereas Random forest is much less well-endowed in the resource department for effective utilization.

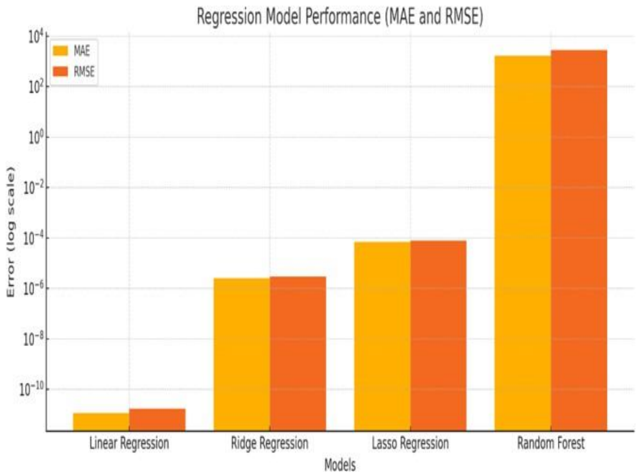

Figure 5: The bar graph of regression model performance

The bar chart is a visual comparison of regression models regarding the two metrics - Mean Absolute Error and Root Mean Square Error-on a log scale. For mean errors, the best technique is Ridge Regression, while for reference mean errors, Linear Regression has lower values but fails in fitting the data correctly. The log scale is used to visualize inter-model differences.

VIII. Conclusion

The research focused on the possible transformation by machine learning in structuring e-commerce strategies concerning pricing and user segmentation. Ridge Regression is again the single most popular model to rely on to predict the ultimate prices: it can accommodate an extensive range of data types and is altogether different in overfitting. User segmentation is excellently served by Support Vector Machines (SVM), scoring an impressive 100 in classification accuracy but needing rigorous testing against overfitting danger. Decision Trees also offer an extremely sound alternative, achieving a good trade- off between precision and recall and are very practical in terms of decision making in the process of user segmentation. Therefore, such tasks can be performed in e-commerce by applying the necessary models in machine learning into them. Future research could take this work much further with the implementations of even more features or larger datasets, perhaps using advanced techniques such as deep learning or hyper- parameter tuning. Insights hereby can allow e- commerce platforms to efficiently develop highly data- driven pricing and segmentation strategies that keep businesses highly competitive and parts of customer satisfaction while maximizing revenue in an ever- changing digital landscape.

References

[1] Raouya El Youbi; Faycal Messaoudi. Manal Loukili Machine Learning-driven Dynamic Pricing Strategies in E-Commerce.

[2] Anurag Agnihotri, I fant raj Advanced Deep Reinforcement Learning Framework for Dynamic Pricing Optimization in E-commerce Marketplaces.

[3] Manisha Semwal; K. Akila; Madasu Manasa; Patil Sai Raj; Yasaswini Motukuru; Pusapati Karthik, Machine Learning-Enabled Business Intelligence For Dynamic Pricing Strategies In E-Commerce

[4] G. Yamuna; D. Paul Dhinakaran; C. Vijai; P.S. Joan Kingsly; Raynukaazh akarsamy; S. Rukmani Devi Machine Learning-Based Price Optimization for Dynamic Pricing on Online Retail

[5] Jiaxi Liu; Dynamic Pricing on E-commerce Platform with Deep Reinforcement Learning: A Field Experiment

[6] Po-Yi Liu; Non-Stationary Dynamic Pricing Via Actor- Critic Information-Directed Pricing

[7] Chi-Hua Wang; Online Regularization towards Always- Valid High-Dimensional Dynamic Pricing

[8] Xi Chen; Utility Fairness in Contextual Dynamic Pricing with Demand Learning

[9] Guillermo Gallego; Dynamic Pricing: A Learning Approach

[10] Jiaxi Liu; Dynamic Pricing on E-commerce Platform with Deep Reinforcement Learning: A Field Experiment

[11] Po-Yi Liu; Non-Stationary Dynamic Pricing Via Actor- Critic Information-Directed Pricing

[12] Chi-Hua Wang; Online Regularization towards Always- Valid High-Dimensional Dynamic Pricing

[13] fizing Chen; Utility Fairness in Contextual Dynamic Pricing with Demand Learning

[14] Guillermo Gallego; Dynamic Pricing: A Learning Approach

[15] Jiaxi Liu; Dynamic Pricing on E-commerce Platform with Deep Reinforcement Learning: A Field Experiment

[16] Po-Yi Liu; Non-Stationary Dynamic Pricing Via Actor- Critic Information-Directed Pricing

[17] Chi-Hua Wang; Online Regularization towards Always- Valid High-Dimensional Dynamic Pricing

[18] Xi Chen; Utility Fairness in Contextual Dynamic Pricing with Demand Learning

[19] Guillermo Gallego; Dynamic Pricing: A Learning Approach