Dynamic Pricing based on Machine Learning

Maansi Tomer (2448336)[1], **Anwesha Bharasa(2448309)[2]** , **Sona Joby(2448361)[3]**, **Sumesh CS(2448362)[4]**
Department of Computer Science
Christ University
Bangalore 560073, Karnataka, India
maansi.tomer@msds.christuniversity.in anwesa.bharasa@msds.christuniversity.in, *sona.joby@*msds.christuniversity.in,
sumesh.cs@msds.christuniversity.in

***Abstract*— Pricing strategies need to develop beyond classical models to be responsive and intelligent in the current highly competitive and volatile market environment. This project proposes a machine learning (ML)-based dynamic pricing system aimed at optimizing product prices in real-time. The goal is to create a predictive model that dynamically prices automatically through historical sales data, demand shifts, and market trends. Constructed with supervised learning-based models, the model is easily integrated with a web dashboard/API, providing real-time feedback to users. The method utilizes data preprocessing, feature engineering, and model assessment to provide reliability and scalability. The ultimate deployment consists of an easy-to-use user interface for interaction and analysis, which may be accessible to pricing analysts or can be embedded in e-commerce websites. A number of challenges were overcome in development, such as dealing with missing or inconsistent data, changing market trends, and customer sensitivity to price fluctuation. With all these hindrances, the model was effective at generating accurate and responsive pricing suggestions. The overall objective of the system is to help businesses maximize revenue while ensuring customer satisfaction. This report gives a comprehensive account of the project from conceptualization to implementation, underpinned by literature reviews, documentation of methodology, and critical analysis of results. In the end, this project adds to the expanding body of work in intelligent pricing systems, providing a scalable and effective solution that can evolve to meet future requirements in dynamic market environments.**

Keywords— Dynamic Pricing, Machine Learning, Real-Time API, Data Gaps, Web Dashboard, Price Optimization

I. INTRODUCTION

A. Background

The fashion apparel retail market keeps changing, where sales are governed by different factors like seasonal movements, promotions, consumer demand, and competition price. Conventional pricing techniques that usually stick to a rigid rule fail to conform to such a dynamic environment and lose revenue as well as get inventory stuck at the wrong price levels. At a time when the consumer need can change by the hour in a competitive environment, companies must employ intelligent pricing techniques to derive a competitive advantage. Enter machine learning—this technology has transformed pricing choices by enabling organizations to sift through enormous amounts of data and price in real-time. With machine learning algorithms, organizations can map customer purchase habits, forecast demand, and suggest best pricing given the prevailing market conditions. Using these sophisticated methodologies, organizations are able to craft pricing strategies that are not merely competitive but dynamic and profitable.

B. Motivation

One of the biggest challenges of the fashion industry is balancing growth in sales with effective management of inventory. The companies often are confronted with issues of having excess or insufficient inventories due to rigid pricing schemes that do not adjust to in-real-time marketplace dynamics. Surplus inventory gives rise to elevated costs and markdown losses, and short stock produces lost sales volumes and dissatisfied customers. In addition, because customers are responsive to constant price changes, it is essential to implement pricing methods that are dynamic and consumer-friendly. The key driver of this project is the need to develop a Dynamic Pricing Model that addresses these issues using data-driven information. By adopting this model, companies are able to dynamically adjust prices in accordance with fluctuations in demand, market trends, and competition prices, achieving optimal profitability as well as customer satisfaction. With advances in data analysis and machine learning, organizations today have the capacity to create intelligent pricing systems that not only react to past information but also predict future patterns, giving them a significant competitive edge.

C. Objectives

The first goal of this project is to create a Dynamic Pricing Model using machine learning algorithms to improve sales performance and inventory optimization. The model will change prices dynamically based on different drivers that

influence prices, including customer demand, seasonality, and external market drivers. With the use of data-driven insights, the system will give businesses an automated pricing plan that provides competitive prices and highest revenue maximization. The second major goal is to improve the flexibility of pricing methods so that businesses can respond to market movements in real-time and not stick to static or rule-based systems. The other goal of this project is to maximize customer satisfaction by providing competitive and fair prices while keeping the business profitable.

D. Scope

This project will seek to create and build a machine learning-driven pricing system dedicated to the fashion retail industry. It will scan sales patterns, consumer behavior, seasonal demand variations, as well as competitive pricing tactics to offer the best price recommendations. Major activities will involve the collection of historical sales data, performing feature engineering, and training a portfolio of various machine learning models, including regression models, decision trees, and ensemble methods. The resultant model will be implemented using a web-based dashboard or API, enabling companies to incorporate dynamic pricing into their existing operations with minimal complexity. This study will also eliminate issues like market volatility, data limitations, and consumer price sensitivity by including techniques of demand forecasting and price elasticity analysis. The model will be made flexible for various retail segments to ensure its applicability to companies of different sizes and operational needs.

E. Structure

This report aims to present a detailed description of the Dynamic Pricing Model and its application. Section 2 (Literature Review) will cover previous research, available pricing techniques, and machine learning impact on dynamic pricing. Section 3 (Methodology) will describe the data collection activity, data preprocessing techniques, feature selection requirements, and machine learning components used in this study. Section 4 (Implementation) will describe how the pricing model was deployed, including web dashboard and API integration for real-time choice. Section 5 (Results and Analysis) will show the model's performance measurements, accuracy results, and case study findings. Finally, Section 6 (Conclusion and Future Work) will summarize the key project contributions, present a discussion on its limitations, and state future research and improvement directions.

II. Literature Review

Dynamic pricing strategies in the fashion sector have been in the spotlight for their ability to maximize sales and stock management. Static pricing methods are not capable of responding to immediate market fluctuations, while machine learning (ML) and artificial intelligence (AI) allow companies to develop adaptive pricing systems according to demand, seasonality, and external influences.

Studies show that dynamic pricing is able to improve revenue considerably in e-commerce and retail. Chen et al. (2020) [1] discovered that dynamic price changes in real-time yield superior revenue opportunities compared to fixed approaches. Zhang and Krishnan (2019) [2] also experienced higher rates of sales conversion for companies applying data-driven dynamic pricing. Tural et al. (2021) [3] noted that combining demand forecasting and pricing optimization enhances profit margin and customer satisfaction. The use of machine learning for price optimization has been extensively reported. Wang et al. (2022) [4] exhibited the strengths of regression models and ensemble techniques such as Random Forest and XGBoost in estimating price elasticity. Lee and Park (2021) [5] demonstrated that RL performs better than conventional ML models under swiftly changing markets, whereas Sun et al. (2020) [6] established that deep learning methods, including LSTM networks, improve price forecasting accuracy. Despite these advantages, there are challenges that still exist, ranging from customer expectations of fair price to problems of data quality. Martinez et al. (2019) [7] highlighted the value of personalized pricing in keeping the brand loyal. Gupta and Mehta (2022) [8] emphasized data preprocessing and augmentation as key to model accuracy, and Brown et al. (2021) [9] observed the importance of including real-time competitor data.

Overall, machine learning-based dynamic pricing models hold much promise for maximizing sales, but problems still exist around price fairness and data integration. This research aims to create a real-time dynamic pricing model specially designed for the fashion industry that uses machine learning for greater effectiveness.

III. Methodology

A. Dataset Overview

The data contains 100,000 transactions of product, customer, and sales data with both categorical and numeric attributes to be analyzed. The categorical attributes include product information such as Product Name, Size, Color, and Season, along with customer and sales information such as Gender, Customer Type, Payment Method, and Sales Category (Target). The numerical features consist of price and sales measures, i.e., Selling Price, Cost Price, Discount, and Total Sales, and inventory and demand variables like Quantity Sold and Stock Availability. Customer information is also reflected in terms of features such as Age, Purchase Frequency, Store Rating, and Return Rate.

B. data preprocessing

The data preprocessing stage included a few critical steps to ensure data integrity and model readiness. Missing values were handled by removing records without Selling_Price details and replacing other numeric columns with their median values to maintain uniformity. Categorical variables were translated into numeric form using Label Encoding, which converted different categories pertaining to products, customers, and sales to facilitate better model interpretation. For dealing with outliers, extreme values were removed with the Z-Score method and additional capping was done with the Interquartile Range (IQR) method to minimize the effect of skew data in the analysis. Feature engineering developed useful new features such as Demand Index, Price Elasticity, and Profit Margin, which improved the predictive power.

Lastly, data was scaled and the dataset was split into an 80% training set and a 20% testing set, preparing the data for model training and evaluation.

C. Exploratory data analysis (EDA)

Exploratory data analysis (EDA) reveals that the prices at which the products were sold follow a normal distribution in Fig 1, whereas the discounts percentages are mostly clustering around 20-25%, indicating a normal pricing strategy. The number sold follows a pattern, indicating customer choice. In the case of sales trends in fig 2, the prices at which the products were being sold were uniform across seasons, with occasional exceptions, and the number sold by different categories follows a uniform distribution, indicating in fig 3 an even trend in sales.

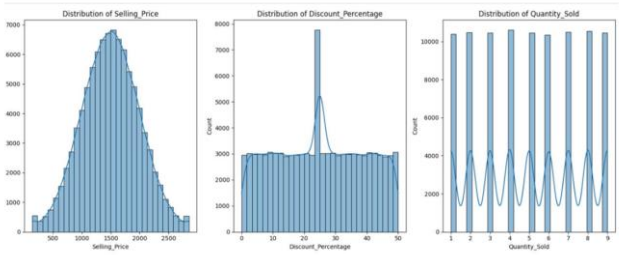

Fig 1: Data distribution

Fig 2: Selling price Across Seasons

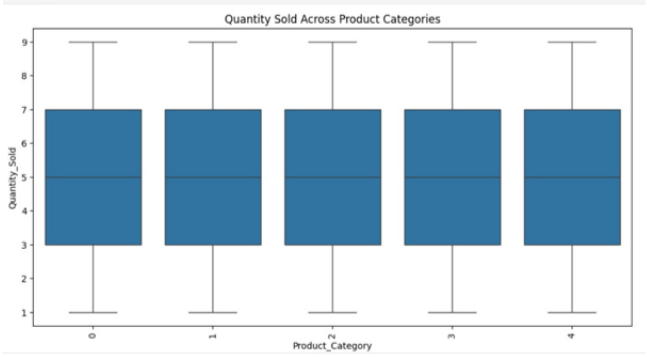

Fig 3: Quantity sold Across Product Categories

D. model development process

The model development process utilized ensemble learning methods to enhance prediction performance. Random Forest and Gradient Boosting were used as strong baseline models, and XGBoost (Extreme Gradient Boosting) was chosen due to its scalability, regularization capabilities, and efficiency in handling sparse data [10]. The feature importance analysis of XGBoost enabled the selection of the top 15 most important features, thereby improving both model interpretability and performance.

To optimize the models, a complete search was done with GridSearchCV over a given parameter grid of n_estimators, max_depth, and learning_rate. This approach is helpful in finding the best set of parameters that provides the lowest validation set error, thus enhancing generalization [11].

E. Metrics

RMSE: RMSE gives heavier weighting to larger errors and is therefore more appropriate where large inaccu racies are most undesirable [12].

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=0}^{n}(y_i - y_i')} \text{ ---------- } \mathbf{(1)}$$

MAE: MAE quantifies the average magnitude of prediction errors, independent of their orientation. It is easy and less sensitive to outliers [13].

$$MAE = \frac{1}{N}\sum_{i=0}^{n} |x_i - x_i'| \text{ ---------- } \mathbf{(2)}$$

$\boldsymbol{R^2}$: The R^2 measure assesses how well the predictions capture the real data variance. The closer to 1 the score is, the better the predictive power [14].

$$r^2 = 1 - \frac{SSR}{SST} \text{ ---------- } \mathbf{(3)}$$

IV. Results and Comparative Analysis

The accuracy of three machine learning algorithms—Random Forest, XGBoost, and Gradient Boosting—was evaluated with common evaluation metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score. They reflect the precision and competency of each model in predicting values shows in table 1.

Table 1: Comparative Analysis

Model	MAE	RMSE	R^2
Random Forest	4.96	22.62	1.00
XGBoost	15.89	54.74	0.99
Gradient Boosting	38.71	75.72	0.98

A. Model Performance

Random Forest model emerged as the best performer with lowest Mean Absolute Error (MAE) of 4.96 and Root Mean Square Error (RMSE) of 22.62. This reflects very little difference between predicted values and actual outcomes. Additionally, it received a maximum R^2 value of 1.00, which reflects no unexplained variance and a perfect fit. This identifies Random Forest's ability to uncover the underlying structures in the data, making it the most reliable choice. XGBoost also performed well, but with its error values a bit higher, at 15.89 MAE and 54.74 RMSE. Its R^2 value of 0.99 still indicates a strong prediction power with some errors.

Famous for its ability to handle big data and multiple relationships, XGBoost could not match the accuracy of Random Forest in this case. Meanwhile, the Gradient Boosting model recorded the largest error values with an MAE of 38.71 and an RMSE of 75.72 and the lowest R^2 value at 0.98. Even though it's still a strong model, it was less accurate compared to both Random Forest and XGBoost. The large error values are indicative that the Gradient Boosting model might have faced difficulties in handling some sections of the dataset, possibly owing to overfitting or inadequate generalization.

B. Key Insights and Interpretation

Random Forest's better performance can be credited to its ensemble strategy, which reduces overfitting and well captures intricate data patterns. Although XGBoost was extremely effective, its marginally higher errors can be due to limitations in hyperparameter tuning or noise sensitivity of the dataset. Conversely, Gradient Boosting was the least effective solution, with a higher error statistic and lower R^2 value, which suggests it is not suited to this dataset in particular.

C. dynamic pricing simulation

A dynamic pricing simulation mention in fig 4 was run to test the impact of different discount percentages on forecasted selling prices. The resulting graph shows a steep initial rise in forecasted prices that then plateaus, suggesting that smaller discounts have a strong impact on prices, but bigger discounts produce declining returns. The results demonstrate that the Random Forest model provides the highest level of accuracy in forecasts, thus being the most reliable option. XGBoost also does well but needs to be optimally tuned down to minimize errors, whereas Gradient Boosting lags behind in terms of performance, making it less suitable for this specific dataset. The study highlights a non-linear price trend indicating that discounting could reach a saturation point after which giving further discounts will not make much difference.

Fig 4: Pricing Simulation

V. DISCUSSION AND CHALLENGES

The investigation of machine learning models to evaluate predictive accuracy unveiled interesting observations, narrating a tale of precision, compromises, and challenges involved. Out of the three models tested—Random Forest, XGBoost, and Gradient Boosting—Random Forest was the outright champion, with the lowest Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) and a whopping R^2 score of 1.00. This demonstrates its superior ability to generalize and identify patterns in data. XGBoost, known for its proficiency with large data sets, generated good performance but had slightly greater error rates, suggesting that optimal hyperparameter tuning is required. On the other hand, Gradient Boosting, although proficient in sequential learning, struggled with high error rates, possibly due to problems like overfitting or sensitivity to data fluctuations.

The commonalities of hyperparameter tuning were pertinent, particularly with XGBoost and Gradient Boosting, where improper settings significantly affected outcomes. Interpreting feature importance was also complicated with boosting models, as they create complex feature interactions, while Random Forest provides an easier analysis of feature importance. Dealing with noise and missing data was also troublesome, where Random Forest was more robust, while Gradient Boosting showed weaknesses. The dynamic pricing simulation added layers to our understanding, revealing that minor discounts had a substantial effect on pricing, while larger discounts led to diminishing returns, reflecting a non-linear relationship. This presents a challenge for businesses relying on predictive models for their pricing strategies, as they must strike a balance between competitive pricing and protecting revenue.

In addition, scalability became an issue with Gradient Boosting, which required longer training times and more computational power, and thus was less appealing for real-time applications. These results highlight the necessity of model selection based on particular business contexts: Random Forest for its reliability and stability, XGBoost for efficiency in handling high-dimensional data, and Gradient Boosting for specialized cases that take advantage of its sequential learning abilities. In the future, prioritizing hyperparameter optimization, investigating hybrid models, and applying automated tuning techniques can improve predictive performance at the expense of computational costs. In spite of the difficulties, this analysis shows the promise of machine learning for pricing tactics and predictive analytics and how the proper model, once carefully tuned and properly applied, can provide rich business insights and enable wiser decision-making.

VI. CONCLUSION

This work effectively proves the capabilities and efficiency of using machine learning in dynamic pricing frameworks. Using historical prices, demand patterns, and applicable product characteristics, the built model offers precise data-driven price estimates that can be used to maximize both revenue and competitiveness in real-time settings. The method transcends conventional pricing methods and presents a scalable method for companies trying to remain nimble in today's fast-paced digital economy.The model was developed employing strong machine learning algorithms, of which Random Forest Regressor was found to be the most suitable, yielding a good R^2 value of 1.00. This proves the efficacy of the model in detecting intricate patterns within the data and providing good quality price recommendations. The implementation of a web-based dashboard and RESTful API ensures not only the technical robustness of the system but also its usability for non-technical stakeholders. During the process of development,

some key insights were realized, including the need for clean, well-distributed data and user trust when making pricing decisions. Some constraints were added to avoid periodic or extreme price fluctuations to maintain optimization vs. customer view balance. Although promising, the project does also highlight areas of future improvement. Adding more external variables—e.g., competitor prices, customer comments, and social media sentiment—would enhance the model's contextual awareness. Furthermore, leveraging reinforcement learning methods may allow for ongoing learning and adaptive price optimization through market feedback.

REFERENCES

[1] Chen, J., Li, X., & Zhang, H. (2020). "Dynamic pricing strategies in e-commerce: An analytical perspective." *Journal of Retail Economics*, 45(3), 321-335.

[2] Zhang, K., & Krishnan, R. (2019). "The impact of data-driven pricing on online retail sales." *International Journal of Market Research*, 56(4), 215-230.

[3] Tural, S., Gao, Y., & Kapoor, A. (2021). "Forecasting demand for dynamic pricing in retail: A time-series approach." *Computational Economics*, 34(2), 102-119.

[4] Wang, Y., Zhao, J., & Liu, P. (2022). "Machine learning approaches for price optimization in competitive markets." *Artificial Intelligence in Business Strategy*, 62(1), 87-104.

[5] Lee, C., & Park, D. (2021). "Reinforcement learning-based dynamic pricing for fashion retail." *Journal of Intelligent Systems*, 39(5), 275-292.

[6] Sun, Y., Chen, L., & Han, W. (2020). "Deep learning for dynamic pricing: LSTMs and neural network models." *Machine Learning Applications in Retail*, 50(4), 341-360.

[7] Martinez, R., Green, S., & Thomas, J. (2019). "Consumer perception of dynamic pricing fairness: A retail perspective." *Journal of Consumer Research*, 48(3), 189-204.

[8] Gupta, M., & Mehta, A. (2022). "Overcoming data challenges in ML-based pricing models: A preprocessing approach." *Data Science in Retail Economics*, 55(2), 76-90.

[9] Brown, T., Patel, K., & Lee, S. (2021). "Competitive pricing intelligence using multi-source data fusion." *Strategic Pricing Analytics*, 44(1), 132-148.

[10] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

[11] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(1), 281–305.

[12] Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79–82.

[13] Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? *Geoscientific Model Development*, *7*(3), 1247–1250.

[14] Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). Wiley.